



AUTOMATION

23. Leitkongress der Mess- und Automatisierungstechnik

AUTOMATION 2022

Automation creates Sustainability

Baden-Baden, 28. und 29. Juni 2022

VDI-BERICHTE

Herausgeber:

VDI Wissensforum GmbH

Bibliographische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet unter www.dnb.de abrufbar.

Bibliographic information published by the Deutsche Nationalbibliothek (German National Library)

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliographie (German National Bibliography); detailed bibliographic data is available via Internet at www.dnb.de.

© VDI Verlag GmbH · Düsseldorf 2022

Alle Rechte vorbehalten, auch das des Nachdruckes, der Wiedergabe (Photokopie, Mikrokopie), der Speicherung in Datenverarbeitungsanlagen und der Übersetzung, auszugsweise oder vollständig.

Der VDI-Bericht, der die Vorträge der Tagung enthält, erscheint als nichtredigierter Manuskriptdruck.

Die einzelnen Beiträge geben die auf persönlichen Erkenntnissen beruhenden Ansichten und Erfahrungen der jeweiligen Vortragenden bzw. Autoren wieder. Printed in Germany.

ISSN 0083-5560

E-ISBN 978-3-18-102399-0

ISBN 978-3-18-092399-4

Vorwort 1

► **Fertigungsautomation – Digitaler Zwilling: Vom Konzept zur Anwendung**

Anwendungsfälle und Ansatz zur Erstellung des Digitalen Zwillings aus Sicht eines
Komponentenherstellers 5

V. Stegmaier, Graduate School of advanced Manufacturing Engineering, Universität Stuttgart;
W. Schaaf, J. Schmalz GmbH, Glatten;
N. Jazdi, M. Weyrich, Institut für Automatisierungstechnik und Softwaresysteme,
Universität Stuttgart

Interaktion zwischen Steuerungen auf der Basis von OPC UA FX und deren Konfiguration
durch Verwaltungsschalen 19

C. Diedrich, T. Werner, Otto-von-Guericke-Universität, Magdeburg;
M. Riedl, ifak e.V. Magdeburg

► **Fertigungsautomation – Einsatz der Verwaltungsschale**

Framework for bringing wireless communication into the context of Industrie 4.0 31

G. Cainelli, L. Underberg, L. Rauchhaupt, Institut für Automation und Kommunikation e.V.
Magdeburg

Verwaltungsschale-basierter Ansatz für die Umsetzung von
auftragsgesteuerter Produktion. 45

C. Urban, A. Belyaev, Otto von Guericke Universität Magdeburg;
C. Diedrich, ifak – Institut für Automation und Kommunikation e.V.

Verwaltungsschalen zum anlagenweiten Testen wandelbarer Produktionsanlagen 59

S. Schäfer, D. Schöttke, T. Kämpfe, V. Denkov, Hochschule für Technik und Wirtschaft Berlin;
B. Tauber, EAW Relais-technik GmbH, Berlin

► **Fertigungsautomation – Konzepttransfer in die Fertigungsautomation**

Dienste-Design für modulare Fertigungsanlagen mit individuellen Produkten – Variantenvielfalt der Dienste	75
P. Habiger, G. Hildebrandt, R. Drath, Hochschule Pforzheim; A. Fay, Helmut-Schmidt-Universität, Hamburg	

► **Prozessautomation – Safety und Integrated Engineering**

Typenbasiertes Engineering für modulare Anlagen: Fallstudie zur Umsetzung eines modularen Neutralisationsprozesses	89
J. Lorenz, A. Klose, I. Viedt, J. Mädler, V. Khaydarov, L. Urbas, Technische Universität Dresden	
Datenbasierte Überwachung von reproduzierbaren, chemischen Batchprozessen für modellbasierte Sicherheitsfunktionen.	101
P. Langlitz, F. Lenhart, M. Roser, BASF SE, Ludwigshafen	

► **Prozessautomation – Security**

Security-Entscheidungen „by Design“ in das Engineering prozesstechnischer Anlagen integrieren – Konzept der „Automation Security by Design Decisions“	115
S. Fluchs, admeritia GmbH, Langenfeld, Helmut-Schmidt-Universität/Universität der Bundeswehr, Hamburg; E. Tastan, R. Drath, Hochschule Pforzheim; M. Mertens, INEOS Manufacturing Deutschland GmbH, Köln; J. Ritter, A. Horch, HIMA Paul Hildebrandt GmbH, Brühl; A. Fay, Helmut-Schmidt-Universität/Universität der Bundeswehr, Hamburg	
AutomationML-basierte Modellierungsansätze für ein Security-Engineering-Informationsmodell.	133
E. Tastan, R. Drath, Hochschule Pforzheim; S. Fluchs, admeritia GmbH, Langenfeld, Helmut-Schmidt-Universität/Universität der Bundeswehr, Hamburg	
IT-Security für Automatisierungssysteme mit Ethernet-APL-Feldgeräten	149
K.-H. Niemann, Hochschule Hannover; S. Merklin, Endress + Hauser Digital Solutions, Reinach, Schweiz	

► Prozessautomation – Digital Plant & Sustainability

Materialverfolgungsbasierter Softsensor für die Überwachung von Treibhausgasemissionen in der Prozessindustrie	161
C. Song, J. Schlake, ABB Corporate Research Center, Ladenburg;	
S. Emmrich, ABB Division Process Industries, Cottbus	

► Prozessautomation – Advanced Control

Anforderungen aus der dynamischen Nutzung von grünem Wasserstoff in Power-to-X-Wertschöpfungsketten	175
I. Viedt, M. Mock, J. Lorenz, J. Mädler, L. Urbas, Professur für Prozessleittechnik und Arbeitsgruppe Systemverfahrenstechnik, Technische Universität Dresden;	
A. Klose, H. Lange, Process-to-Order-Lab, Technische Universität Dresden;	
D. Erdmann, Verband Deutscher Maschinen- und Anlagenbau (VDMA), Frankfurt am Main;	
L. Scholz, A. Markaj, A. Fay, Helmut-Schmidt-Universität / Universität der Bundeswehr Hamburg	
Steuerung und Regelung von Abwasserkanalnetzen – ein praxisgerechter Ansatz zur modell-prädiktiven Regelung.	187
M. Schütze, J. Alex, ifak e. V., Magdeburg	

Nonlinear Model Predictive Control: from Digital Twin to Real-World Application in Polymerization Process	205
B.-M. Pfeiffer, C. Leingang, C. Busse, M. Oppelt, Solutions Process Automation, Simulation & Digitalization, Siemens AG, Erlangen and London, Great Britain;	
O. Lorenz, Siemens AG, Karlsruhe,	
A. Cano, Siemens PSE, Parsippany, USA	

► Prozessautomation – KI in der Prozessautomation

Kognitive Automations-Architektur für Cyber-Physische-Produktionssysteme (CPPS).	219
A. Mankowski, D. Antonow, N. Moriz, H. Trsek, inIT – TH OWL, Lemgo	

► **Prozessautomation – KI in der Prozessautomation**

Data Mining in der Prozesstechnik als Key-Enabler intelligenter Digitaler Zwillinge für eine datengetriebene Optimierung der Prozessführung. 231

M. Krüger, B. Vogel-Heuser, K. Land, Lehrstuhl für Automatisierung und Informationssysteme, Technische Universität München;
G. Grim, J. Lorenzer, M. Hanf, Andritz Separation GmbH, Vierkirchen

Künstliche Neuronale Netze (KNN) zur Plausibilisierung verfahrenstechnischer Prozessdaten 245

T. Unglaube, Bayer AG, Haan;
L. Däubler, A. Aljandali, Ostfalia Hochschule, Salzgitter

► **Methoden und Synergien – Verwaltungsschale und deren Anwendung**

Konzept und Realisierung für die automatisierte Erstellung von Verwaltungsschalen in Unternehmen basierend auf bestehenden Daten aus verschiedenen Informationssystemen . 263

I. Garmaeve, T. Miny, T. Kleinert, Lehrstuhl für Informations- und Automatisierungssysteme für die Prozess- und Werkstofftechnik (IAT), RWTH Aachen University;
A. Schüller, P. Bitterlich, Yncoris GmbH & Co. KG, Hürth

Supporting Product Lifecycle Information Management and Distributed Engineering with an Integration Model 277

M. Freund, C. Barth, Festo SE & Co. KG, Esslingen;
P. Juhlin, ABB Corporate Research Center Germany, Ladenburg;
S. Lehnhoff, EngRoTec Consulting GmbH, Esslingen;
M. Winter, G. Fachinger, RWTH Aachen University, Aachen

► Methoden und Synergien – Modulare Anlagen und Orchestrierung

Vergleichende Untersuchung der Orchestrierung modularer Anlagen in der Prozessindustrie, der Fertigungsindustrie und der produktionsnahen Logistik. 289

M. Blumenstein, A. Fay, L. Beers, Helmut-Schmidt-Universität, Hamburg;

A. Stutz, M. Maurmaier, Siemens AG, Karlsruhe

Orchestrierung hybrider Anlagen – auf dem Weg zur Integration von PackML Komponenten in Modulare Anlagen 307

S. Grüner, M. Hoernicke, ABB Corporate Research Center, Ladenburg

Selbstorganisation von Softwarekomponenten für produktive Aufgaben 319

M. Riedl, ifak e.V., Magdeburg;

H. Zipper

► Methoden und Synergien – Security

Vertrauensbasierte Integration von Geräten mittels OPC UA Device Provisioning 329

F. Volkmann, N. Baur, S. Höme, Siemens AG, Nürnberg;

A. Palmin, M. Madsen, Siemens AG, Karlsruhe

► Methoden & Synergien – Modellierung und Simulation

FAIRness in der Automatisierungstechnik am Beispiel Modellierung und Simulation 343

M. Stüber, G. Frey, Universität des Saarlandes, Saarbrücken

Strukturgraph der Produktion – Ein Konzept für Prozessdatenintegration und -analyse mit Anwendungsbeispielen 355

W. Li, M. Winter, C. Roos, T. Kleinert, RWTH Aachen University, Aachen

► Methoden & Synergien – Sustainability

Konzeption eines vereinheitlichenden Energieinformationsmodells zur Integration unterschiedlicher Energieprofile.	371

L.-T. Reiche, A. Fay, Helmut-Schmidt-Universität / Universität der Bundeswehr, Hamburg
M. Runge, K.-H. Niemann, Hochschule Hannover, Hannover

Nachhaltigkeit in der Wasserstoffwirtschaft – Eine Methode zur Bewertung von Wasserstoff-Wertschöpfungsketten.	387

J.-P. Beck, A. Fay, Helmut-Schmidt-Universität, Hamburg

► KI & Autonome Systeme – Machine Learning in der Produktion

Chancen und Herausforderungen für Künstliche Intelligenz in kleinen und mittelständischen Unternehmen	399

D. Vranješ, P. Topalis, O. Niggemann, Helmut-Schmidt-Universität, Hamburg

Bring your Machine Learning Intelligence to the Edge – An Example on Edge Analytics for Cement Strength Prediction	413

M. Platenius-Mohr, J. Schlake, M. Vach, ABB AG Corporate Research Center, Ladenburg;
R. Koorts, ABB Switzerland Ltd, Daettwil, Switzerland

Approach to Data Pre-processing for Predictive Quality of Hydraulic Test Results in a Dynamic Manufacturing Environment	425

C. Neunzig, Bosch Rexroth AG, Ruhr-University Bochum;
D. Möllensiep, S. Fahle, B. Kühlenkötter, Ruhr-University Bochum;
M. Möller, J. Schulz, Bosch Rexroth AG, Homburg

► KI & Autonome Systeme – Machine Learning im Engineering

Methoden der künstlichen Intelligenz für die automatisierte Planung von modularen Produktionsprozessen	439

R. Heesch, N. Widulle, A. Köcher, A. Nordhausen, L. M. Vieira da Silva, J. Putzke, O. Niggemann, Helmut-Schmidt-Universität Hamburg

Auf dem Weg zur Digitalisierung von R&I – Evaluierung von Few Shot Learning Techniken zur einfachen Erkennung bisher unbekannter Elemente	453

M. Gärtler, B. Schmidt, ABB AG, Ladenburg

Digitale Zwillinge als Enabler für Agentensysteme zur Realisierung autonomer Wertschöpfungsnetzwerke mit erhöhter Produktivität	465

E. Bayrhammer, Fraunhofer IFF, Magdeburg;

B. Vogel-Heuser, F. Ocker, Technische Universität AIS, München;

H. Agoustinos, R. Thrift, EXPO21XX GmbH, Hannover

► **KI & Autonome Systeme – Hitze erkennen und behandeln**

Hot Spot Detection in Electrical Equipment using the Inverse Pixel Count Algorithm	479

R. Gitzel, ABB, Ladenburg;

T. Kozel, ABB, Brno, Tschechien

Automatisierung des manuellen Löschvorgangs bei Holzbränden	491

F. Stoller, A. Fay, Helmut-Schmidt-Universität, Hamburg;

F. Kümmerlen, Wehrwissenschaftliches Institut für Schutztechnologien – ABC-Schutz (WIS),
Münster

► **KI & Autonome Systeme – Automatische Extraktion**

Verwendung Neuronaler Netze zur automatisierten Analyse und Modellierung elasto-plastischer Materialeigenschaften von metallischen Zugproben.	505

H. S. Steude, D. Vranješ, J. L. Augustin, O. Niggemann, Helmut-Schmidt-Universität, Hamburg;

M. Lange-Hegemann, Technische Hochschule Ostwestfalen-Lippe;

D. Höche, Helmholtz-Zentrum Geesthacht Hereon, Geesthacht

Automatisierte Modell- und Testgenerierung aus textuellen Anforderungen	521

R. Gröpler, V. Sudhi, L. Kutty, ifak – Institut für Automation und Kommunikation e.V.

Magdeburg

► KI & Autonome Systeme – Datengenerierung

Eine Simulationsumgebung für flexible Cyber-Physische Produktionssysteme zur Generierung realistischer Datensätze für maschinelle Lernverfahren 533

N. Widulle, J. Ehrhardt, M. Krantz, A. Liebert, A. Nordhausen, O. Niggemann,
Institut für Automatisierungstechnik, Helmut-Schmidt Universität, Hamburg

Anwendungsoptimierte Modelladaption des Digitalen Zwillings eines modularen Produktionssystems während der Betriebsphase. 545

D. Dittler, T. Müller, V. Stegmaier, N. Jazdi, M. Weyrich, Universität Stuttgart, Institut für Automatisierungstechnik und Softwaresysteme, Stuttgart

► KI & Autonome Systeme – Sicherheit von KI und autonomen Systemen

Validierung von sicherheitsrelevanten KI-Systemen in der Pharmabranche. 561

A. Charizanis, Universität Stuttgart und Exyte GmbH, Stuttgart;
M. Müller, M. Weyrich, Institut für Automatisierungstechnik und Softwaresysteme, Universität Stuttgart;
H. Mettler, Exyte GmbH, Stuttgart

Strategisches Drohnenverkehrsmanagement: Konzept für eine automatisierte Fluggenehmigung und -freigabe von Drohnenflügen. 581

T. Grebner, A. Fay, Helmut-Schmidt-Universität, Hamburg

Erweiterte Risikobeurteilung für Maschinen und Betriebsmittel mit künstlicher Intelligenz 595

W. Varro, TÜV SÜD Product Service GmbH, München

► KI & Autonome Systeme – Batch Phasen Erkennung

Hidden Markov Models und Active Learning zur automatisierten Kennzeichnung von Batchphasen in der Prozessindustrie. 615

G. Just, V. Khaydarov, L. Urbas, Professur für Prozessleittechnik/AG Systemverfahrenstechnik, Technische Universität Dresden;
B. Klöpper, ABB;
F. D. Böhner, Bayer AG

Phasenidentifikation von Batch-Prozessen durch aktives Lernen 625

C. Song, M. Gärtler, B. Klöpper, ABB Corporate Research Center, Ladenburg

► Industrielle Kommunikation – Drahtlose Kommunikation

Performance Evaluation of a Campus Network emphasizing the ICT-OT-relation 637

L. Underberg, G. Cainelli, Institut für Automation und Kommunikation e.V. Magdeburg;
T. Neugebauer, Götting KG, Lehrte/Röddensen

► Industrielle Kommunikation – OPC UA in der Anwendung

Konfigurierbare Controller-Controller-Kommunikation für modulare Produktionssysteme. . . 651

A. Stutz, T. Kempin, M. Maurmaier, Siemens AG, Karlsruhe;
M. Blumenstein, A. Fay, Helmut-Schmidt-Universität, Hamburg

Gestaltung einer Forschungsinfrastruktur für die Anwendung digitaler Werkzeuge in Cyber-Physischen Produktionssystemen 667

I. Gräßler, M. Hieb, D. Roesmann, Heinz-Nixdorf-Institut, Lehrstuhl für Produktentstehung,
Universität Paderborn

Sichere interoperable Maschine-zu-Maschine- Kommunikation basierend auf OPC UA Field Exchange 679

S. Höme, T. Fischer, G. Biehler, S. Kerschbaum, Siemens AG, Nürnberg

► Digitale Geschäftsmodelle – Umsetzungsansätze für digitale Geschäftsmodelle

Stärkung der Cyber-Resilienz in KMU. 691

C. Siegart, F. Scherhag, ZeMA gGmbH, Saarbrücken;
K. Pervözl, Fraunhofer IAIS, Sankt Augustin,
G. Frey, Universität des Saarlandes, Saarbrücken

► **Gebäudeautomation**

Prognostizierte Energieverbräuche durch Deep Learning als Basis zur Integration in ein Smart Grid am Beispiel des Flughafen Köln Bonn 703

M. Stange, M. Both, J. Müller, Institut für Technische Gebäudeausrüstung,
Technische Hochschule Köln

Verbesserte Workflow-Prozesse in der Gebäudetechnik und Gebäudeautomation im Kontext der Transformationspfade Energiewende und Digitalisierung 715

P. Knoll, M. Becker, Hochschule Biberach (HBC), Biberach

Prototypische Implementierung von Capability- und Feasibility-Checks in I4.0-Verwaltungsschalen der technischen Gebäudeausrüstung 729

N. Maisch, R. Ulmer, J. Müller, Institut für Technische Gebäudeausrüstung, Technische Hochschule Köln;
S. Schmitt, D. Uecker, F. Schnicke, M. Conradi, Fraunhofer IESE, Kaiserslautern

**VDI NACHRICHTEN RECRUITING TAG –
DEUTSCHLANDS FÜHRENDE
KARRIEREMESSE FÜR INGENIEURE.****Ideal für Ihre erfolgreiche Jobsuche:**

- Renommierete Unternehmen
- Direkter Kontakt mit Entscheidern
- Karriereberatung und -vorträge
- Job Board

Wir machen Ingenieurkarrieren. Vor Ort. Und Online.

VDI nachrichten Recruiting Tag und VDI nachrichten Recruiting Tag Online. Deutschlands führende Karrieremessen für Ingenieure und IT-Ingenieure.

Für alle Studierenden der Ingenieurwissenschaften, Absolventen und Young Professionals ein absolutes Muss. Knüpfen Sie Kontakte zu renommierten Unternehmen und sprechen Sie direkt mit Entscheidern aus den Fachabteilungen. Viele Serviceangebote wie Karriereberatung und -vorträge unterstützen Sie bei Ihrem erfolgreichen Einstieg ins Berufsleben.

**Jetzt informieren und kostenfrei teilnehmen: www.ingenieur.de/recruitingtag****Mehr Informationen?**

Silvia Becker, Telefon: +49 211 6188-170
Franziska Opitz, Telefon: +49 211 6188-377
recruiting@vdi-nachrichten.com

VDI nachrichten recruiting tag

Vorwort

Erschreckende Studien zur Erderwärmung und mit dem Klimawandel verbundene dramatische Brände und Hochwasserereignisse sensibilisieren unsere Gesellschaft einmal mehr für die Nachhaltigkeit unseres Handelns. Ein ausgeprägtes Systemdenken ist der Fachcommunity der Automation vertraut. Wie können wir die großen Systeme so gestalten, dass diese im Sinne der Nachhaltigkeit ein bestimmtes Maß an Ressourcennutzung dauerhaft aushalten, ohne Schaden zu nehmen? Wie gelingt eine Erhaltung der Ressourcen trotz deren Nutzung?

Die Zeichen der Zeit, wirtschaftlichen Erfolg und Nachhaltigkeit zu verbinden, sind erkannt. Über das bloße Erkennen hinaus, weist das Motto „Automation creates Sustainability“ auf die Gestaltungskraft der Automatisierungstechnik hin. Automation als Wegbereiter für nachhaltiges Wirtschaften, das nicht als Verzicht, sondern als Fortschritt und Innovation verstanden wird. Die Folgen von Ressourcenverknappung und die Nachhaltigkeitsforderung lassen sich „digital“ abfedern. Methoden der Automation sind ein Schlüssel für das Überwinden von Hemmnissen für Wachstum und Wohlstand. In Zeiten großer gesellschaftlicher Herausforderungen sind Austausch und gemeinsames Nachdenken besonders gefragt.

Das jährliche Networking-Event des AUTOMATIONKongresses bietet dafür eine ideale Plattform. In bewährter Weise umfasst das Kongressprogramm Fachbeiträge zu Methoden und Konzepten, zu Technologie und Engineering sowie den vielfältigen interdisziplinären Anwendungen.

Wir freuen uns, Sie auf der AUTOMATION 2022 willkommen zu heißen.

Die Kongressleiter



Prof. Dr.-Ing. Ulrich Jumar,
Vorstandsvorsitzender und Institutsleiter,
ifak – Institut für Automation und
Kommunikation e. V., Magdeburg



Dr.-Ing. Felix Hanisch,
Head of Process and Plant Safety,
Bayer AG, Leverkusen

Programmausschuss

Dipl.-Ing. Heiko Adamczyk, Senior Expert Industrial Security, TÜV SÜD Product Services GmbH, München

Schirin Baer, Area Manager, Amazon, Frankenthal

Prof. Dr.-Ing. Mike Barth, IOS³ – Institut für Smart Systems und Services, Hochschule Pforzheim, Chefredakteur „science“ atp magazin

Prof. Dr.-Ing. Martin Becker, Institut für Gebäude- und Energiesysteme, Hochschule Biberach

Prof. Dr. Joachim Birk, Vice President, Executive Expert of Automation Technology, BASF SE, Ludwigshafen

Dr.-Ing. Michael Deilmann, Managing Director, KROHNE Messtechnik GmbH, Duisburg

Sascha Dessel, M. Sc., Geschäftsführer VDI/VDE-GMA, VDI e. V., Düsseldorf

Prof. Dr.-Ing. Christian Diedrich, Lehrstuhl für Automatisierungstechnik (IFAT), Otto-von-Guericke-Universität, Magdeburg

Dr.-Ing. Dagmar Dirzus, Vice President AI & Plattform Business, KROHNE Innovation GmbH, Duisburg

Univ.-Prof. Dr.-Ing. Alexander Fay, Leiter des Instituts für Automatisierungstechnik, Helmut-Schmidt-Universität/Universität der Bundeswehr, Hamburg

Univ.-Prof. Dr.-Ing. Georg Frey, Lehrstuhl für Automatisierungs- und Energiesysteme, Universität des Saarlandes, Saarbrücken

Hans Joachim Fröhlich, Director Technology and Portfolio, Endress+Hauser Group Services AG, Reinach, Schweiz

Dr.-Ing. Stefan Gehlen, Geschäftsführer, VMT Vision Machine Technic Bildverarbeitungssysteme GmbH, Mannheim

Dr. Martin Gerlach, Head of Process Analytical Technologies, Engineering & Technology, Bayer AG, Leverkusen

Prof. Dr.-Ing. Iris Gräßler, Fachgruppe Produktentstehung, Universität Paderborn

Christian Gülpen, Bereichsleiter Digitalisierung am Institut für Technologie und Innovationsmanagement, RWTH Aachen

Dipl.-Ing. Axel Haller, Global Segment Manager, ABB AG, Mannheim

Prof. Dr.-Ing. Michael Heizmann, Institutsleiter, Institut für Industrielle Informationstechnik – IIIT, Karlsruher Institut für Technologie (KIT)

Dipl.-Ing. Ulrich Hennen, Head of Market Management Industry & Process, WAGO Kontakttechnik GmbH & Co. KG, Minden

Dipl.-Ing. Tim Henrichs, Director Marketing Europe, Yokogawa Deutschland GmbH, Ratingen

Prof. Dr.-Ing. i. R. Hartmut Hensel, ehemals Fachbereich Automatisierung und Informatik, Hochschule Harz, Wernigerode

Prof. Dr.-Ing. Jörg Kiesbauer, Leiter des Instituts für Mechanik und Mechatronische Systeme (im2s), Hochschule Darmstadt

Dr.-Ing. Niels Kiupel, Vice President Smart Operations, Evonik Operations GmbH, Marl

Prof. Dr.-Ing. Tobias Kleinert, Lehrstuhl für Prozessleittechnik, RWTH Aachen

Prof. Dr. Steffi Knorn, Prozess- und Verfahrenstechnik, Fachgebiet Mess- und Regelungstechnik, Technische Universität Berlin

Gunther Koschnick, Geschäftsführer Fachverband Automation, ZVEI e. V., Frankfurt/Main

Univ.-Prof. Dr.-Ing. Bernd Kuhlentötter, Lehrstuhl für Produktionssysteme (LPS), Fakultät für Maschinenbau, Ruhr-Universität Bochum

Dr. Felix Loske, Head of Corporate Technology Development, HARTING Stiftung & Co. KG, Espelkamp

Dr. Christine Maul, Head of Advanced Process Control, Global Process Control Technology, Covestro Deutschland AG, Leverkusen

Dipl.-Ing. Martin Müller, Leiter Business Unit I/O and Networks, Phoenix Contact Electronics GmbH, Bad Pyrmont

Dr.-Ing. Jörn Oprzynski, Leiter Portfolio & Innovation, Siemens AG Digital Industries, Karlsruhe

Dr. Thomas Paulus, Global Executive Officer – CDO, Digital Transformation, KSB SE & Co. KGaA, Frankenthal

Prof. Dr. rer. Pol. Frank T. Piller, Institutsleiter, Institut für Technologie- und Innovationsmanagement, RWTH Aachen

Dr. Thorsten Pötter, Chief Digital Officer, SAMSON AG, Frankfurt am Main

Dr.-Ing. Eckhard Roos, ehem. Leiter Prozessautomation, Festo AG & Co. KG, Esslingen/Neckar

Prof. Dr.-Ing. Martin Ruskowski, Vorstandsvorsitzender der Technologie-Initiative Smart-Factory KL e. V., Kaiserslautern

Dipl.-Kfm. Felix Seibl, Geschäftsführer, ZVEI-FB Messtechnik und Prozessautomatisierung, ZVEI e. V., Frankfurt/Main

Prof. Dr.-Ing. habil. Olaf Simanski, Fachgebiet Automatisierungstechnik, Hochschule Wismar

Dr.-Ing. Lisa Underberg, Leiterin der AG Validation und Test der 5G-ACIA, Forscherin Funk in der Automation, ifak - Institut für Automation und Kommunikation e. V. Magdeburg

Prof. Dr.-Ing. Leon Urbas, Professur für Prozessleittechnik und AG Systemverfahrenstechnik, Technische Universität Dresden

Nils Weber, Geschäftsführer NAMUR e. V., Leverkusen

Prof. Dr.-Ing. Dr. h.c. Michael Weyrich, Institutsleitung, Institut für Automatisierungstechnik und Software-systeme, Universität Stuttgart

SPONSOREN

Wir danken unseren Sponsoren für die freundliche Unterstützung.

GOLDSPONSOREN:



BONZESPONSOR:



SPONSOR:



Veranstalter:
VDI Wissensforum GmbH

Anwendungsfälle und Ansatz zur Erstellung des Digitalen Zwillings aus Sicht eines Komponentenherstellers

V. Stegmaier, M.Sc., Graduate School of advanced Manufacturing Engineering, Universität Stuttgart;
Dr.-Ing **W. Schaaf**, J. Schmalz GmbH, Glatten;
Dr.-Ing **N. Jazdi**, Prof. Dr.-Ing. Dr. h. c. **M. Weyrich**,
Institut für Automatisierungstechnik und Softwaresysteme,
Universität Stuttgart

Kurzfassung

Neben Herausforderungen bei der Erstellung bieten sich Komponentenherstellern durch den Digitalen Zwilling viele potenzielle Mehrwerte. Auf die Anwendungsfälle, um die entsprechenden Mehrwerte zu realisieren, soll im Folgenden näher eingegangen werden. Für eine effiziente Erstellung aus Sicht des Komponentenherstellers wird ein Konzept präsentiert. Dieses nutzt bestehende Informationen und erstellt häufig fehlende Verhaltensmodelle automatisiert. Zur Veranschaulichung potenzieller Mehrwerte der Verhaltensmodelle wird der Luftverbrauch eines Vakuum-Greifsystems beispielhaft analysiert.

Abstract

In addition to challenges in creation, the Digital Twin also offers component manufacturers many potential added values. The use cases for realizing the corresponding added values will be discussed in more detail below. A concept is presented for efficient creation from the component manufacturer's point of view. This uses existing information and creates frequently missing behavior models in an automated way. To illustrate potential added values of the behavior models, the air consumption of a vacuum gripping system is analyzed as an example.

1. Einleitung

Die Mehrwerte des Digitalen Zwillings in der produzierenden Industrie sind hinreichend bekannt [1, 2]. Dieser kann in den unterschiedlichsten Anwendungsfällen wie der virtuellen Produktentwicklung, der virtuellen Inbetriebnahme oder der vorausschauenden Wartung eingesetzt werden [3]. Die daraus folgende große Bedeutung bringt dem Digitalen Zwilling große Aufmerksamkeit in der Literatur. Dies führt zu einem stark variierenden Verständnis darüber, was genau ein Digitaler Zwilling ist und welche Eigenschaften er aufweist. Für diesen Beitrag soll die Definition nach Ashtari et al. [4] verwendet werden, wonach ein Digitaler Zwilling die

digitale Repräsentation eines physischen Objektes (Assets) ist, welches möglichst realitätsnahe Modelle und alle verfügbaren Daten des Assets einschließt. Die enthaltenen Daten müssen dabei neben organisatorischen und technischen Informationen auch Betriebsdaten umfassen. Zudem muss der Digitale Zwilling immer mit dem physischen Asset synchronisiert sein und die Simulation des Verhaltens des Assets ermöglichen [4]. Neben den zahlreichen Vorteilen eines Digitalen Zwillings, gibt es auch damit verbundene Herausforderungen. Eine dieser ist die effiziente Erstellung Digitaler Zwillinge, sodass der dabei entstehende Aufwand den späteren Nutzen nicht übersteigt [5].

Betrachtet man die Assets welche durch Digitale Zwillinge beschrieben werden, verbleiben diese selten bei dem Unternehmen, von dem sie entwickelt und produziert werden. Viel häufiger überschreiten die Assets Unternehmensgrenzen. Die daraus resultierende unternehmensübergreifende Zusammenarbeit wird auch als industrielle Wertschöpfungskette bezeichnet. In dieser kommen häufig die drei Hauptgruppen Komponentenhersteller, Maschinen- und Anlagenbauer und produzierende Unternehmen vor [6, 7]. Die Erstellung Digitaler Zwillinge wird dabei überproportional stark von Komponentenherstellern ausgeführt, da diese häufig früh in der industriellen Wertschöpfungskette aktiv sind [8, 9]. Aus deren Sicht ergeben sich nicht nur Mehraufwände durch die Erstellung des Digitalen Zwillings, sondern auch eine Vielzahl von Möglichkeiten zu Effizienzsteigerungen und neuen Geschäftsmodellen.

Auf die angesprochenen Anwendungsfälle und Herausforderungen durch den Digitalen Zwilling aus Sicht eines Komponentenherstellers (Kapitel 2), sowie Ansätze zur Erstellung Digitaler Zwillinge (Kapitel 3) und ein Anwendungsbeispiel (Kapitel 4) wird in diesem Beitrag eingegangen. Abschließend wird zusammengefasst und ein Ausblick gegeben (Kapitel 5).

2. Anwendungsfälle und Herausforderungen durch den Einsatz Digitaler Zwillinge bei Komponentenherstellern

Bild 1 zeigt links die Unternehmensgruppen der industriellen Wertschöpfungskette [7, 10] eingeordnet über der klassischen Automatisierungspyramide. Darin ist zu erkennen, dass produzierende Unternehmen und Maschinenbauer ihre Kernkompetenzen hauptsächlich in den mittleren und oberen Ebenen der Automatisierungspyramide besitzen. Komponentenhersteller hingegen sind vor allem in der Steuergeräteebene und darunter, bis zum Produkt bzw. dem Prozess aktiv. Im rechten Bereich von Bild 1 sind unterschiedliche Gruppen von Simulationen, sowie die allgemeine Vielfalt an verwendeten physikalischen Grundprinzipien und die relevante Modellierungstiefe aufgetragen. Die unterschiedlichen Ansätze zur Simulation für verschiedene Anwendungsfälle lässt sich in vier Gruppen unterteilen. Simulationen von Geschäftsprozessen stellt dabei die Gruppe mit der größten Abstraktion und Vielzahl modellierter

Prozesse dar. Bei der Betrachtung ganzer Geschäftsprozesse werden einzelne Komponenten und sogar ganze Maschinen zu größeren Systemen abstrahiert. Der Fokus liegt auf einer makroskopischen Analyse ganzer Unternehmen. Die nächst detailliertere Gruppe ist jene der Anlagen- und Prozesssimulationen. Auch diese betrachten viele Komponenten noch sehr abstrahiert, der Fokus liegt jedoch nicht mehr auf ganzen Unternehmen, sondern auf Produktionsanlagen oder deren Verbund zu Produktionsprozessen. Die darin beinhalteten Maschinen sind Fokus der dritten Gruppe von Simulationen. Ein bekannter Vertreter dieser Gruppe ist die virtuelle Inbetriebnahme (VIBN). Allgemein fokussiert sich die Gruppe der Maschinensimulationen zwar auf das Zusammenspiel einzelner Komponenten hin zur Funktionsweise einer Maschine, die detaillierte Funktionalität der einzelnen Komponenten ist jedoch nicht Kernaspekt dieser Gruppe. Dies ist vielmehr der Fall für technisch-physikalische Simulationen die sich auf einzelne Komponenten oder gar einzelne Funktionsprinzipien fokussieren. Dabei ist die dementsprechend abgebildete Vielfalt an Funktionen oder Prozessen einer solchen Simulation stark eingeschränkt. Ein Beispiel hierfür ist die Strömungssimulation der Strahldüse eines pneumatischen Vakuumejektors. Diese betrachtet das technisch-physikalische Verhalten der Düse, bildet jedoch keinerlei weitere Funktionalität des Ejektors ab. [11–13]

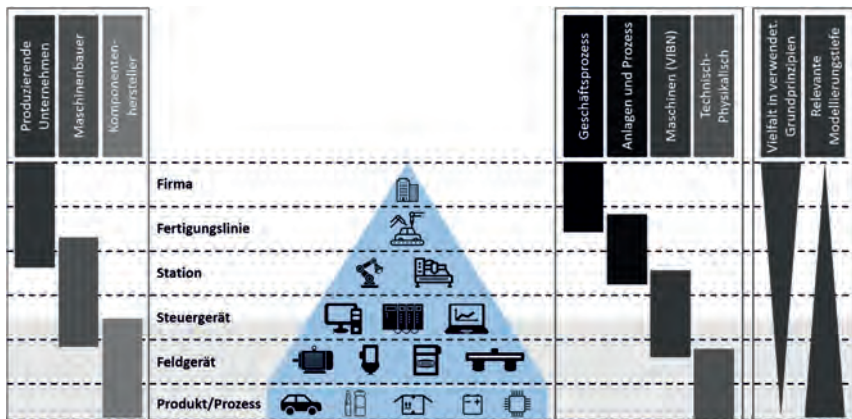


Bild 1: Einordnung von Unternehmensgruppe der industriellen Wertschöpfungskette anhand der klassischen Automatisierungspyramide

Der rechte Kasten in Bild 1 hebt die soeben beschriebene Einordnung hervor. Darin ist die Vielfalt verwendeter Grundprinzipien über die Automatisierungspyramide aufgetragen. Dieser nach unten abnehmende Balken veranschaulicht die Quantität unterschiedlicher Funktionali-

täten der einzelnen Ebenen. Betrachtet man ein gesamtes Produktionsunternehmen mit diversen Produktionslinien, werden in den unterschiedlichsten Anlagen und Maschinen die unterschiedlichsten Grundprinzipien eingesetzt. Beispiel hierfür ist ein Unternehmen oder eine Produktionslinie, in denen Roboter mit Schweißzangen, Roboter mit Greifsystemen, CNC-Bearbeitungszentren und Lackierstationen eingesetzt werden. Es werden also unterschiedlichste Komponenten und Prozesse mit sehr vielfältigen Funktionsweisen verwendet. Zum Vergleich kommt bei der Betrachtung einer einzelnen Komponente oder eines Prozesses auf der Feldgeräte- oder Produkt-/Prozessebene nur eine deutlich geringere Anzahl an Grundprinzipien zum Einsatz. Der Vielfalt an verwendeten Grundprinzipien steht die relevante Modellierungstiefe gegenüber. Auf den oberen Ebenen der Automatisierungspyramide wird häufig eine eher makroskopische Betrachtung der Assets durchgeführt. Die entsprechende Modellierungstiefe der einzelnen Betrachtungsgegenstände wie Fertigungslinien oder Anlagen ist gering. Am unteren Ende der Pyramide hingegen ist die Quantität der betrachteten Grundprinzipien stark verringert, dafür werden jedoch sehr detaillierte Betrachtungen einzelner Vorgänge durchgeführt. Die entsprechende Modellierungstiefe ist sehr hoch. Die Überdeckung der Einordnung des Komponentenherstellers mit den allgemeinen Kriterien zur Einordnung von Modellen wird durch einen grauen Kasten in Bild 1 veranschaulicht. Dieser hebt hervor, dass der Fokus für Komponentenhersteller auf technisch-physikalischen und Maschinensimulationen mit einer geringen Anzahl an verwendeten Grundprinzipien in großer Modellierungstiefe liegt. Eine alternative Betrachtungsweise dieser Sachverhalte kann auf Basis des Maschinen- und Anlagenlebenszyklus durchgeführt werden. Bild 2 stellt einen solchen beispielhaft dar [14–16]. Über die 7 Phasen des dargestellten Anlagenlebenszyklus sind zum einen klassische Industrie-4.0-Anwendungsfälle aus Sicht eines Komponentenherstellers dargestellt. Zum anderen präsentiert Bild 2 die Einordnung von Datenumfang und benötigter Modellierungstiefe der verwendeten Modelle. Dabei stellt die Einordnung der einzelnen Anwendungsfälle keine absolute und unverrückbare Klassifizierung dar. Vielmehr wird damit eine grobe Verortung in die entsprechende Phase visualisiert, in der ein entsprechender Anwendungsfall Mehrwerte stiften kann. [17, 18]

Anwendungsfälle bestehen, wie in Bild 2 dargestellt, nicht nur in den frühen Phasen des Anlagenlebenszyklus, in denen Komponentenhersteller heutzutage häufig aktiv sind. Viele Anwendungsfälle sind auch den späteren Phasen vorzufinden. Beispielsweise können durch Optimierungen in der Betriebsphase oder präventive Wartung Mehrwerte für den Komponentenhersteller sowie dessen Kunden mithilfe des Digitalen Zwillings geschaffen werden. Dabei nimmt der Umfang an Daten mit steigender Lebensdauer eines Assets immer weiter zu. Bestehen die Daten und Modelle zu Beginn noch hauptsächlich aus Engineering-Informationen

wie CAD-Modellen, Elektroschaltplänen oder Betriebsanleitungen, werden diese im Laufe der Zeit vor allem durch Betriebs- und Umgebungsdaten angereichert. Anders verhält es sich mit der benötigten Modellierungstiefe der verwendeten Modelle im Digitalen Zwilling. Ist diese zu frühen Phasen noch eher gering, werden mit fortlaufender Konzeptphase und vor allem in der Entwicklungsphase größere Modellierungstiefen benötigt. Anschließend geht die benötigte Modellierungstiefe wieder zurück. Ab der Betriebsphase wird eine annähernd gleichbleibende Modellierungstiefe benötigt. Dies kann unter anderem durch die verfügbaren Rechenkapazitäten in den entsprechenden Phasen des Anlagenlebenszyklus erklärt werden. Modelle mit einer größeren Modellierungstiefe sind rechenintensiver. Vor allem in der Entwicklungsphase sind ausreichend Rechenkapazitäten vorhanden. Später im Anlagenlebenszyklus nimmt die verfügbare Rechenkapazität für die einzelnen Modelle durch die Zunahme an betrachteten Komponenten jedoch ab, was häufig durch eine geringere Modellierungstiefe kompensiert wird.

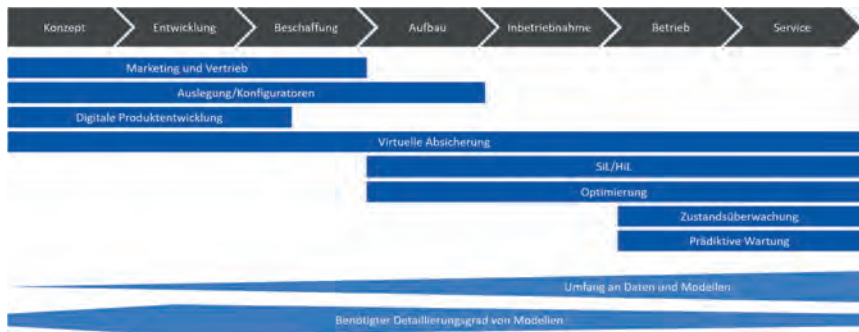


Bild 2: Einordnung von Industrie-4.0-Anwendungsfällen und Eigenschaften von Daten und Modellen über die Phasen des Anlagenlebenszyklus angelehnt an [14–16]

Zusammenfassend lässt sich festhalten, dass Anwendungsfälle aus Sicht eines Komponentenherstellers entlang des gesamten Anlagenlebenszyklus bestehen. Die daraus erzielbaren Mehrwerte verteilen sich somit auch über alle Phasen. Dabei gilt es vor allem Daten und Modelle entlang des gesamten Anlagenlebenszyklus und auch über Unternehmensgrenzen hinweg konsistent zu erfassen und zur Verfügung zu stellen. Zudem ist es notwendig Verhaltensmodelle unter Berücksichtigung der zur Verfügung stehenden Ressourcen in entsprechender Modellierungstiefe bereitzustellen.

3. Ansatz zur Erstellung Digitaler Zwillinge aus Sicht eines Komponentenherstellers

Der Ansatz zur effizienten Erstellung Digitaler Zwillinge aus Sicht eines Komponentenherstellers soll zuerst makroskopisch in der industriellen Wertschöpfungskette verortet werden. Dies wird durch Bild 3 veranschaulicht. Darin sind die bereits genannten drei Gruppen der industriellen Wertschöpfungskette dargestellt. Zudem ist noch eine zusätzliche Unternehmenskategorie aufgeführt, der Zulieferant. Komponentenhersteller sind zwar sehr früh in der industriellen Wertschöpfungskette aktiv, entwickeln und produzieren jedoch nicht alle Aspekte ihrer Komponenten komplett selbst. Häufig greifen sie auf bereits bestehende Komponenten wie z.B. Elektronikbaugruppen oder Schrauben zurück. Aus diesem Grund kann die in Bild 1 eingeführte Gruppe der Komponentenhersteller exemplarisch in die zwei in Bild 2 dargestellten Gruppen unterteilt werden. In den Unternehmensgruppen ist jeweils ein Enterprise Service Bus (ESB) dargestellt. Dieser repräsentiert die unternehmensspezifische IT-Infrastruktur der entsprechenden Unternehmen, da diese häufig auf einem ESB aufbauen. Alternativ können auch IT-Systeme mit proprietären Schnittstellen eingesetzt werden. Beispiele für die in diesem Zusammenhang wichtigen IT-Systeme sind in der Box des Komponentenherstellers eingezeichnet. Product Lifecycle Management (PLM), Product Information System (PIM) oder das Enterprise Resource Planning System (ERP) sind Systeme die viele relevante Informationen für den Digitalen Zwilling enthalten. Wichtig für die konsistente Verwendung von Daten und Modellen im Digitalen Zwilling entlang der Wertschöpfungskette ist ein einfacher Austausch von Daten und Modellen. Dies ist in Bild 3 durch die Schnittstelle für externe Kommunikation berücksichtigt. Als Standard für eine solche Kommunikation eignet sich beispielsweise die Verwaltungsschale (VWS) da diese in der Industrie bereits weit verbreitet ist [5, 19]. Beispielhafte Informationen, relevant für einen Austausch zwischen den Unternehmen, sind entlang der grauen Pfeile im oberen Bereich der Abbildung dargestellt. Austauschplattform für diese Informationen kann beispielsweise ein Digitaler Zwilling in einer Cloud oder auch private Cloud mit entsprechenden Zugriffsrechten und Sicherheitsvorkehrungen sein.

Über einen solchen Aufbau kann das geplante Assistenzsystem, dargestellt für den Komponentenhersteller, die benötigten Informationen zur Erstellung des Digitalen Zwillings aus internen und externen Quellen beziehen. Auf Basis dieser Informationen kann es den unterschiedlichen Stakeholdern Digitale Zwillinge zur Verfügung stellen. Die Informationen reichen dabei beispielsweise von Qualitäts- und Herstellungsdaten über Updates und Betriebsdaten hin zur Anforderung eines Digitalen Zwillings in der entsprechenden Modellierungstiefe und dessen Bereitstellung.

Bild 4 stellt eine Detaillierung des Ansatzes aus Bild 3 mit Fokus auf dem Assistenzsystem dar. Darin ist die bereits erläuterte Schnittstelle für eine externe Kommunikation sowie die unternehmensspezifische IT-Infrastruktur dargestellt. Zentrale Eingangsinformationen für das Assistenzsystem sind Anforderungen aus internen oder externen Quellen. Diese Anforderungen werden vorverarbeitet und in ein entsprechend einheitliches Format überführt. Diese vereinheitlichten Anforderungen sind die Eingangsgrößen für die zwei zentralen Bestandteile des Assistenzsystems.

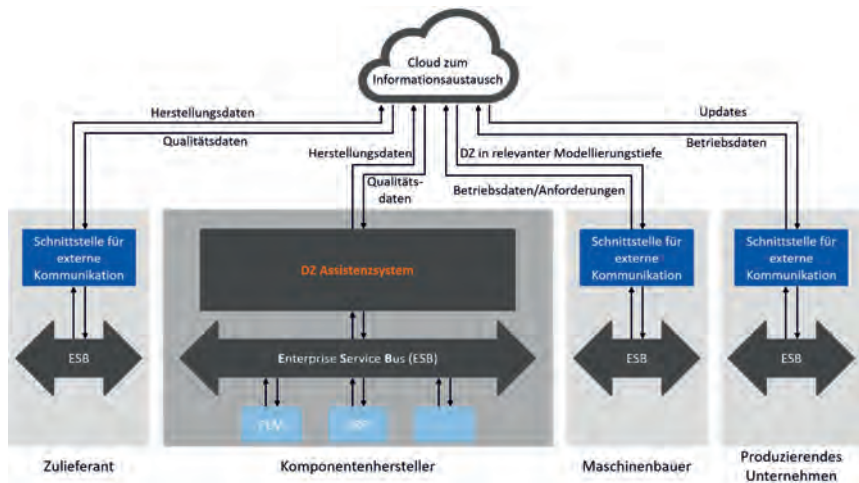


Bild 3: Einordnung des Assistenzsystem zur effizienten Erstellung Digitaler Zwillinge in die industrielle Wertschöpfungskette

Zum einen gibt es Daten und Modelle, die bereits im Unternehmen vorliegen. Diese müssen semantisch aufbereitet werden um im Anschluss eine automatisierte Weiterverarbeitung gewährleisten zu können. Dazu werden die Daten aus den entsprechenden IT-Systemen über den ESB angezogen und mithilfe einer semantischen Übersetzungsbasis aufbereitet. Diese Übersetzungsbasis muss vorher mithilfe von Informationen aus den entsprechenden IT-Systemen assistiert erstellt werden und ist unternehmensspezifisch. Ein möglicher und industriell weit verbreiteter Standard für eine solche Übersetzungsbasis ist ECLASS [20].

Zum anderen existieren Verhaltensmodelle bei Komponentenherstellern häufig nicht in allen prozessrelevanten Modellierungstiefen. Die Verhaltensmodelle sollen daher auf Basis der Anforderungen mithilfe einer Bibliothek für Verhaltensmodelle und einer Datenbasis zur Paramet-

rierung erstellt werden. Die Verhaltensmodellbibliothek muss hierfür im Voraus aufgebaut werden. Die Datenbasis hingegen kann mithilfe von Versuchsdaten oder Ergebnissen aus tiefgreifenden domänenspezifischen Simulationen befüllt werden, die häufig im Unternehmen schon vorhanden sind. Dazu müssen die Daten entsprechend vorverarbeitet werden. Nach diesen beiden Schritten werden die Informationen zu einer Struktur zusammengefügt, die als Digitaler Zwilling bezeichnet werden kann. Wie bereits erwähnt bieten sich hierfür verbreitete Standards an. Aus genannten Gründen soll bei späteren beispielhaften Implementierungen auf die VWS zurückgegriffen werden. Die somit erstellten Digitalen Zwillinge können anschließend in den entsprechenden IT-Systemen für interne Anwendungsfälle verwendet werden. Alternativ dazu steht die Übermittlung an andere Unternehmen. Den soeben vorgestellten Funktionalitäten sind in Bild 4 Symbole zugewiesen, durch die dargestellt wird, ob es sich dabei um eine assistierte Funktionalität, eine voll automatisierte Funktionalität oder einen Daten- oder Modellspeicher handelt. In der rechts dargestellten Legende sind die einzelnen Funktionalitäten erläutert.

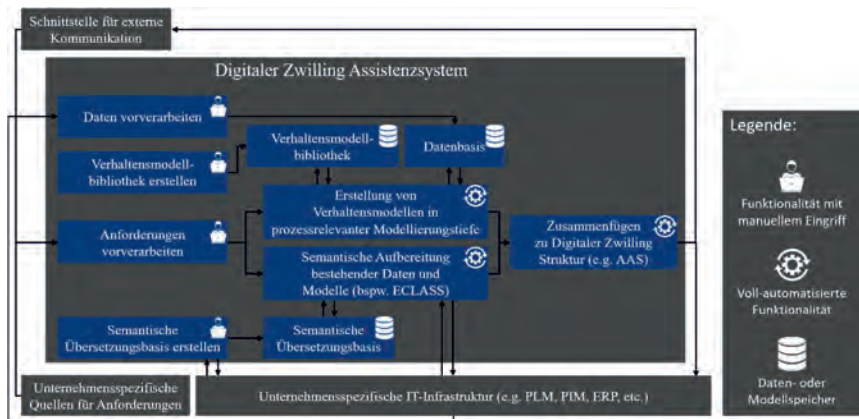


Bild 4: Struktur des Assistenzsystems

Der Fokus im vorgestellten Konzept liegt auf der automatisierten Erstellung und Parametrierung von Verhaltensmodellen. Um die Mehrwerte dieser Verhaltensmodelle zu verdeutlichen, wird im folgenden Abschnitt ein Beispiel aus der Vakuum-Handhabungstechnik präsentiert.

4. Anwendungsbeispiel

Vakuum-Greifsysteme werden wegen ihrer hohen Robustheit und einfachen Inbetriebnahme vielfältig in Produktionssystemen und für automatisierte Handhabungsaufgaben eingesetzt [21]. Ein zentraler Vorteil ist dabei deren Fähigkeit das Handhabungsobjekt von nur einer Seite

zu greifen, was eine hohe Flexibilität zur Folge hat [22]. Ein beispielhafter Anwendungsfall eines solchen Greifsystems ist der automobiler Karosseriebau, wie links in Bild 5 dargestellt. Abgeleitet aus diesem Anwendungsfall soll für die nachfolgenden Versuche ein konfigurierbarer Ersatzaufbau, dargestellt in der Mitte von Bild 5, verwendet werden. Der Fokus der nachfolgenden Analyse liegt auf der fluidischen Disziplin, da diese die meisten Key-Performance-Indicators (KPIs) des Vakuum-Greifsystems maßgeblich beeinflussen. Fluidisch besteht dieses Greifsystem aus vier Sauggreifern, Schläuchen und Verbindern, einem Drucksensor und einem pneumatischen Vakuum Ejektor. Das entsprechende Verhaltensmodell wird basierend auf dem Ansatz aus [23] automatisch erstellt und kann anschließend in MATLAB Simscape ausgeführt werden. Das resultierende Modell ist rechts in Bild 5 dargestellt.

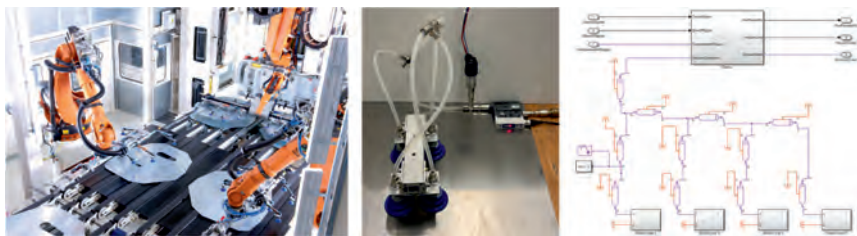


Bild 5: Beispielhafter Anwendungsfall eines Vakuum-Greifsystems (links) [24], konfigurierbarer Ersatzaufbau (Mitte) und Verhaltensmodell (rechts)

Ein Parameter, der das Verhalten eines Vakuum-Greifsystems stark beeinflusst ist die Leckage, also der unerwünschte Eintrag von Fremdluft ins System während des Handhabungsprozesses. Diese Leckage kann beispielsweise durch unterschiedlich beschaffenen Handhabungsobjekte und deren unterschiedliche Oberflächenrauigkeiten hervorgerufen werden. Das vorgestellte Vakuum-Greifsystem wird auf zwei unterschiedlichen Oberflächen vermessen. Zum einen einem Stahlblech, welches nahezu keine Leckage verursacht („saugdicht“) und zum anderen einem Holzbrett, welches für eine gut feststellbare Leckage sorgt („leckagebehaftet“). Der Leckage-Faktor für das Holzbrett wird geschätzt, da hierfür aktuell noch keine funktionalen Zusammenhänge bekannt sind. Wichtiger Schritt bei der ersten Verwendung eines Verhaltensmodells aus Bausteinen, die bisher noch nicht validiert sind, ist der Abgleich mit dem realen Verhalten des Systems. Hierzu wird das vorgestellte Vakuum-Greifsystem auf den beschriebenen zwei Oberflächen mit einer Zykluslänge von 5 Sekunden vermessen. Die obere Regelschwelle der Zweipunktregelung [22] ist dabei auf 750 mbar,rel eingestellt und

die untere auf 600 mbar,rel. Die Messung wird mithilfe einer speicherprogrammierbaren Steuerung (SPS) Fabrikat Beckhoff CX9020 und einer Zykluszeit von 1 Millisekunde (ms) durchgeführt.

Um die Streuung der Messungen abschätzen zu können, wird die Messung 10-mal wiederholt. Der Vergleich aus Messung und Simulation auf dem Stahlblech ist dabei links in Bild 6 dargestellt, jener für das Holzbrett rechts.

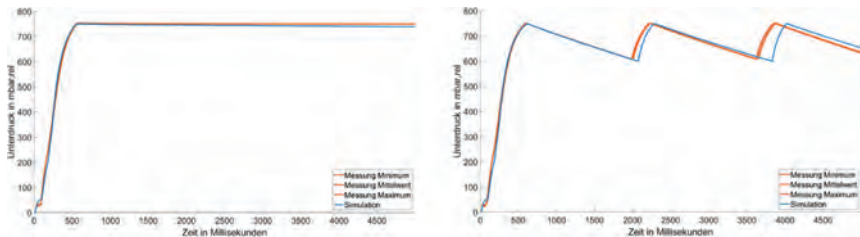


Bild 6: Abgleich zwischen Messung und Simulation für den Anwendungsfall „saugdicht“ (links) und „leakagebehaftet“ (rechts)

Dabei lässt sich zunächst gut erkennen, dass die Streuung der Messung auf beiden Oberflächen sehr gering und visuell in den Diagrammen kaum wahrnehmbar ist. Zudem ist die Übereinstimmung zwischen Messung und Simulation vor allem im linken Grafen sehr gut. Es lässt sich kaum eine Abweichung feststellen. Selbiges gilt auch für den ersten Bereich, bis ungefähr 2000 ms im rechten Grafen. Anschließend driften Messung und Simulation leicht auseinander. Dies liegt vor allem am gewählten Leakage-Faktor für das Holzbrett in Kombination mit dem verwendeten Sauger. Allgemein lässt sich jedoch feststellen, dass das prinzipielle Verhalten des Vakuum-Greifsystems mithilfe des präsentierten Modells gut beschrieben werden kann. Das validierte Modell kann im Anschluss beispielsweise zur Untersuchung des Energieverbrauchs bei gegebenen Handhabungszyklen verwendet werden. Dieser stellt in Zeiten steigender Energiepreise eine signifikante Größe dar, vor allem beim Einsatz energieintensiver Druckluft. Variable Parameter sind speziell für diese Untersuchung der Durchmesser der angelegten Leckage (links), der Innendurchmesser der Schläuche (Mitte links) sowie die untere (Mitte rechts) und obere Schwelle (rechts) der Zweipunktregelung. Zudem werden für die entsprechenden Parameter insgesamt vier unterschiedliche Düsengrößen der pneumatischen Vakuum Ejektoren untersucht. Verhaltensmodelle als integraler Bestandteil eines Digitalen Zwillings bieten die Möglichkeit zur effizienten Durchführung von Parameterstudien. Diese liefern für Experten der Vakuum-Handhabungstechnik bekannte Zusammenhänge und nachvollziehbare Erkenntnisse. Einige Effekte, deren Quantifizierung in realen Versuchen nicht so

ohne weiteres möglich wäre können damit aufgezeigt werden. Solche Effekte sind in den dargestellten Diagrammen in Bild 7 beispielsweise ein energetisches Optimum für bestimmte Innendurchmesser von Schläuchen (Mitte links), ein stark ansteigender Luftverbrauch ab einem gewissen Wert der Schaltschwelle H2 (Mitte rechts) oder ein energetisches Optimum für bestimmte Schaltschwellen H1 (rechts).

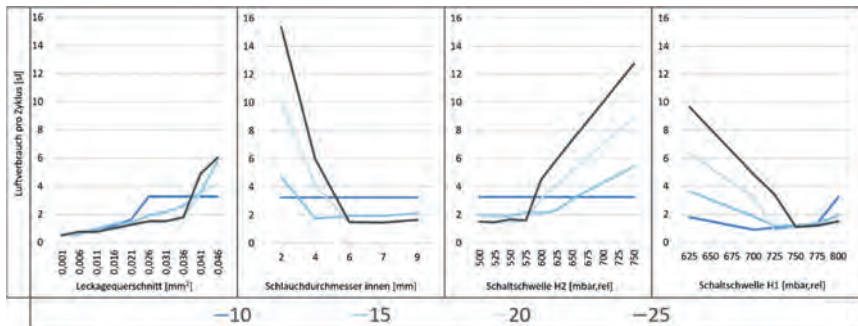


Bild 7: Luftverbrauch in Standardlitern über unterschiedliche Leckagedurchmesser [mm] (links), unterschiedliche Schlauchinnendurchmesser [mm] (Mitte links), unterschiedliche Schaltschwellen H2 [mbar,rel] (Mitte rechts) und unterschiedliche Schaltschwellen H1 [mbar,rel] (rechts)

Diese soeben beschriebenen Erkenntnisse können so zum einen zum allgemeinen Aufbau von Expertenwissen genutzt werden, was vor allem durch Schulungen am Digitalen Zwilling ermöglicht werden kann. Zum anderen können diese Luftverbräuche beispielsweise in einem automatischen Konfigurator ausgewertet um entsprechend der Anforderungen des Kunden die bestmögliche Konfiguration auszuwählen. Diese aktuell häufig manuell durchgeführte Arbeit ist stark getrieben durch Erfahrungswerte und benötigt dadurch ein großes Expertenwissen. Durch die schiere Produktvarianz und die dargestellten komplexen Parameter-zusammenhänge ist es jedoch nahezu unmöglich ein optimales System lediglich auf Basis von Expertenwissen aufzubauen. Werden jedoch für die Auslegung eines Vakuum Greifsystems automatisiert erstellte Verhaltensmodelle verwendet, kann ein effizienteres System schneller und kostengünstiger ausgelegt werden und bietet dabei zudem eine höhere Sicherheit, dass die gestellten Anforderungen erfüllt werden.

5. Zusammenfassung und Ausblick

Komponentenhersteller sind häufig früh in der industriellen Wertschöpfungskette aktiv, wodurch bei diesen auch häufig erste Digitale Zwillinge entstehen. Dabei ergeben sich für

Komponentenhersteller Möglichkeiten auf Basis des Digitalen Zwillings Mehrwerte in neuen Anwendungsfällen zu schaffen. Einige dieser Anwendungsfälle wurden beispielhaft präsentiert und deren Anforderungen hinsichtlich verfügbarer Daten und benötigter Modellierungstiefe dargestellt. Den Mehrwerten gegenüber steht jedoch die Herausforderung, den Digitalen Zwilling zu erstellen, ohne dass der dabei entstehende Aufwand den späteren Nutzen übersteigt. Hierfür wurde ein Konzept zur Erstellung Digitaler Zwillinge aus Sicht eines Komponentenherstellers zuerst makroskopisch in die industrielle Wertschöpfungskette eingeordnet und später detailliert erklärt. Wichtig bei der Erstellung Digitaler Zwillinge ist zum einen die semantische Aufbereitung bereits im Unternehmen vorliegender Daten und zum anderen die Erstellung von Verhaltensmodellen in prozessrelevanten Modellierungstiefen. Die Relevanz von Verhaltensmodellen konnte durch die energetische Analyse unterschiedlicher Parametereinflüsse auf ein Vakuum Greifsystem verdeutlicht werden.

Zukünftig soll das vorgestellte Konzept weiter verfeinert und ausgestaltet werden. Der Fokus liegt dabei auf Ansätzen zur Abstraktion von Verhaltensmodellen in die entsprechend prozessrelevanten Modellierungstiefen.

6. Literaturverzeichnis

- [1] M. Klein *et al.*, "Architektur und Technologiekomponenten eines digitalen Zwillings," *VDI-Kongress Automation 2019*, no. 2351, 89-102, 2019.
- [2] W. Kritzinger, M. Karner, G. Traar, J. Henjes, and W. Sihn, "Digital Twin in manufacturing: A categorical literature review and classification," *IFAC-PapersOnLine*, vol. 51, no. 11, pp. 1016–1022, 2018, doi: 10.1016/j.ifacol.2018.08.474.
- [3] B. Boss *et al.*, *Digital Twin and Asset Administration Shell Concepts and Application in the Industrial Internet and Industrie 4.0: An Industrial Internet Consortium and Plattform Industrie 4.0 Joint Whitepaper*. [Online]. Available: https://www.plattform-i40.de/PI40/Redaktion/DE/Downloads/Publikation/Digital-Twin-and-Asset-Administration-Shell-Concepts.pdf?__blob=publicationFile&v=11 (accessed: Apr. 22 2021).
- [4] B. Ashtari Talkhestani *et al.*, "An architecture of an Intelligent Digital Twin in a Cyber-Physical Production System," *at - Automatisierungstechnik*, vol. 67, no. 9, pp. 762–782, 2019.
- [5] F. Ocker, C. Urban, B. Vogel-Heuser, and C. Diedrich, "Leveraging the Asset Administration Shell for Agent-Based Production Systems," *IFAC-PapersOnLine*, vol. 54, no. 1, pp. 837–844, 2021, doi: 10.1016/j.ifacol.2021.08.186.

- [6] R. Rasor, D. Göllner, R. Bernijazov, L. Kaiser, and R. Dumitrescu, "Towards collaborative life cycle specification of digital twins in manufacturing value chains," *Procedia CIRP*, vol. 98, pp. 229–234, 2021, doi: 10.1016/j.procir.2021.01.035.
- [7] C. Wagner *et al.*, "The role of the Industry 4.0 asset administration shell and the digital twin during the life cycle of a plant," in *2017 22nd IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, Limassol, Sep. 2017 - Sep. 2017.
- [8] R. Rosen *et al.*, "Die Rolle der Simulation im Kontext des Digitalen Zwillings," *atp*, vol. 63, no. 04, pp. 82–89, 2021, doi: 10.17560/atp.v63i04.2534.
- [9] S. Thongnuch, "An approach to generating high-fidelity models for the virtual commissioning of specialized production machines and cells using MCAD models," Helmut-Schmidt-Universität / Universität der Bundeswehr Hamburg, 2021.
- [10] Plattform Industrie 4.0, "Die Verwaltungsschale im Detail: von der Idee zum implementierbaren Konzept," Berlin, 2019. Accessed: Jan. 14 2021. [Online]. Available: https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/verwaltungsschale-im-detail-pr%C3%A4sentation.pdf?__blob=publicationFile&v=14
- [11] S. Dierssen, "Systemkopplung zur komponentenorientierten Simulation digitaler Produkte," ETH Zurich, 2002.
- [12] A. Kufner, "Automatisierte Erstellung von Maschinenmodellen für die Hardware-in-the-Loop-Simulation von Montagemaschinen," Dissertation, Institut für Steuerungstechnik der Werkzeugmaschinen und Fertigungseinrichtungen der Universität Stuttgart, Universität Stuttgart, Stuttgart, 2012. Accessed: Jan. 27 2021.
- [13] Lenze SE, *Von der Simulation bis zur virtuellen Inbetriebnahme: Konkrete Anwendungsmöglichkeiten im Digital Engineering*. [Online]. Available: <https://www.lenze.com/de-de/virtuelle-inbetriebnahme-von-der-simulation-bis-zur-virtuellen-inbetriebnahme/> (accessed: Apr. 19 2022).
- [14] *Digitale Fabrik: Grundlagen*, 4499 Blatt1, Verein Deutscher Ingenieure e.V., Berlin, Feb. 2018.
- [15] *Entwicklung mechatronischer und cyber-physischer Systeme*, 2206, Verein Deutscher Ingenieure e.V.; Verband der Elektrotechnik, Elektronik und Informationstechnik, Berlin, Nov. 2021.
- [16] S. M. Scheifele, *Generierung des Digitalen Zwillings für den Sondermaschinenbau mit Losgröße 1*. Stuttgart: Fraunhofer Verlag, 2020.

- [17] B. Boss et al., *Digital Twin and Asset Administration Shell Concepts and Application in the Industrial Internet and Industrie 4.0: An Industrial Internet Consortium and Plattform Industrie 4.0 Joint Whitepaper*. [Online]. Available: https://www.plattform-i40.de/PI40/Redaktion/DE/Downloads/Publikation/Digital-Twin-and-Asset-Administration-Shell-Concepts.pdf?__blob=publicationFile&v=11 (accessed: Apr. 22 2021).
- [18] M. Liu, S. Fang, H. Dong, and C. Xu, "Review of digital twin about concepts, technologies, and industrial applications," *Journal of Manufacturing Systems*, vol. 223, September, p. 451, 2020, doi: 10.1016/j.jmsy.2020.06.017.
- [19] A. Deuter and S. Imort, "Product Lifecycle Management with the Asset Administration Shell," *Computers*, vol. 10, no. 7, 2021, doi: 10.3390/computers10070084.
- [20] S. Suess, C. Diedrich, and F. Auris, "An approach to automatically assign mechatronic components to an existing or identified classification," in *2018 IEEE 14th International Conference on Automation Science and Engineering (CASE)*, Munich, Germany, 82018, pp. 1525–1529.
- [21] F. Gabriel, P. Bobka, and K. Dröder, "Model-Based Design of Energy-Efficient Vacuum-Based Handling Processes," *Procedia CIRP*, vol. 93, pp. 538–543, 2020, doi: 10.1016/j.procir.2020.03.006.
- [22] D. Straub, *Methode zur technischen Auslegung von Vakuumgreifsystemen mit einer Mindesthaltedauer auf Basis fluidischer Untersuchungen*. Stuttgart: Fraunhofer Verlag; Universitätsbibliothek der Universität Stuttgart, 2021.
- [23] V. Stegmaier, W. Schaaf, N. Jazdi, and M. Weyrich, "Efficient creation of behavior models for Digital Twins exemplified for vacuum gripping systems," 2022, submitted.
- [24] J. Schmalz GmbH, *Vakuum-Komponenten von Schmalz ermöglichen effiziente Blech-handhabung*. [Online]. Available: <https://www.schmalz.com/de/vakuumtechnik-fuer-die-automation/produktnews/n/vakuum-komponenten-von-schmalz-ermoeglichen-effiziente-blechhandhabung/> (accessed: Jan. 5 2022).

Interaktion zwischen Steuerungen auf der Basis von OPC UA FX und deren Konfiguration durch Verwaltungsschalen

Prof. Dr. Ing **Ch. Diedrich**, Dipl.-Ing. **Th. Werner**,
Otto-von-Guericke-Universität, Magdeburg;
Dr.-Ing. **Matthias Riedl**, ifak e.V. Magdeburg

Kurzfassung

Die Nutzung der Informations- und Kommunikationstechnologien (IuK) in der industriellen Fertigung ist unverzichtbar geworden. Dadurch kann den hohen Anforderungen an Flexibilität, Qualität und Zuverlässigkeit der Produkte besser Rechnung getragen werden. Trotz hohem Digitalisierungsgrad in der Fertigung liegt die Gestaltung der Interaktionen zwischen Steuerungen hinter den Erwartungen weit zurück. Dies soll mit der OPC UA FX-Spezifikation verbessert werden, um dem Konzept der „Smart Factory“ auch auf diesem Gebiet näher zu kommen. Die Spezifikation beschreibt das OPC UA-Informationsmodell und die Mechanismen der Konfiguration der OPC UA-Server. Jedoch müssen die Informationen für die Konfiguration durch Anwendungen des Engineerings bereitgestellt werden.

In diesem Beitrag werden die Konzepte und die Konfiguration der Controller-to-Controller-Kommunikation erläutert. Daraus ergeben sich erforderliche Informationen, die die auszutauschenden Daten der Steuerungsprogramme und die QoS und andere Kommunikationsinformationen betreffen. Im Beitrag wird gezeigt, wie die Verwaltungsschale mit den OPC UA FX-Automation Components und dem Connection Manager zusammenarbeitet, um sich verändernde Steuerungskopplungen zur Laufzeit zu ermöglichen.

1. Motivation

Die Digitale Fabrik und Smart Manufacturing sind durch hohe Flexibilität gekennzeichnet. Dazu gehört auch, dass der Datenaustausch zwischen den Steuerungen zur Synchronisierung oder zur Weitergabe vom Zustand eines vorgelagerten Produktionsschritt nicht mehr vorkonfiguriert werden kann. Dies gilt im Besonderen, wenn der Anlagenbetreiber Maschinen von unterschiedlichen OEMs einsetzt und keinen Zugriff auf die Projekte der Maschinen hat. Außerdem können die Vorgänger-, Nachfolgerbeziehungen der Fertigungsschritte sich erst im operativen Betrieb ergeben.

Im Stand der Technik gibt es dafür keine standardisierten Lösungen. Eines der Probleme ist, dass die Client/Server-Kommunikationsarchitektur vorherrscht, obwohl hierfür eine logische Peer-to-Peer-Interaktion vorliegt.

Die OPC Foundation bietet mit dem OPC UA FX-Konzept [1...4] neben anderen Anwendungsfällen auch eine Controller-to-Controller-Kommunikation an, die hierfür eine standardisierte Lösung ermöglicht. OPC UA FX definiert ein harmonisiertes Informationsmodell für Automatisierungskomponenten, sowie die Orchestrierung von Informationsflüssen zwischen den Automatisierungskomponenten. Für deren Konfiguration sind eine Reihe von Informationen notwendig, die sich einerseits aus den Kommunikationsanschlüssen aber auch aus den Anwendungsstrukturen ergeben.

In diesem Beitrag werden die Konzepte und die Konfiguration der OPC UA FX-Controller-to-Controller-Kommunikation erläutert. Daraus ergeben sich die erforderlichen Informationen, die die auszutauschenden Daten der Steuerungsprogramme und die QoS und andere Kommunikationsinformationen betreffen. Im Beitrag wird gezeigt, wie die Verwaltungsschale mit den OPC UA FX-Servern und dem Connection Manager zusammenarbeitet, um sich verändernde Steuerungskopplungen zur Laufzeit zu ermöglichen.

Der Beitrag ist wie folgt gegliedert: Zunächst werden in Kapitel 2 das OPC UA FX-Konzept und das der Verwaltungsschale dargestellt. Kapitel 3 geht im Anschluss auf die Konfiguration des OPC UA FX-Servers mittels des Connection Managers ein. Daraus leitet sich der Informationsbedarf der Verwaltungsschale ab, der für die Konfiguration benötigt wird. Dieser ist in sogenannten Teilmodellen zu hinterlegen, die überblicksweise dargestellt werden. Der Beitrag schließt in Kapitel 6 mit einer Zusammenfassung und einem Ausblick ab.

2. Thematische Grundlagen

2.1 Controller-2-Controller-Kommunikation mit OPC UA FX

In der Automatisierungstechnik werden zurzeit industrielle Kommunikationssysteme wie z.B. PROFIBUS und PROFINET eingesetzt. Daneben etabliert sich OPC UA, wobei momentan im wesentlichen Client/Server und Ereignis-Notifier zum Einsatz kommen. FLC (Field Level Communication) ist eine weltweite Initiative, in der OPC UA für die IIoT-Kommunikation erarbeitet wird. Dazu gehört auch die vertikale und horizontale Kommunikation zwischen Steuerungen sowie von Steuerungen zu Feldgeräten. In der Version 1.0 steht die Controller-to-Controller-Kommunikation (C2C-Communication) im Mittelpunkt. Für die Umsetzung sind die neu entwickelten Pub/Sub-Kommunikationsdienste von OPC UA vorgesehen.

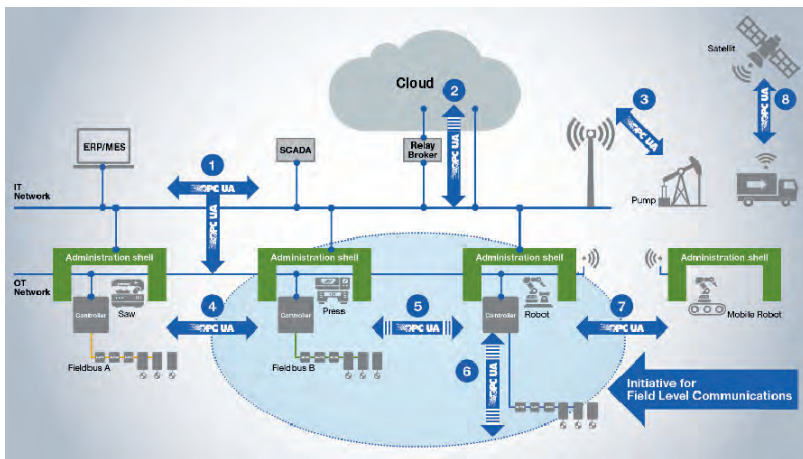


Bild 1: Einordnung der C2C-Kommunikation in die Automatisierungsarchitektur

Konzeptionell werden die Maschinen mit ihren Steuerungen als „Automation Component“ abgebildet (Abbildung 2). Dies ist eine eins zu eins (1:1)-Beziehung, d.h. eine Steuerung ist eine Automation Component, die im Informationsmodell eines OPC UA FX-Servers abgebildet wird. Die Variablen und Parameter, die für den Kommunikationsaustausch vorgesehen sind, werden innerhalb der Automation Component als sogenannte „Functional Entity“ dargestellt. Das ist eine ein zu viele (1:n)-Beziehung, da in einer Automation Component mehr als eine Functional Entity vertreten sein kann. Dies resultiert daraus, dass im operativen Betrieb Kommunikation zwischen mehreren anderen Stationen möglich sein soll, die unterschiedliche Ein- und Ausgangsdaten haben. Des OPC UA FX-Konzept sieht vor, dass die Konfiguration der OPC UA FX-Automation Components über eine standardisierte Schnittstelle mittels „Connection Manager“ erfolgt. Die Kommunikation hat dynamische und strukturelle Anforderungen, die in den OPC UA FX-Servern für eine konkrete Beziehung zwischen zwei Functional Entities zu konfigurieren sind.

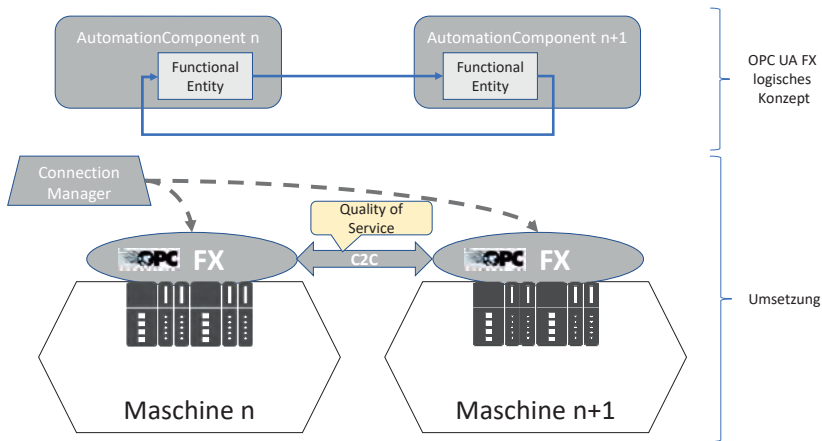


Bild 2: Konzept der C2C-Kommunikation

2.2 Maschinenauswahl durch VWS-Ausschreibungsverfahren

Die Verwaltungsschale (VWS) hat die Konzeptphase verlassen und verschiedene, meist Open Source-Lösungen stehen für die Umsetzung zur Verfügung. Die weitere Verbreitung wird in Zukunft besonders durch die neu gegründete IDTA (Industrial Digital Twin Association) vorangetrieben. Im Rahmen der AG1 der Plattform Industrie 4.0 wurde in Zusammenarbeit des ZVEI, des VDMA und der OPC Foundation eine Abbildung des Metamodells der Verwaltungsschale auf das Informationsmodell von OPC UA definiert. Das VWS-Metamodell entspricht der „Specification Details of the Administration Shell – Part 1: The exchange of information between partners in the value chain of Industrie 4.0; Version 3.0 RC01 Plattform Industrie 4.0, Berlin 2019“ [5] und [6].

Im Rahmen der Arbeiten der IDTA werden für ausgewählte Anwendungsfälle sogenannte Teilmodell-Templates erarbeitet, die für diese eine standardisierte Menge an Parametern und anderen Informationen von Assets bereitstellen. Beispiele sind das Digitale Typenschild, die technische Dokumentation, die Simulation, Fähigkeitsbeschreibung und vieles mehr. Die Inhalte dieser Teilmodelle werden dann von den Herstellern ausgefüllt und eine standardisierte Abfrage über die Interfaces ist möglich.

Ein Ausschreibungsverfahren zur Auswahl von Maschinen auf der Basis ihrer Fähigkeiten ist in VDI/VDE 2193 Teil 2 [10] definiert und verwendet die in der gleichen Richtlinie definierte I4.0-Sprache [9]. Basis ist es, dass mehrere Verwaltungsschalen interagieren und zu einem in

den Nachrichten definierten Sachverhalt ein Ausschreibungsverfahren abwickeln. Die Verwaltungsschalen verwenden dabei das Ausschreibungsprotokoll, in dem die verschiedenen Rollen definiert sind. Dabei existieren die Rolle ServiceProvider für eine Verwaltungsschale, die einen Dienst erbringt und ein Angebot dafür abgibt und die Rolle ServiceRequester, welche den Dienst abfordert und die Ausschreibung durchführt. Die Struktur ist in Abbildung 3 verdeutlicht. Abbildung 4 zeigt die dazugehörigen Abläufe¹. Nach diesem Ablauf erfolgt die Konfiguration der OPC UA FX-Automation Components.

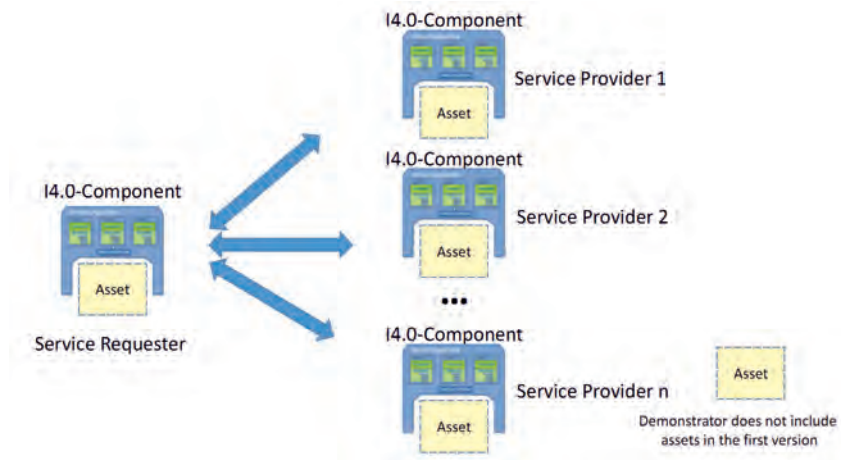


Bild 3: Struktur des Ausschreibungsverfahrens

Das Ausschreibungsverfahren wird nach der Beauftragung des ServiceProviders mit der Konfiguration der C2C zwischen den UA FX OPC UA-Servern fortgesetzt. Die dafür benötigten Informationen werden aus den Verwaltungsschalen gewonnen, der Prozess läuft jedoch außerhalb der Verwaltungsschalen im OPC UA-Kontext ab.

¹ Für weitere Informationen siehe auch ausführlichere Beschreibungen unter www.vwsvernetzt.de

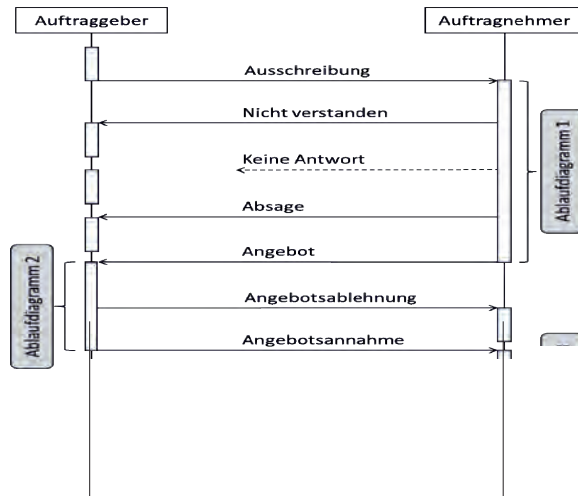


Bild 4 – Prinzipieller Ablauf des Ausschreibungsverfahrens ohne Auftragserteilung

3. Konfiguration der OPC UA FX-Connection Manager

Der wesentliche Use Case für OPC UA FX ist, dass die Steuerungen von je zwei oder mehreren (1:n) Maschinen (das können Werkzeugmaschinen oder andere zur Produktion gehörende Komponenten und Assets sein, wie Roboter, fahrerlose Transportsysteme, Lager, etc.) basierend auf der OPC UA FX-Spezifikation Daten miteinander austauschen (siehe auch [8]). Diese Kommunikation wird als Controller-to-Controller-Kommunikation (C2C-Kommunikation) bezeichnet (Abbildung 5). Zweck dieser Kommunikation ist es, dass beim Übergang eines zu produzierenden Produktes von einer Maschine zur anderen sowohl die für die logische Synchronisierung (Produktionsschritt der Vorgängermaschine abgeschlossen, der nächste Produktionsschritt kann beginnen) erforderlichen Statusinformationen (Status - Abbildung 5) als auch für den Transport evtl. benötigte Daten (workpiecetransport - Abbildung 5) ausgetauscht werden. Es wird davon ausgegangen, dass die auszutauschenden Daten in den Steuerungsprogrammen enthalten und gegenseitig bekannt sind (verschieden farbige Boxen in den grauen Kästen, die die Maschinen darstellen - Abbildung 5).

Zusätzlich kann eine Koordinatorsteuerung mit den Maschinensteuerungen interagieren, um werkstückspezifische Daten in die jeweilig aktive Steuerung einzuspielen bzw. Daten aus dieser auszulesen und die Maschinen untereinander zu koordinieren. Ziel ist es, die erforderliche

C2C-Kommunikationsbeziehungen einzurichten und zu aktivieren. Dafür wird die OPC UA FX-Spezifikation angewendet.

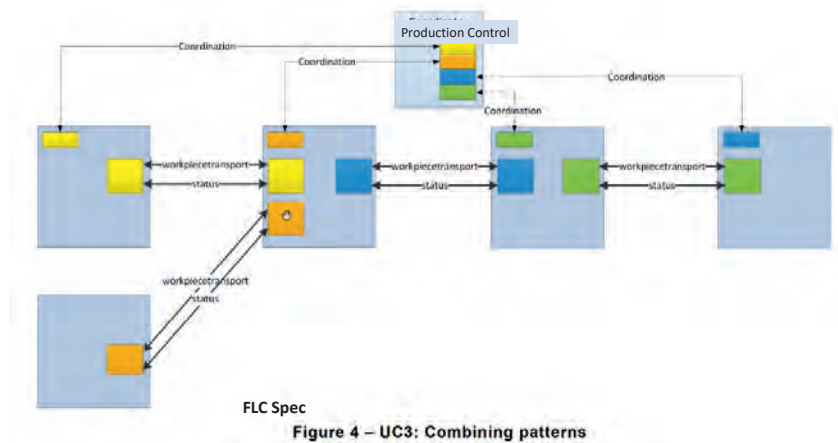


Bild 5: Einsatzszenario der UA FX-C2C-Kommunikation – beispielhaft [1]

Das Grundkonzept für die Konfiguration sieht vor, dass die Bereitstellung und die Akquisition der für die C2C-Konfiguration benötigten Informationen mittels Verwaltungsschalen organisiert wird. Der Konfigurationsprozess der OPC UA FX-Server und die operative Kommunikation läuft ausschließlich zwischen den OPC UA-Komponenten ab. Dafür wird wie folgt vorgegangen (Abbildung 6):

- Die Maschinen und wenn vorhanden der Koordinator werden je mit einer VWS ausgestattet
- Der Arbeitsplan für jedes individuelle Produkt liegt in einer entsprechenden Entität, z.B. Produktionsplanungs- oder -steuerungswerkzeug, die für eine VWS zugänglich sind
 - o Der Arbeitsplan besteht aus Arbeitsschritten, die jeweils auf einer Maschine ausgeführt werden, d.h. der Arbeitsplan beschreibt die Fähigkeiten, die von den Maschinen bereitzustellen sind.
- Ausschreibungsverfahren
 - o Zwischen VWS des Arbeitsplans (Service Requester) und VWS der Arbeitsschritt-ausführenden Maschinen (Service Provider) findet das I4.0-Ausschreibungsverfahren statt (blaue durchgezogene Linien in Abbildung 6).

- o Im Ergebnis ist die Zuordnung des jeweils nächsten Arbeitsschrittes zu der durch das I4.0-Ausschreibungsverfahren ausgewählten Maschine bekannt
 - o In der VWS des Arbeitsplanes werden die Vorgänger-Nachfolgerbeziehungen auf die entsprechenden OPC UA FX-Server abgebildet so dass die C2C-Kommunikation eindeutig identifiziert ist. Die Vorgängermaschine ist aus dem vorherigen Arbeitsschritt bekannt.
- C2C-Konfiguration
 - o Für diese Steuerungen (Vorgänger-Nachfolgerbeziehungen) sind die OPC UA FX-Deskriptoren aus den Steuerungen zu holen oder sie sind bereits bekannt
 - o Mit diesen und der bestimmten Vorgänger-Nachfolgerbeziehungen sind die Konfigurationsinformationen für den Connection Manager zu ermitteln.
 - o Diese müssen in das für den Connection Manager benötigte Notationsformat gebracht werden und an diese übertragen werden.
 - o Connection Manager können in einer Steuerung, in einem separaten System oder einer eigenen Komponente, z.B. Edge Device deployed werden. Es können auch mehrere Connection Manager gleichzeitig im System agieren.
 - o Jede einzelne C2C-Beziehung muss durch einen Connection Manager konfiguriert werden.
 - o In diesen Connection Manager werden die C2C-Konfigurationsinformationen hineingeschrieben (grüne gestrichelte Linien in Abbildung 6).
- OPC UA FX-Beziehungsmanagement
 - o Die Konfiguration erfolgt durch den Connection Manager wie sie in der UA FX-Spezifikation [2] beschrieben ist (graue gestrichelte Linien in Abbildung 6).
- Es kann sein, dass Maschinen eine für das Produkt und für diesen konkreten Arbeitsschritt spezifische Parametrierung/Konfiguration erhalten müssen. Dies ist zusätzlich zu organisieren, bevor die Maschine mit der operativen Arbeit beginnt.

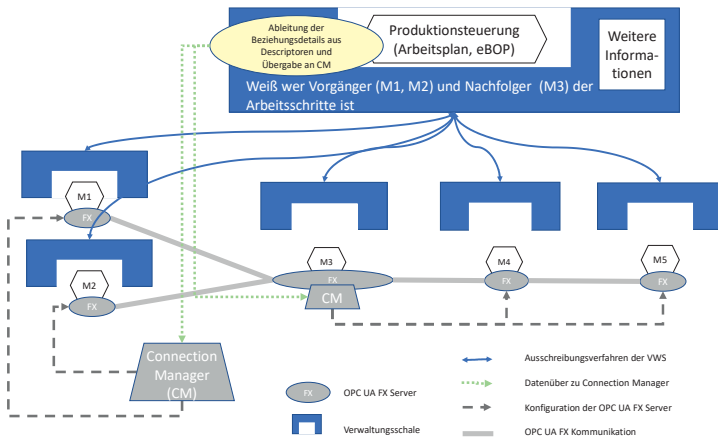


Bild 6: Konzeptioneller Ansatz für das Zusammenwirken von Connection Manager und VWS

4. Informationsbedarf in den Verwaltungsschalen

Die Analyse des Informationsbedarfs für die Konfiguration zeigt folgende wesentliche Informationsgruppen:

1. Beschreibung des Arbeitsplanes mit den einzelnen Arbeitsschritten
2. Identifizierende Informationen zu den Fähigkeiten der Maschinen entsprechend der Anforderungen aus dem Arbeitsplan
3. Zuordnung der OPC UA FX-Automation Components zu den Maschinen
4. Identifikation der Functional Entities, die für die jeweils ausgewählte Vorgänger-Nachfolger-Beziehung zutreffend sind
5. Bestimmung der OPC UA FX-Beziehungsdetails wie Adressierung und QoS.

Wie oben erwähnt werden in der IDTA Teilmodell-Templates erarbeitet, die einen standardisierten Zugriff auf Informationen der Assets erlauben. In der folgenden Tabelle 1 werden existierende Teilmodell-Templates mit dem Informationsbedarf für die C2C-Konfiguration verglichen. Daraus leiten sich fehlende Informationsanteile ab. Diese müssten dann manuell in den Konfigurationsprozess eingebracht werden.

Tabelle 1: Vergleich Informationsbedarf für die OPC UA FX-Konfiguration und existierende Teilmodell-Templates

Informationsbedarf	Teilmodell-Template (hier werden auch Teilmodelle aufgeführt, die noch nicht fertiggestellt sind)	Anmerkung
Zu 1: Arbeitsplanbeschreibung mit Fähigkeiten	-	Diese Aufgabe wurde noch nicht begonnen
Zu 2: Fähigkeitsbeschreibung	Submodel Template „Capability“	Die Konzepte wurden in verschiedenen Projekten ausgearbeitet und werden derzeit in ein Teilmodell-Template überführt
Zu 3: OPC UA FX-Automation Components	Submodel Template „Digital Nameplate“	Eine Zuordnung zwischen OPC UA 100 (Herstelleridentifikation) und den Angaben des Digitalen Nameplate sind eindeutig abbildbar. Eine Zuordnung ist dadurch eindeutig gegeben
Zu 4: Functional Entities	-	Im Fähigkeits-Teilmodell sind auch Details der Parameter enthalten. Eine Zuordnung zu den Functional Entities ist jedoch noch eindeutig möglich
Zu 5: OPC UA FX-Identifikation	Submodel Template „OPC UA Server Identification“	Dieses Template befindet sich derzeit in Bearbeitung
Zu 5: QoS	-	In der VDI/VDE2192 existiert eine QoS beschreibung in detaillierter Form. Eine Übertragung zum Teilmodell-Template-vorschlag wurde noch nicht vorgenommen.

6. Zusammenfassung und Ausblick

Die vorliegende Arbeit zeigt, wie die Informationen für die Konfiguration der C2C-Kommunikation des OPC UA FX-Konzeptes im Zusammenwirken mit der Verwaltungsschale zusammengestellt werden können. Dabei stellt sich heraus, dass diese unterschiedlichen planerischen und operativen Aspekte des Produktionsablaufes zuzuordnen sind. Eine assistierte oder automatische Konfiguration auch zur Laufzeit im operativen Betrieb gelingt nur, wenn diese Informationen an standardisierten Schnittstellen bereitstehen. Hier liegt die Stärke der Verwaltungsschale, da diese den lebenszyklus-übergreifenden Zugriff ermöglicht. Die Teilmodelle der Verwaltungsschale, die als Templates standardisiert vorgehalten werden, sind Träger dieser Informationen. Bei einer detaillierten Analyse stellt sich heraus, dass weitere Standardisierungen von Templates vorgenommen werden müssen, damit der gesamte Informationsbedarf abdeckbar ist.

Quellen

- [1] OPC 10000-80: OPC Unified Architecture - Field eXchange (UA FX), Part 80: UA FX Over-view and Concepts. OPC Foundation 2021
- [2] OPC Foundation, „OPC 10000-81 - UA Specification Part 81 - UA FX Connecting Devices and Information Model 1.00.00 RC2,“ 2022.
- [3] OPC Foundation, „OPC 10000-82 - UA Specification Part 82 - UA FX Networking 1.00.00 RC2,“ 2022.
- [4] OPC Foundation, „OPC 10000-83 – UA Specification Part 83 - UA FX OfflineEngineering 1.00.00 RC2,“ 2022.
- [5] Federal Ministry for Economic Affairs and Energy (BMWi). (Juni 2020). Details of the Asset Administration Shell Part 1 - The exchange of information between partners in the value chain of Industrie 4.0. V3.0 RC01. Berlin: Public Relations (publisher). Plattform I4.0 Text and Editing.
- [6] Federal Ministry for Economic Affairs and Energy (BMWi). (November 2020). Details of the Asset Administration Shell Part 2 Interoperability at Runtime – Exchanging Information via Application Programming Interfaces. V1.0 RC01 Berlin: Public Relations (publisher). Plattform I4.0 Text and Editing.
- [7] Industrial Digital Twin Association – WG2 Submodels: Liste von Submodel-Templates in Entwicklung. IDTA interner Arbeitsbereich. Eingesehen 02.05.2022.
- [8] S. Höme, T. Fischer, G. Biehler, S. Kerschbaum. Sichere interoperable Maschine-zu-Maschine-Kommunikation basierend auf OPC UA Field Exchange. Automation Kongress 2022. Konferenzunterlagen.
- [9] VDI/VDE 2193 Blatt 1: Sprache für I4.0-Komponenten. Düsseldorf: VDI, 2019
- [10] VDI/VDE 2193 Blatt 2: Sprache für I4.0-Komponenten. Interaktionsprotokoll für Ausschreibungsverfahren. Düsseldorf: VDI, 2019

Framework for bringing wireless communication into the context of Industrie 4.0.

Gustavo Cainelli, Dr.-Ing. Lisa Underberg, Dr.-Ing. Lutz Rauchhaupt,
Institut für Automation und Kommunikation e.V. Magdeburg

1. Kurzfassung

In Industrie 4.0 (I4.0) muss das Produktionssystem flexibel an Kundenanforderungen angepasst werden können, ohne dass jedes Detail im Voraus geplant werden muss wie in einem starren Produktionssystem. Außerdem kann das Kommunikationssystem nicht für den schlimmsten Fall geplant und installiert werden. Sind aufgrund von Produktanforderungen oder aufgrund von Störungen im Funkkanal Änderungen notwendig, muss das Kommunikationssystem automatisch angepasst werden. Daher ist es notwendig, dass das Kommunikationssystem und das Produktionssystem interagieren können, um Ressourcen auszuhandeln. Um dieses Maß an Anpassungsfähigkeit anzugehen, schlägt diese Arbeit ein Framework vor, das eine digitale Domäne mit einer physischen Domäne verbindet und zum Entwerfen, Validieren und Untersuchen von Anwendungsfällen für Resilienz, Interferenzrobustheit oder Datenanalyse verwendet werden kann. Die digitale Domäne enthält Asset Administration Shells (AAS) von 5G-System- und Produktionssystemelementen. Die physikalische Domäne besteht aus einem 5G-Simulationsmodell. Jedes Asset im Simulationsmodell hat sein AAS in der digitalen Domäne. Das AAS enthält die Modelle mit den relevanten Parametern des drahtlosen Kommunikationssystems und kann Nachrichten mit anderen AAS und auch mit dem Asset austauschen. Anhand der Ergebnisse eines Anwendungsfalls wird exemplarisch gezeigt, wie Kommunikationsparameter von den digitalen Zwillingen genutzt werden können, um den Automatisierungsprozess anzupassen, wenn eine Störung im Funkkanal auftritt.

2. Abstract

In Industry 4.0 (I4.0) the production system must be flexible to adapt to customer requirements, without the need to plan every detail in advance as in a rigid production system. Also, the communication system cannot be planned and installed for the worst case. If changes are necessary due to product requirements, or due to disturbances in the wireless channel, the communication system must be adapted automatically. Therefore, it is necessary that the communication system and the production system can interact to negotiate resources. To address this level of adaptability this work proposes a framework that connects a digital domain with a physical domain and can be used to design, validate, and investigate use cases for resilience,

interference robustness or data analytics. The digital domain contains Asset Administration shells (AAS) of 5G system and productions system elements. The physical domain is composed of a 5G simulation model. Each asset in the simulation model has its AAS in the digital domain. The AAS contains the models with the relevant parameters of the wireless communication system and can exchange messages with other AAS and also with the asset. The results of a use case are presented to exemplify how communication parameters can be used by the digital twins to adapt the automation process if a disturbance in the radio channel happens.

3. Introduction

Usually, the traditional manufacturing systems have mass production and the number of products that will be produced is known. This kind of system is not flexible and is static, which means the production system is built to produce a certain kind of product and will produce it until the system is reconfigured, usually by a human. It is also well-engineered, which means that everything is planned and in some cases, it is planned for the worst-case which leads to a non-efficient use of resources, for example, the wireless communication medium. The Industry 4.0 (I4.0) concept corresponds to adapting the use of the production system to the needs of customers, without planning every detail of the production system in advance aiming to enable mass customization of manufactured products. Moreover, in case external or internal disturbances occur, the production system must cope with that and reconfigure its parameters to avoid downtime.

The wireless communication system (WCS) is one of the key elements of I4.0 and it has to cope with the level of adaptability required by Industrie 4.0. For this reason, it cannot be planned and installed for the worst case, it means that if changes are necessary, the WCS and/or its use of the medium should be adapted accordingly. It includes changes due to the production process requirements or changes due to external disturbances, for example, interferences. Moreover, a WCS brings the possibility of a high number of devices and mobility which means that in some cases the only way to optimize the network configuration process is if the system can adjust itself. So to put WCS in the context of I4.0, self-configuration is required and for that collaboration between the production system and communication systems is necessary. To enable this collaboration between the parties, we propose a digital twin approach.

The digital twin is a virtual representation of an asset. Assets are things that are valuable to an organization and they can be devices, machines, components, documents or even software.

The digital twin is not only used for the pre-operation phase but also for the operation phase. It can monitor the physical system and changes its parameters when necessary [1] [2].

Fig. 1 shows a representation of the assets in the physical domain and their virtual counterparts in the digital domain. The figure shows three layers. The first one, the production system (grey layer) includes the mechanical parts of the system. In the second one (green layer) are the automation elements that control the production system. The automation system relies on the wireless system (blue layer) to exchange productive data [3]. These three layers are present in both domains. The physical domain is the set of physical and non-physical assets and their relationships while the digital domain is the set of virtual counterparts of elements from the physical domain.

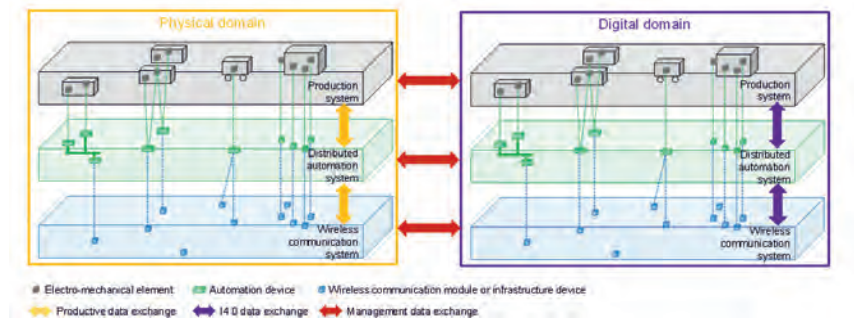


Fig. 1: Physical and digital domain (adapted from [3]).

The data exchange is distinguished into three types:

Productive data exchange: It is data exchange between elements within the physical domain. It is the data traffic of the production process. For example, a sensor sends its variable process to a PLC that uses it to calculate the control action and sends a command to an actuator also using the way.

I4.0 data exchange: it is data exchange within the digital domain. For example, a digital twin sends a request to another one to adapt a system parameter.

Management data exchange: It is data exchange between domains. It is used for example for the digital twin to reconfigure its asset. Moreover, the digital twin can monitor the current status of the physical domain and make decisions acting on the production system, automation system or wireless communication system.

To implement the Digital Twin, Plattform Industrie 4.0 proposed the Asset Administration Shell **Fehler! Verweisquelle konnte nicht gefunden werden.** The AAS is a concept to implement the Digital Twin for Industrie 4.0. The AAS contains information that represents the characteristics and behaviours of an entity here called an asset. Therefore, the AAS is a tool to create fully digital versions of any kind of asset in a factory. Through the AAS an asset becomes an I4.0 component. The AAS provides functionalities of data gathering, collaboration, negotiation, reconfiguration, and decision making [5].

4. Related works

In [6] a review of the current AAS's status of the art is presented. Moreover, the authors propose a general template for an AAS. Finally, they present a use case for interoperability where assets of different types can interact with each other through an AAS gateway. The work of [7] presents the implementation of the AAS in an industrial scenario. The goal is to test the interoperability between a robotic arm and grinding machine by integrating them through the AAS. Between the defined submodels were identification, documentation, job status and axis positions. In [8] an AAS approach is implemented to allow plug-and-produce (PnP) of industrial field devices. The use case contains three robots and a turntable. The authors propose I4.0 adaptors to connect the assets to them. Each robot system can be operated independently and the AASs are visually demonstrated using a WebUser Interface (Web UI). The defined submodels were documentation, interaction (between the AAS), and pick-and-place, among others. The work presented in [9] addresses the integration between product lifecycle management (PLM) and application lifecycle management (ALM). A discussion about the potential of the AAS as a standardized concept for PLM/ALM is performed. The submodels considered here are the product submodel and the application submodel. The authors show how to link elements between these submodels.

Although the works [6] and [8] indicate that a communication submodel is necessary when modelling the AAS, they do not provide detailed information on that. Moreover, these works do not have the same point of view as our proposal which consists of modelling the communication submodel not just allowing communication between assets but using this information to manage and adapt the production system.

5. Digital Twin Approach

This section presents the proposed approach. The general architecture that connects the digital and physical domain is discussed. Moreover, the main aspects of the implementation of each domain are shown.

a) General Architecture

Fig. 2 shows the general architecture of the framework. It is composed of the digital and physical domains. The physical domain contains the field assets that can be both hardware as ASSET_M1 and software as control applications running in pcs like ASSET_APP1. The digital domain contains the AAS. Between the assets (solid yellow line), there is productive data exchange, while between the AASs (dashed purple line) there is I4.0 data exchange. Moreover, there is management data exchange between the digital and physical domains.

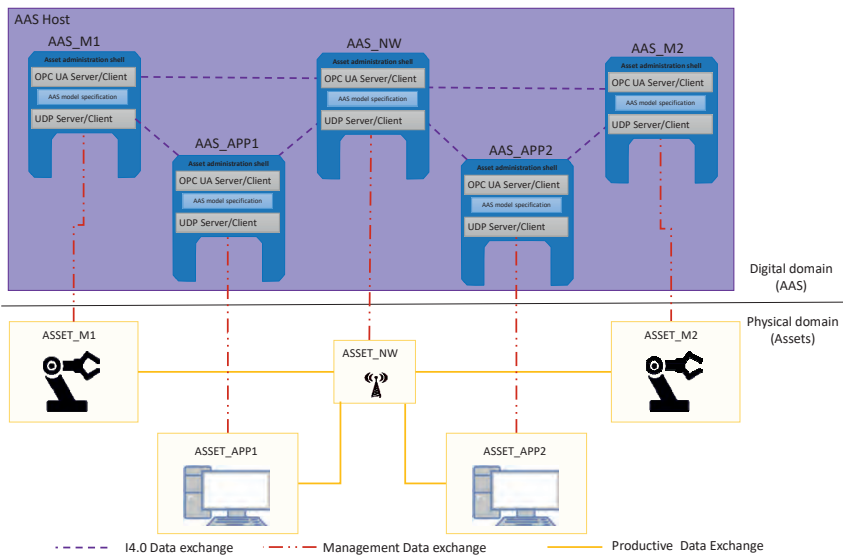


Fig. 2: General architecture connecting physical and digital domain.

The physical domain is composed of a 5G simulation model called Simu5G developed by [10]. The assets were prepared for being digitalized by exposing information on their parameters to the AASs. The digital domain is composed of OPC UA servers and clients. The OPC UA serves as a communication interface for connecting various AAS. Moreover, the information model that describes the asset is within the OPC UA server.

The AAS can interact and exchange data with one another. The AAS can also collect data from the assets. The AASs are implemented by developing an OPC UA server with open62541 [1]. First, the information model of the asset is developed using SIOME 0. After that, the a NodeSet2.xml file is exported and integrate into the OPC UA server.

OPC UA provides interfaces for data exchange between industrial devices and systems. The client-server communication scheme provided by OPC UA is used as an AAS communication interface as OPC UA organizes static and dynamic data into structured information. Data exchange between the two OPC UA applications proceeded via an OPC UA client.

The AAS is composed of a passive part and an active part. The passive part is the properties of the asset which are readable and/or modifiable. For example, Static device configuration and status parameters. On the other hand, the active part consists of procedures and algorithms performed by the AAS. The active part is composed of methods and can be used for example to read and write properties values in the AAS or the asset. Moreover, the active part of an AAS has decision-making functionalities. Some examples are Status assessment and Status control.

b) Communication submodel

The asset's information is described in the AAS as submodels. The AAS may incorporate general submodels (e.g. identification) and also specific submodels, for example, communication. There is no limit to the number of submodels of an AAS, as they are defined according to the level of detail required for the model. Below the submodel level, there are the submodel elements like SubmodelCollection and Properties. The submodel elements store specific data related to the submodel. For example, a property, which is a submodel element type, can contain a value that represents a parameter of the asset. It can be of several types, for example, INT, BOOL or STRING. An example is the current status of a pump (ON / OFF). For a wireless system, this can be values of parameters that indicate the status of radio channel as SINR, RSSI, among others.

Fig. 3 shows an example of an AAS of a 5G-enabled device which can be any device that can communicate using 5G technology. This includes sensors, actuators or even more complex systems such as a machine. In this example, it is an AAS of a machine. Some submodels examples are identification and localization of the machine. Other possible submodels could be related to the machine's functions (bending, drilling, moving...). The focus of this work is on the communication submodel.

The communication submodel is responsible for storing all the relevant information related to communication. For that, several submodels and properties within the communication submodel should be defined to group the information. Currently, within the communication submodel, two main submodels are being considered: *WirelessModule* and *QoS*. The *WirelessModule* refers to the characteristics of the 5G communication module itself considering

datasheet parameters and the current values of parameters. The QoS stores information related to the quality of service, including the requested, assured and current values. Other examples of submodels that can be defined for a 5G-enabled device are the SIM card that will aggregate all information related to the SIM card used. Parameters such as IMSI, ICCID, and PIN should be stored here.

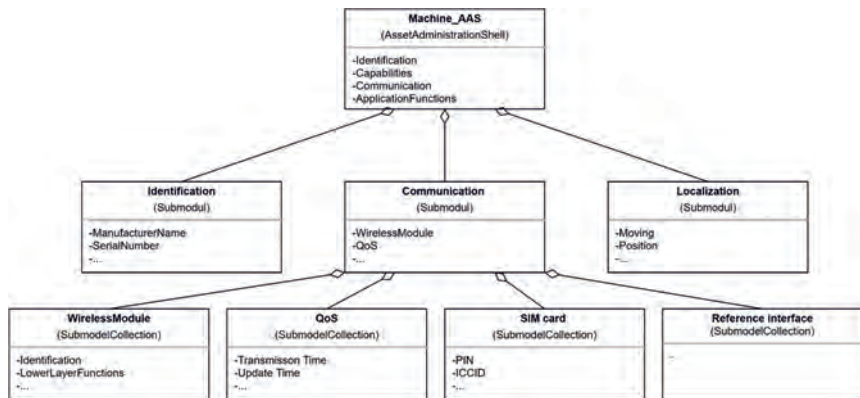


Fig. 3: Asset administration shell of a machine with wireless communication.

Fig. 4 shows two submodels of AAS_M1 in an OPC UA client viewer. The WirelessModule contains all the relevant information related to the wireless communication module. The data stored in this submodel can be divided into static and dynamic. Static ones are the capabilities of the 5G module as delivered. This data generally can be found in the module's documentation (datasheets). Examples are frequency band and maximum transmits power. Information related to identification, for example, manufacturer, and serial number, among others are also static. The dynamic ones are updated during the operational process. It includes RSSI, RSRP and current output power. Among the submodels and properties defined, there is the submodel TransceiverFunction. Two submodels element collections are defined within this submodel element: *Transmitter* and *Receiver*. Under the Receiver, there are properties such as receiver sensitivity, and SINR, among others.

The QoS currently has a submodel template specification under discussion on Platform I4.0. In this template, some characteristics parameters described in VDI/VDE 2185 part 4 [12] are considered parameters such as transmission time (TT) and update time (UT) are considered.

Furthermore, each of the characteristic parameters is composed of required value, guaranteed value and current value.

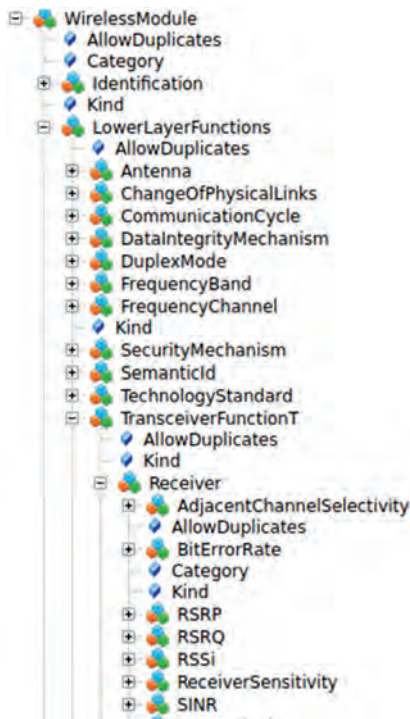


Fig. a: Wireless Module.

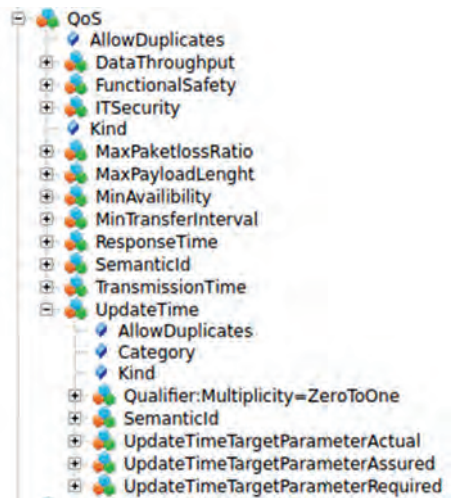


Fig. b: Quality of service.

Fig. 4: Submodels of wireless devices in an OPC UA client viewer.

A management system can use information from the AAS related to the wireless signal (e.g. RSSI, RSRP, SINR) to monitor the conditions of the radio channel and take actions, for example, adapt the production process, based on these values.

6. Use cases

Below we briefly describe two use cases where the communication submodel is required:

Use case 1: “Managing the WCS based on production requirements”. In this use case, the parameters of the wireless communication system are adjusted based on requirements that

come from the production system. For example: If a product has priority, then more communication resources are allocated to the machine that produces it. In other words, the WCS is adapted based on production and automation system requirements.

Use case 2: "Production process management based on the quality of communication". In this use case, the parameters of a production process element (e.g. machine, device, AGV) are adapted based on the quality of communication. For example: if the wireless channel is under interference and messages are being lost, the production system is adapted to cope with that. In other words, the production system is adapted based on the WCS.

7. Experiments and Results

An experiment based on use case two (previous section) was conducted to prove the concept. Three parameters are monitored or adapted in the experiment:

Transfer Interval (TI): time between two transmitted messages on the source side.

Update Time (UT): time between two received messages on the destination side.

Signal-to-interference-plus-noise ratio (SINR): it is the ratio between received signal power and interference. It is a way to measure the quality of wireless connections.

The scenario of the experiment is the following: the ASSET_APP1 exchanges productive data with the ASSET_M1 cyclically. The transfer interval of the application is 131 ms. The machine uses this data for control purposes. The UT has timing constraints and cannot exceed 132 ms. If messages are lost in the wireless channel due to interferences or disturbances, the UT may exceed this threshold. If this happens, then the machine enters a failure state and stops producing. In an ideal case, the UT should follow the transfer interval [1]. However, if disturbances occur messages might be lost in the physical medium and retransmission will be necessary. This would consequently increase the UT of some messages. The scenario was programmed to have two moments in which disturbances on the wireless channel happen.

The experiment is divided into two test cases: test case 01 (TC01) without AAS and test case 2 (TC02) with AAS.

The results of TC01 can be seen in Fig. 5. The chart shows the UT of messages in ASSET_M1. Most of the messages have UT equal to TI, however, there are two main areas of the chart where the UT is toggling below and above 131 ms.

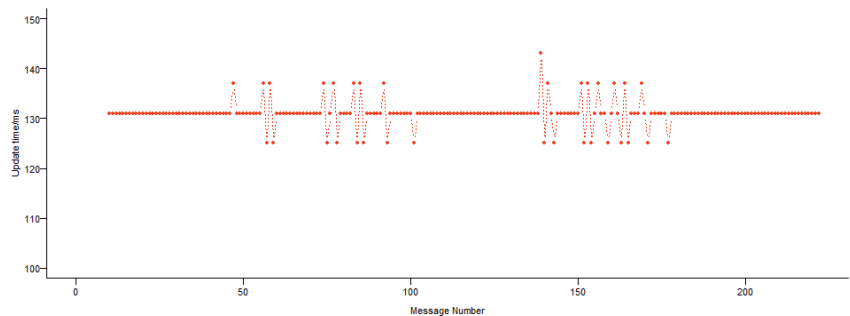


Fig. 5: Update time on machine 1 with disturbances.

Fig. 6 shows sequence diagram of the interaction between the AASs and Assets in TC02. In this test case, the AAS_M1 monitors the UT of ASSET_M1. If it detects that the UT is higher than the threshold, the AAS_M1 sends a request to AAS_APP1 asking to reduce the TI. If possible the AAS_M1 reduces the TI. Now even with retransmission in the wireless medium, the ASSET_M1 receives messages in a period shorter than 132 ms. The AAS_M1 monitors the SINR, and when it is back to a normal value it informs the AAS_APP1 that the TI can return to normal.

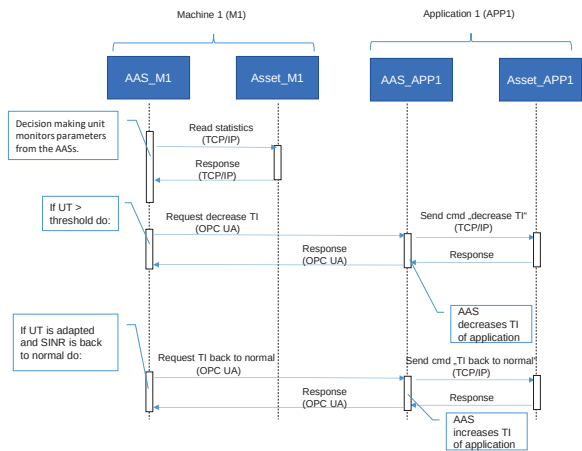


Fig. 6: Sequence diagram of AAS's and Asset's interaction.

The result of TC02 can be seen in Fig. 7. The UT is still toggling but even the higher value is below the threshold.

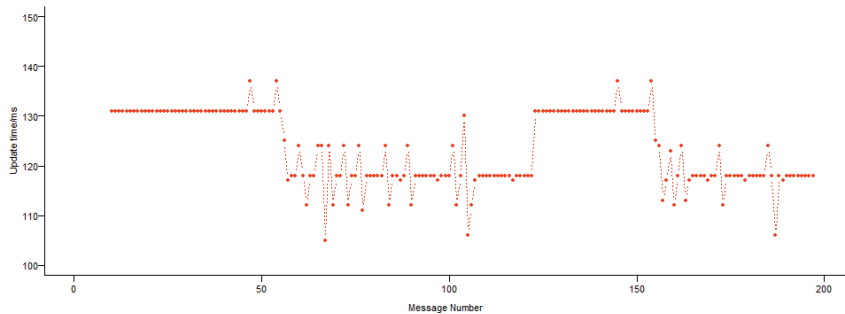


Fig. 7: Update time on machine 1 when transfer interval is adapted.

One alternative to avoid the UT exceeding the threshold is to set the transfer interval of the control application to a value that ensures the machine will receive messages each 131 ms. This would solve the problem but implies setting the system for the worst case. To use the medium more efficiently, the system must be able to detect possible problems and adapt itself. This method avoids that the production system and wireless system are built for the worst case. It also increases flexibility as now the system can adapt application parameters based on wireless communication parameters making the system more resilient [12].

8. Conclusion and Future Works

This paper shows that in Industry 4.0 the WCS cannot be planned for the worst case. It has to adapt autonomously to cope with production system requirements or external disturbances. The digital twin approach following the Asset administration shell concept is a way to cope with this level of flexibility. To allow the collaboration between automation and the communication system the communication parameters must be included in the virtual description of the assets. For that, the manufacturers of 5G devices should provide access to some device parameters such as SINR, and RSRP among others. These parameters were described using the asset administration shell and uploaded in OPC UA servers.

The method was tested using a 5G simulation model and the authors are now working to validate this approach in a real 5G network. The presented WirelessModule submodel can be further developed and become a submodel template to be used for developers that want to represent wireless devices in the AAS.

9. Acknowledgement

This work was funded in parts by the project "5GIWCoW" under contract 165GU041D, funded by the Federal Ministry of Transport and Digital Infrastructure, Germany.

10. References

- [1] 5G-ACIA (2021). Using Digital Twins to Integrate 5G into Production Networks (White Paper). URL <https://5g-acia.org/wp-content/uploads/2021/05/5G-ACI A Using Digital Twins to Integrate 5G into Production Networks single-pages.pdf>.
- [2] Schroeder, G.N., Steinmetz, C., Pereira, C.E., and Espindola, D.B. (2016). Digital twin data modelling with automationml and a communication methodology for data exchange. *IFAC-PapersOnLine*, 49(30), 12 { 17. DOI:<https://doi.org/10.1016/j.ifacol.2016.11.115>. 4th IFAC Symposium on Telematics Applications TA 2016.
- [3] 5G-ACIA, "Performance Testing of 5G Systems for Industrial Automation (White Paper)," Tech. Rep. April, 2021, <https://www.5gacia.org/publications/>.
- [4] Plattform I4.0 (2018a). Details of the asset administration shell - from idea to implementation. URL <https://www.plattform-i40.de/>.
- [5] X. Ye, M. Yu, W. S. Song and S. H. Hong, "An Asset Administration Shell Method for Data Exchange Between Manufacturing Software Applications," in *IEEE Access*, vol. 9, pp. 144171-144178, 2021, doi: 10.1109/ACCESS.2021.3122175
- [6] Ye, X. and Hong, S.H. (2019). Toward industry 4.0 components: Insights into and implementation of asset administration shells. *IEEE Industrial Electronics Magazine*, 13(1), 13{25. doi:10.1109/MIE.2019.2893397. Ye, X., Jiang, J., Lee, C., Kim, N., Yu, M., and Hong, S.H. (2020). Toward the plug-and-produce capability for industry 4.0: An asset administration shell approach *IEEE Industrial Electronics Magazine*, 14(4), 146{157. doi:10.1109/MIE.2020.3010492.
- [7] Inigo, M.A., Porto, A., Kremer, B., Perez, A., Larrinaga, F., and Cuenca, J. (2020). Towards an asset administration shell scenario: a use case for interoperability and standardization in industry 4.0. In *NOMS 2020 -2020 IEEE/IFIP Network Operations and Management Symposium*, 1{6. doi:10.1109/NOMS47738.2020.9110410.
- [8] Ye, X., Jiang, J., Lee, C., Kim, N., Yu, M., and Hong, S.H. (2020). Toward the plug-and-produce capability for industry 4.0: An asset administration shell approach. *IEEE Industrial Electronics Magazine*, 14(4), 146{157. doi:10.1109/MIE.2020.3010492.
- [9] Deuter, A. and Imort, S. (2020). Palm/ALM integration with the asset administration shell. *Procedia Manufacturing*, 52, 234{240. doi:10.1016/j.promfg.2020.11.040.
- [10] Nardini, G., Stea, G., and Sabella, D. (2020). Simu5g: A system-level simulator for 5g networks. In *In Proceedings of the 10th International Conference on Simulation and Modeling Methodologies, Technologies and Applications (SIMULTECH 2020)*, 68{80.

- [11] Open62541 OPC UA Library, URL: <https://open62541.org/>
Siemens OPC UA Modeling Editor (SiOME) zur Umsetzung von OPC UA Companion Spezifikationen. URL: https://support.industry.siemens.com/cs/attachments/109755133/109755133_SiOME_MAN_V25_de.pdf
- [12] VDI/VDE-Gesellschaft Mess- und Automatisierungstechnik, VDI/VDE 2185 Part 4, "Radio-based communication in industrial automation - Metrological performance rating of wireless solutions for industrial automation applications.", 2019-03
- [13] Cainelli, G. and Rauchhaupt, L. (2021). Introducing resilience in industrial 5g systems using a digital twin approach. In 2021 17th IEEE International Conference on Factory Communication Systems (WFCS), 33{36.doi:10.1109/WFCS46889.2021.9483618.

Verwaltungsschale-basierter Ansatz für die Umsetzung von auftragsgesteuerter Produktion

M.Sc. **Chris Urban**, M.Sc. **Alexander Belyaev**,
Otto von Guericke Universität Magdeburg;
Prof. Dr.-Ing. **Christian Diedrich**,
ifak – Institut für Automation und Kommunikation e.V.

1. Kurzfassung

In diesem Beitrag wird ein Ansatz für die Umsetzung von auftragsgesteuerter Produktion (AGP) basierend auf Konzepten der Plattform I4.0 vorgestellt. Es werden die dabei auftretenden Ausprägungen und Wechselwirkungen von relevanten Konzepten und Spezifikationen diskutiert. Aus einzelnen Teilaspekten und Bausteinen wird ein Gesamtbild erschaffen, das einen Beitrag für die Umsetzung von AGP leisten soll. Darüber hinaus diskutiert der Beitrag weitere Elemente der unternehmensübergreifenden, auftragsgesteuerten Produktion, die weiterhin zu erforschen und standardisieren sind, um diesen Anwendungsfall zu einem höheren Reifegrad zu führen. Des Weiteren beschreibt der Beitrag den Aufbau des ersten lauffähigen Verwaltungsschale-basierten I4.0-Ökosystems in einem gemeinsamen, deutschlandweiten Demonstrator.

2. Einführung in das Anwendungsszenario „Auftragsgesteuerte Produktion“

Zur Veranschaulichung der von Industrie 4.0 angestrebten digitalen Transformation der produzierenden Industrie formuliert die Plattform Industrie 4.0 Anwendungsszenarien [1]. Eines der relevanten Anwendungsfälle ist die Auftragsgesteuerte Produktion (AGP). Dieses Szenario beschreibt den Übergang von Massenfertigung zu kosteneffizienter Herstellung von individualisierten und einzigartigen Produkten [2]. Dahinter verbirgt sich die Idee, dass die Produktionsmittel nicht wie gewohnt in starre und immer wiederkehrende Produktionsabläufe eingebunden sind, sondern bezogen auf den aktuellen Auftrag ausgewählt, konfiguriert, parametrisiert und in den Produktionsprozess eingeplant werden. Basis für die Gestaltung der angestrebten Transformation sollen Konzepte und Spezifikationen werden, die im Rahmen der Initiative Plattform I4.0 erarbeitet werden. Einige Konzepte, die aus Sicht der Plattform I4.0 für die Umsetzung der von I4.0 skizzierten Visionen maßgeblich sind, konnten bereits identifiziert werden. Ihre Standardisierung und Entwicklung schreiten voran. Eines der wichtigsten Konzepte ist dabei die Verwaltungsschale (VWS). Zur Operationalisierung und weite-

rer Technologieentwicklung der Verwaltungsschale wurde die Industrial Digital Twin Association (IDTA) gegründet.

Jedoch ist die Lösung für die Umsetzung der AGP in aktuellen technischen Spezifikationen und Standardisierungsaktivitäten nicht direkt adressiert. In den von Plattform I4.0 publizierten Dokumenten findet sich keine Beschreibung des ganzheitlichen Ansatzes für die Umsetzung der AGP, der die dafür bereits vorhandenen und benötigten Bausteine identifiziert, beschreibt und die Rolle von zentralen I4.0-Konzepten wie die VWS einordnet.

Dies motiviert das Forschungsziel dieser Arbeit. In diesem Beitrag wird ein Ansatz für die Umsetzung von AGP basierend auf Konzepten der Plattform I4.0 vorgestellt. Aus einzelnen Teilaspekten und Ausprägungen von relevanten I4.0-Konzepten wird ein Beitrag für die Umsetzung von AGP geschaffen. Darüber hinaus diskutiert das Papier weitere Aspekte der unternehmensübergreifenden AGP, die weiterhin zu standardisieren sind, um diesen Anwendungsfall zu einem höheren Reifegrad zu führen.

3. Ausprägungen der auftragsgesteuerten Produktion

Für die Ableitung und Diskussion von für die Umsetzung der AGP notwendigen Bausteinen ist es zielführend, zwischen zwei Ausprägungen dieses Anwendungsfalls zu unterscheiden (siehe Abbildung 1).

Liegt der Fokus auf der Automatisierung und Flexibilisierung unternehmensinterner Planungsprozesse, der Gestaltung von modularen und wandelbaren Produktionsanlagen, der oft diskutierten Auflösung der Automatisierungspyramide und horizontaler Vernetzung von Maschinen, bietet es sich an, diese Ausprägung als „unternehmensinterne AGP“ zu bezeichnen. Eines der meistbekannten Beispiele der unternehmensinternen AGP ist die Idee, dass die smarten Produkte ihre Produktionsschritte kennen und dann durch fortlaufende Auswahl der richtigen Produktionsmittel einen zu diesem Zeitpunkt günstigen Weg durch die Produktion finden [2, 3].

Digitalisierung und Automatisierung von unternehmensinternen Produktionsprozessen ist eine Grundlage für die zunehmende Vernetzung und horizontale Integration in der gesamten Wertschöpfungskette und der angestrebten Wandlung von klassischen Wertschöpfungsketten zu vernetzten Wertschöpfungsnetzwerken.

Die evolutionäre Weiterentwicklung von unternehmensinterner AGP, d.h. die automatisierte Einbindung von externen Produktionskapazitäten und die Erstellung von dynamischen Kooperationsbeziehungen über die Standort-, bzw. Fabrik- und Unternehmensgrenzen hinweg, kann als „unternehmensübergreifende AGP“ bezeichnet werden. „Production as a Service“ wird in dieser Arbeit als Synonym zu unternehmensübergreifender AGP betrachtet. Unter-

nehmensübergreifende AGP bedeutet, dass Smarte Fabriken selbst zu Knoten der zukünftigen Produktionsnetzwerke und digitalen Ökosysteme werden. Die Produktion erfolgt im Verbund von internen und externen Fertigungsfähigkeiten und -kapazitäten. Fertigungspartner finden sich auftragsbezogen und ad hoc.

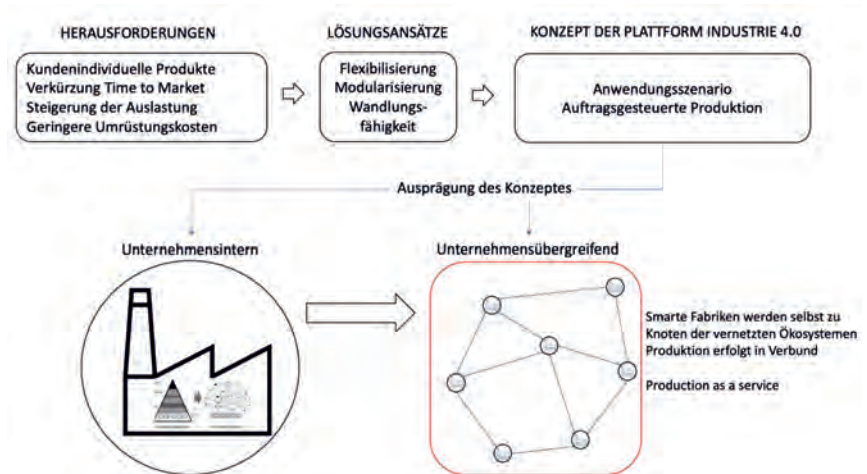


Bild 1: Ausprägungen des Anwendungsszenarios "Auftragsgesteuerte Produktion"

Es steht somit nicht mehr im Voraus vollständig fest, welche Kooperationspartner in Produktion eingebunden werden und in welcher Weise diese miteinander interagieren. Je nach Auftrag und Konfiguration eines konkreten Produktes, Serie oder Charge, können die Produktionsstandorte eigene Fertigungsfähigkeiten und Kapazitäten flexibel erweitern und die Fähigkeiten von externen Partnern in eigene Produktionsprozesse automatisiert einbinden. Andererseits können die Unternehmen, die freie Fertigungskapazitäten haben, diese anbieten und auf diese Weise die Auslastung eigener Maschinen erhöhen¹.

Ein Beispiel: eine Produktionslinie erhält von einem Lieferanten ein Werkstück mit Lochbohrung. Nun wurde der Lieferant gewechselt und es wurde ein Werkstück ohne Lochbohrung geliefert. Damit der Produktionsprozess fortgesetzt werden kann, ist dementsprechend eine neue Produktionsfähigkeit „Loch bohren“ erforderlich. Der Bedarf nach einer neuen Produktionsfähigkeit kann nicht mit eigenen Maschinen gedeckt werden. Also stellt die Produktions-

¹ <https://www.youtube.com/watch?v=rwcsa8489RY>

steuerung eine entsprechende Anfrage in einem Produktionsnetzwerk, zum Beispiel in dem diese eine Ausschreibung im Netzwerk initiiert. Andere Teilnehmer des Produktionsverbundes erhalten die Anfrage, evaluieren diese und können mit einem Angebot antworten. Anschließend wird das beste Angebot vom Initiator ausgewählt. Danach erfolgt der durch ein ähnliches Prozedere ausgewählte Transport des Werkstückes zum jeweiligen Dienstleister, die Erbringung der Dienstleistung „Bohren“ und schließlich der Rücktransport des gebohrten Werkstückes.

4. Skizzierung des Konzeptes

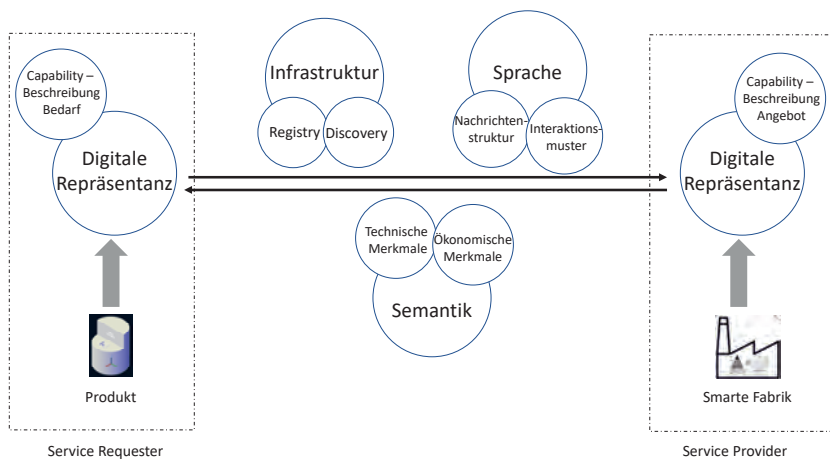


Bild 2: Bausteine, die für die Umsetzung von AGP notwendig sind

Das Ziel der unternehmensübergreifenden AGP ist, dass unterschiedliche Akteure befähigt werden, untereinander autonom Dienstleistungen zu verhandeln. Im Rahmen einer Verhandlung lassen sich zwei Rollen abstrahieren. Zum einen kann es einen Akteur geben, der einen materiellen oder immateriellen Bedarf hat, der durch Anforderungen ausgeprägt ist – der Service Requester (SR) oder auch Service Consumer. Zum anderen kann es auch einen Akteur geben, der diesen Bedarf durch Bereitstellung eines materiellen oder immateriellen Guts befriedigt – der Service Provider (SP). Im Kontext dieses Papers werden nur Dienstleistungen im Sinne produktionsnaher Fähigkeiten betrachtet.

Der Anwendungsfall unternehmensübergreifende AGP setzt hohe Anforderungen an Interoperabilität. Zur Umsetzung sind daher verschiedenartige Vereinbarungen zwischen den

Akteuren zu treffen und einzuhalten. Folgende Aspekte wurden identifiziert, aus denen sich Anforderungen an diese ableiten lassen (siehe Abbildung 2):

- **Digitale Repräsentanz** – Da Interaktionen zwischen Assets nur im virtuellen Raum möglich sind, bedarf es einer digitalen Repräsentanz dieser Assets in Form eines digitalen Zwillings.
- **Beschreibung der Dienstleistung** – Beide Rollen müssen das gleiche semantische und strukturelle Verständnis für die Beschreibung des Verhandlungsgegenstandes besitzen. Das betrifft auch die technischen und ökonomischen Merkmale einer Dienstleistung.
- **Infrastruktur** – Die Akteure müssen in der Lage sein sich in einer möglichst kurzen Zeit zu finden. Idealerweise kann ein SR vorselektieren, welche SP seine Anforderungen erfüllen. Außerdem müssen sie in der Lage sein, Informationen auszutauschen. Weitere Anforderungen sind typische IT-Aspekte wie Security, Vertraulichkeit, Verfügbarkeit, Integrität und Authentizität.
- **Sprache** – Die Akteure müssen die gleiche Sprache sprechen, sowohl in Bezug auf die Struktur als auch die Abfolge der Nachrichten

In den folgenden Abschnitten sollen zu diesen Aspekten Lösungsansätze in Form von I4.0-Bausteinen präsentiert werden oder zumindest Anforderungen an diese formuliert werden. Die Zusammensetzung ergibt ein Konzept für die Realisierung der unternehmensübergreifenden AGP.

4.1. Digitale Repräsentanz

Teilnehmer des Produktionsnetzwerkes/Verbundes werden als autonome entscheidungsfähige Einheiten aufgefasst. Diese sind im zukünftigen I4.0 Ökosystem gleichrangig und gleichberechtigt. Eine Ausprägung eines digitalen Zwillings ist die Verwaltungsschale (VWS), welche zusammen mit dem zugehörigen Asset eine I4.0-Komponente bildet. Die VWS ist ein zentraler Bestandteil der von Plattform Industrie 4.0 und IDTA verfolgten I4.0-Konzepte und besitzt ein standardisiertes Informationsmodell in Form von Submodels und eine Zugriffsschnittstelle (API) [4, 5]. Aufgrund des VWS-Metamodells können sie den tatsächlichen Aufbau, Struktur und interne Komplexität ihres Assets durch eine einheitliche Schnittstelle abstrahieren. Diese Schnittstelle ist die Beschreibung ihrer Fähigkeiten, bzw. Dienstleistungen, welche die VWS suchen oder den anderen Teilnehmern des Netzwerkes anbieten.

Um AGP umsetzen zu können, ist sicherzustellen, dass die I4.0-Komponenten eine gewisse Intelligenz und Autonomie besitzen. Die I4.0-Komponenten müssen an mehreren Stellen im Rahmen des Ausschreibungsverfahrens Entscheidungen treffen, z.B. muss ein SR entschei-

den, welches Angebot das Beste ist oder ein SP muss überprüfen, ob er in der Lage ist die Anforderungen des SR zu erfüllen. Daher wurden verschiedene Erscheinungsformen der Verwaltungsschale identifiziert, welche sich hauptsächlich in Hinblick auf deren Interaktionsmuster unterscheiden [6]. Im Rahmen dieses Konzepts wird die proaktive oder auch Typ-3-Verwaltungsschale betrachtet, da nur diese in der Lage ist Informationen mittels nachrichtenbasierter Kommunikation auszutauschen. Ein Grund dafür ist, dass die Spezifikationen der VWS nur ein Informationsmodell und eine Zugriffsschnittstelle definieren und somit die VWS im eigentlichen Sinne eine Ausprägung einer Datenstruktur darstellt.

Der angestrebte Anwendungsfall erfordert allerdings auch zusätzliche Software-Komponenten, die einen kontrollierenden Charakter besitzen, z.B. in Form eines Zustandsautomaten und somit ein autonomes Verhalten ermöglichen. Spricht man von einer proaktiven VWS, meint man also eine aktive Software-Komponente, welche in der Lage ist Peer-to-Peer mit anderen proaktiven VWS zu interagieren und eine VWS als Wissensbasis oder als Beschreibungsmittel für Interaktionsgegenstände zu nutzen (siehe Abbildung 3). Mögliche Architekturen und Ansätze der proaktiven VWS wurden bereits identifiziert und können auch agentenbasiert sein [6–8]. Dementsprechend können diese autonomen Einheiten die Rolle von Service Requester und die Rolle der Service Provider annehmen.

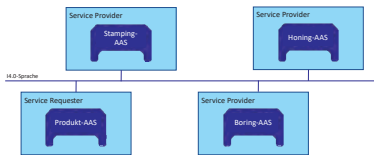


Bild 3: Proaktive AAS im I4.0-System

4.2. Sprache

Die Sprache zwischen den Akteuren werden durch die I4.0-Sprache und das Ausschreibungsverfahren nach VDI 2193 beschrieben [9]. Die VDI-Richtlinie 2193 definiert einen dreiteiligen Regelsatz für die I4.0-Sprache (siehe Abbildung 4). Die Struktur der Nachrichten ist in VDI 2193-1 definiert. Jede Nachricht besteht aus einem Frame und einem Interaktionselement. Der Frame enthält Attribute, die zum einen der Identifizierung (z.B. Sender, Empfänger) und zum anderen der Kontextualisierung dienen (z.B. Nachrichtentyp, ConversationID). Das Interaktionselement ist aus dem Metamodell der Verwaltungsschale gegriffen und stellt ein beliebiges Metamodellelement dar, z.B. Submodel oder SubmodelElement. Ein weiterer Bestandteil der I4.0-Sprache sind die Interaktionsprotokolle.

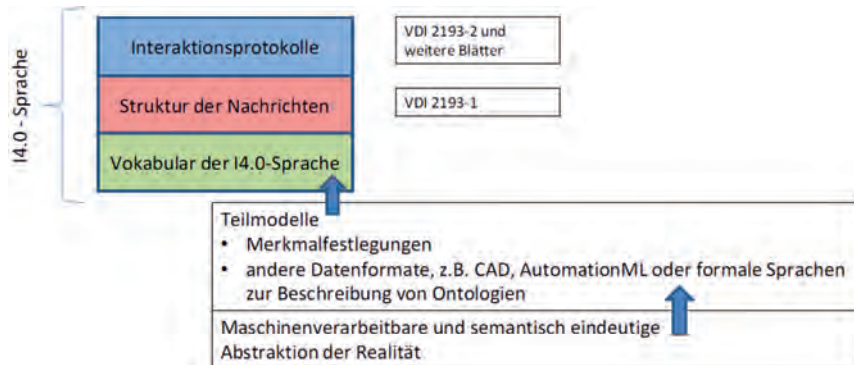


Bild 4: Dreiteiliger Regelsatz der I4.0-Sprache

Diese definieren den Ablauf eines Dialogs zwischen zwei oder mehreren I4.0-Komponenten. Ein bereits definiertes Interaktionsprotokoll ist das Ausschreibungsverfahren nach VDI 2193-2. Dieses basiert auf dem Contract Net Protocol der FIPA und legt die einzuhaltende Nachrichtensequenz fest, welche notwendig ist um eine Dienstleistung zwischen einem Service Requester und einem Service Provider zu verhandeln.

4.3. Beschreibung der Dienstleistung

Eine der in dieser Arbeit verfolgten Arbeitsthemen ist die Annahme, dass es möglich ist, die exakte Verwendung von Produktionsressourcen (einzelnen Maschinen oder gar Smarten Fabriken) als ihre „Fähigkeit“ (eng.: Capability) zu beschreiben. Eine Capability beschreibt die von einem Asset zu erfüllende Funktion, beschreibt, wozu ein Asset geeignet ist und definiert seine Verwendung. Weitere in dieser Arbeit getroffene Annahme ist, dass es möglich ist, eine generische Beschreibung wie ein Produkt zu produzieren ist, als eine Sequenz von dafür (von Produktionsmitteln) benötigten Capabilities darzustellen. Dies könnte wie ein für eine Instanz des Produktes in einem Workflow Management System erstellten Workflow vorgestellt werden. In einer Arbeitsgruppe der IDTA wird zurzeit ein von Fachexperten konsolidiertes Submodel Template für die semantisch eindeutige Beschreibung von Capabilities definiert.

Da die unternehmensübergreifende AGP eine Art der automatisierten Geschäftsprozesse darstellt, sind neben den Capabilities auch andere Merkmale und Aspekte maßgeblich, die über die Beschreibung rein technischer Fähigkeiten hinausgehen. Dazu zählen zum Beispiel die Angabe des Ortes und der Zeit der Erbringung der Fähigkeiten, Angaben zu Qualität und

eventueller fachlichen Zertifizierungen, mögliche Arbeitsbewertung durch bisherige Geschäftspartner etc. Diese zusätzlichen Merkmale werden im Kontext dieser Arbeit als kommerzielle Aspekte bezeichnet. Die um die Beschreibung der kommerziellen Aspekte ergänzte Capability-Beschreibung bildet eine Beschreibung der Dienstleistung (siehe Abbildung 5). Für das Gelingen der AGP empfiehlt es sich, eine einheitliche und möglichst Capability-unabhängige Beschreibung von kommerziellen Aspekten von Dienstleistungen zu definieren und darauf aufbauend ein konsolidiertes Submodel Template zu verabschieden.

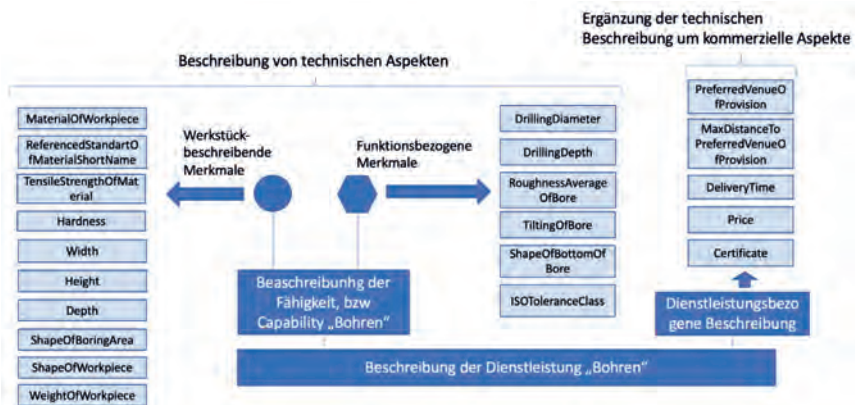


Bild 5: Beispielhafte Beschreibung der Dienstleistung "Bohren"

4.4. Infrastruktur

Die I4.0 – Sprache sowie die Verwaltungsschale sind wichtige Konstrukte, um die Visionen der AGP zu ermöglichen, jedoch reichen dafür allein nicht aus. Der andere wichtige Baustein ist die gemeinsame Infrastruktur oder Plattform, damit sich die Service Requester und Service Provider, bzw. ihre Verwaltungsschalen im Netzwerk finden und interagieren können.

Die entsprechenden Konzepte der Plattform Industrie 4.0 befinden sich zurzeit noch in der Entwicklung und werden oft auf eine Diskussion rund um Registry reduziert. Die diskutierten Konzepte basieren auf den klassischen zentralisierten Ansätzen, die die unabhängigen Akteure zusammenbringen und die Registrierung, das Identitätsmanagement, die Discovery- und Marktplatzfunktionalitäten übernehmen [5].

Aus Sicht der Autoren ist der zentralisierte Ansatz ein valider erster Schritt zur kurzfristigen Realisierung von Industrie 4.0-Anwendungsszenarien. Die langfristigen Visionen der Industrie 4.0 gehen jedoch vom dezentralen Charakter zukünftiger digitaler Ökosysteme aus. Diese sollen keine Komponenten beinhalten, die eine zentrale Monopolstellung gegenüber anderer

Ökosystem-Teilnehmer einnehmen können und deren Nichtverfügbarkeit den sicheren Betrieb des Gesamtsystems beeinträchtigen kann.

Die von Autoren dieses Beitrages durchgeführte Literaturrecherche und die in den relevanten Fachgremien geführten Diskussionen zeigen, dass es in der Community kein einheitliches Konzept dafür gibt, wie die angestrebte „Dezentralität“ der I4.0-Infrastruktur zu gestalten ist. Es lassen sich völlig entgegengesetzte Ansätze für die Gestaltung der dezentralen Infrastruktur identifizieren. Diese reichen von der Weiterentwicklung der zentralisierten Ansätze, nämlich der Annahme, dass die Unternehmen ihre eigenen unabhängigen firmenspezifischen Plattformen (z.B. für Registry, Discovery, Authentifizierung etc) aufbauen bis zu Ansätzen, die für den Aufbau von hochdynamischen ad hoc Kooperationsbeziehungen in P2P-Netzwerken ausgelegt sind und gänzlich ohne separate Instanzen auskommen, welche Registry- und Discovery-Funktionalitäten bereitstellen.

Aus Sicht der Autoren bietet diese Tatsache ausreichend Motivation, um in weiteren Forschungsaktivitäten Taxonomie und Gestaltungsprinzipien von dezentralen Infrastrukturen herauszuarbeiten und einzelne Typen auf die auftretenden Ausprägungen und Wechselwirkungen im Kontext der AGP zu diskutieren. Dies könnte einen wichtigen Beitrag der Plattform I4.0 für die Gestaltung von zukünftigen digitalen I4.0-Ökosystemen darstellen.

5. Deutschlandweiter Demonstrator der AGP

Dieses Kapitel beschreibt eine Erprobung der im Kapitel 3 diskutierten Ansätze und die dabei entstandene Ausprägung der von Plattform I4.0 und IDTA erarbeiteten Konzepte. Für die Überführung der AGP in die industrielle Praxis wird im Projekt „VWS vernetzt“ ein deutschlandweiter Demonstrator aufgebaut, mit dem Ziel das erste lauffähige VWS-basierte I4.0-Ökosystem zu etablieren.

Die Projektpartner beteiligen sich an dem gemeinsamen Vorhaben mit ihren eigenen Demonstratoren. Die fertigungstechnischen Anlagen der Projektpartner erhalten proaktive Verwaltungsschalen, welche diese Anlagen in einem I4.0-System repräsentieren. Zusammen mit proaktiven Verwaltungsschalen bilden die einzelnen Anlagen, die durch Deutschland verteilt, autonomen Dienstleistungseinheiten.

Das angestrebte Szenario ist in zwei Phasen aufgeteilt. In der ersten Phase (siehe Abbildung 7) gibt es im I4.0-System genau einen Service Requester – das Produkt. Das Produkt (siehe Abbildung 6) ist fiktiv, hat keine spezifische Verwendung und dient als Projektionsfläche für die Fähigkeiten der Demonstratoren der Projektpartner.

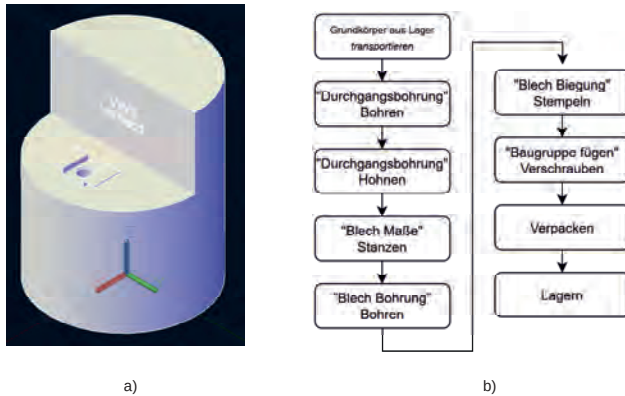


Bild 6: a) fiktives Produkt, b) Abfolge der zur Produktion benötigten Fähigkeiten

Das hat den Vorteil, dass weitere Teilnehmer mit anderen Fähigkeiten sich dem Projekt anschließen können, da das Produkt einfach und flexibel adaptiert werden kann.

Um dieses zu produzieren ist eine Menge verschiedener Produktionsfähigkeiten notwendig, z.B. bohren, honen, stempeln bis hin zu unterstützenden Fähigkeiten wie transportieren und verpacken. Der SR kennt die Reihenfolge dieser Fähigkeiten und die Anforderungen, die an diese bestehen. Über eine Webapplikation lässt sich schließlich auf den SR zugreifen und der Prozess initiieren.

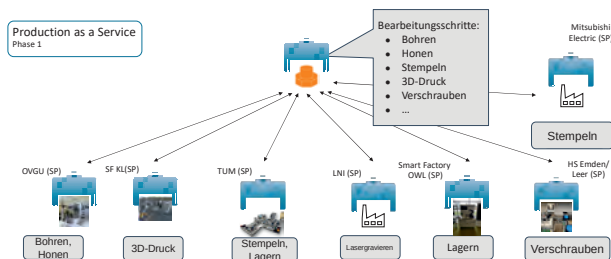


Bild 7: Phase 1 im deutschlandweiten Demonstrator

Im I4.0-System gibt es weitere autonome I4.0-Komponenten in der Rolle als Service Provider, welche durch die Demonstratoren der Projektpartner realisiert werden. Die Verwaltungsschalen der Service Provider kennen die Produktionsfähigkeiten der Demonstratoren, wie z.B. "Bohren", "Honen", "Laserschneiden", "3D-Drucken", "Schrauben", "Verpacken" usw.

und bieten diese den anderen Teilnehmern des digitalen I4.0-Ökosystems als Dienstleistung an.

Eine getroffene Vereinfachung ist, dass der Produktionsablauf starr im SR des Produkts implementiert ist. Im Sinne einer flexiblen AGP gibt es noch Forschungsbedarf, wie die Sequenzen der geforderten Capabilities generisch aus Produktbeschreibungen (z.B. aus CAD-Daten) erstellt werden. Außerdem werden derzeit keine geometrischen Daten in den Dienstleistungsbeschreibungen integriert. Das ist aber zwingend notwendig, damit die Positionen der Dienstleistungen am Werkstück verortet werden können.

Im zweiten Schritt erfolgt die Vernetzung der Demonstratoren aller Projektpartner untereinander, so dass diese zusätzlich zu der Rolle des SP die Rolle eines SR einnehmen und gegenseitig auf die Fertigungsfähigkeiten zugreifen können (siehe Abbildung 8).

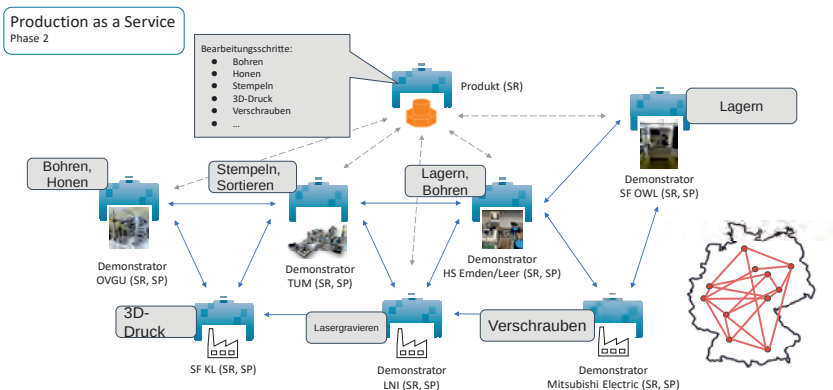


Bild 8: Phase 2 im deutschlandweiten Demonstrator

Ein Grund für einen solchen Rollenwechsel könnte sein, dass die eigentlich bereits verhandelte Dienstleistung bei Maschinenausfall nicht durchgeführt werden kann und daher der SP versucht eine Vertragsstrafe zu vermeiden.

Das Ziel ist es, einen Demonstrator eines offenen, dynamischen, lebendigen I4.0-Ökosystem aufzubauen. Damit soll ein wesentlicher Meilenstein auf dem Weg zur Verwirklichung der Visionen der Plattform I4.0 und Adoption der VWS gesetzt werden.

6. Zusammenfassung und Ausblick

Flexible Produktion der Losgröße 1, dynamische Wertschöpfungsnetzwerke und souveräne, gleichberichtigte Marktakteure sind die Visionen der Industrie 4.0. Ein Anwendungsfall, in

dem diese Visionen münden ist die unternehmensübergreifende Auftragsgesteuerte Produktion. Die Autoren ordnen die Ausprägungen der AGP ein und arbeiten Aspekte heraus, die für deren Umsetzung relevant sind. I4.0-Konzepte, wie die VWS, I4.0-Sprache und das Capability-Modell werden als Bausteine zur Realisierung dieser Aspekte identifiziert. Des Weiteren wird aufgezeigt, welche Anforderungen an Aspekten bestehen, zu denen noch bislang keine gängigen I4.0-Konzepte bestehen wie die Infrastruktur. Zusätzlich wird der deutschlandweite Demonstrator vorgestellt, welcher als offenes Testbed der Evaluation und Verfeinerung des vorgeschlagenen Gesamtkonzepts dient. Bestehende Lücken im Konzept wie die Zuordnung von CAD-Daten zu Dienstleistungen, Integration von Infrastruktur oder die Generierung von Produktionsablaufplänen aus Produktbeschreibungen müssen in weiteren Forschungsarbeiten geschlossen werden.

7. Referenzen

- [1] K. Bauer *et al.*, "Anwendungsszenario trifft Praxis: Auftragsgesteuerte Produktion eines individuellen Fahrradlenkers," Berlin, 2017.
- [2] D. Schulz, C. Lithoxopoulos, and L. Simora, "'Massenfertigung' mit Losgröße Eins," *atp*, vol. 63, no. 05, pp. 54–63, 2021.
- [3] I. Kovalenko, D. Tilbury, and K. Barton, "The model-based product agent: A control oriented architecture for intelligent products in multi-agent manufacturing systems," *Control Engineering Practice*, vol. 86, pp. 105–117, 2019.
- [4] Plattform Industrie 4.0 AG 1, "Details of the Asset Administration Shell: Part 1 - The exchange of information between partners in the value chain of Industrie 4.0 (Version 3.0RC01)," Berlin, 2021.
- [5] Plattform Industrie 4.0 AG 1, "Details of the Asset Administration Shell: Part 2 – Interoperability at Runtime – Exchanging Information via Application Programming Interfaces (Version 1.0RC02)," Berlin, 2021.
- [6] Alexander Belyaev and Christian Diedrich, "Aktive Verwaltungsschale von I4.0-Komponenten: Erscheinungsformen von Verwaltungsschalen," in *VDI-Berichte*, vol. 2351, *Automation 2019*, VDI Wissensforum GmbH, Düsseldorf: VDI Verlag, 2019.
- [7] S. Wein, Y. Dassen, C. Pallasch, T. Miny, S. Storms, and C. Brecher, "Konzept und Anwendung Autonomer Industrie 4.0-Komponenten auf Basis Agenten-basierter Ansätze," *at - Automatisierungstechnik*, vol. 69, no. 6, pp. 430–441, 2021.
- [8] A. Fay, F. Gehlhoff, M. Seitz, B. Vogel-Heuser, H. Baumgärtel, C. Diedrich, A. Lüder, T. Schöler, G. Sutschet and R. Verbeet, "Agenten zur Realisierung von Industrie4.0," VDI-Statusreport, Düsseldorf: VDI Verlag, 2019.
- [9] C. Diedrich *et al.*, "Sprache für I4.0-Komponenten: Semantik der Interaktionen von I4.0-Komponenten," in *Automation 2018 - Leitkongress der Mess- und Automatisierungstechnik*, pp. 235–248.

Verwaltungsschalen zum anlagenweiten Testen wandelbarer Produktionsanlagen

Stephan Schäfer, Dirk Schöttke, Thomas Kämpfe, Vasil Denkov,
Hochschule für Technik und Wirtschaft Berlin;
Bernd Tauber,
EAW Relaistechnik GmbH, Berlin

Kurzfassung

Das wesentliche Merkmal wandelbarer Industrieanlagen ist ihre Fähigkeit, sich auf verschiedene Anwendungsfälle flexibel konfigurieren lassen. Hierfür benötigen sie einen zuverlässigen Engineering-Prozess, der reproduzierbare Ergebnisse erzeugt. Das Vertrauen in den Prozess und die zu funktionalen Merkmale der eingebundenen Komponenten entsteht durch Testverfahren. Definitionen und Konzepte für Softwaretests werden in der ISO/IEC IEEE 29119 standardisiert. In IEEE 29119 Teil 2 unterscheidet die Norm in der aktuellen Version von 2021 bei dynamischen Tests innerhalb eines Testprojekts zwischen verschiedenen Testebenen (z. B. Unit-, Integration-, System- und Abnahmetests) oder zwischen bestimmten Testtypen (z. B. Leistungstests, Sicherheitstests und Funktionstests) [Ie21].

Während Unit-Tests im Wesentlichen auf einzelne Komponenten begrenzt sind, stehen Integrationstests, die Testverfahren für die Prüfung anlagenweit benötigter Fähigkeiten und Merkmale anbieten, zunehmend im Fokus. In diesem Beitrag wird daher eine Integrations-Testmethode vorgestellt, die anlagenweit eingesetzt werden kann. Der in diesem Beitrag vorgeschlagene Integrationstest verwendet die Verwaltungsschale als eine genormte Variante eines digitalen Zwillings.

1. Engineering – wandelbare Umgebungen

Laut [Ba19] ist eine Verwaltungsschale (Verwaltungsschale, engl. AAS) definiert als digitale Repräsentation eines physischen Gegenstands dem sogenannten Asset. Aus diesem Grund lässt sich die Verwaltungsschale auch als ein digitales Typenschild für einen physisch vorhandenen Gegenstand bezeichnen, welche basierend auf Metadaten als maschinenlesbare Selbstbeschreibung der entsprechenden physikalischen oder logischen Komponente dient und über eine technische Schnittstelle für Menschen, I4.0 Systemen und Komponenten sowie weitere Entitäten verfügbar [Zi22].

Gemäß Diskussionspapier: "Verwaltungsschale in der Praxis" besteht das Informationsmodell einer Verwaltungsschale aus den folgenden Elementen

- Metainformationen zur Verwaltungsschale, Asset, Teilmodelle, Concept Dictionary und Views,
- Ergänzende Metainformationen zur Verwaltungsschale, Asset, Teilmodelle, Concept Dictionary und Views,
- Asset-unabhängigen Teilmodelle und
- Asset-spezifischen Teilmodelle [VP21].

Dabei können Submodelle je nach Sicht, bestimmte Aspekte des entsprechenden Assets abbilden sowohl statische (Merkmale, Identifikationsdaten) als auch dynamische bzw. ausführbare Elemente (Operationen, Skills, Fähigkeiten) beinhalten.

Das im Folgenden beschriebene Engineering-Konzept fußt auf Verwaltungsschalen für die Beschreibung der Eigenschaften von Assets in wandelbaren Produktionsanlagen. Dabei ermöglichen die Verwaltungsschalen die dynamische Anpassung der Anlagen an sich ändernde industriellen Prozesse. Unter dynamischer Anpassung werden hier unter anderem die Konfiguration, Parametrisierung, Auswahl, Buchung, Integration und Ersetzung von Ressourcen, Zuordnung und Überprüfung der Durchführbarkeit von Fähigkeiten ohne oder mit einer möglichst unerheblichen Unterbrechung des Produktionsprozesses verstanden. Bei dem hier zugrundeliegenden Engineering-Prozess wird davon ausgegangen, dass eine Menge von Ressourcen bereits vorhanden und verfügbar ist. Darüber hinaus wird die Existenz eines Produktionsrezeptes bzw. einer Konfiguration vorausgesetzt, aus welcher die vom Prozess angeforderten Fähigkeiten abgeleitet werden können. Eine Schlüsselrolle für die Umsetzung nimmt die möglichst standardisierte maschinenlesbare Beschreibung der vom Prozess angeforderten und von den Ressourcen angebotenen Fähigkeiten ein. Laut [Ba22, S.13] und [PSB13] eignen sich dafür beispielsweise Ontologien als formale Beschreibungsformen von Konzepten und deren Beziehungen zueinander.

2. Phasen des Engineerings

Der Engineering-Prozess einer Industrie 4.0 Anlage lässt sich nach [Ba22] durch drei automatisierte Phasen beschreiben. In der ersten Phase, hier der Capability Checking Phase, werden die zur Herstellung des entsprechenden Produktes notwendigen Fähigkeiten mit den bereits verfügbaren Fähigkeiten abgeglichen. Sind alle für die Funktion benötigten Fähigkeiten verfügbar, so erfolgt eine simulative und systematische Überprüfung ihrer Durchführbarkeit.

Die folgende Abbildung gibt einen Überblick über die zuvor genannten drei Phasen des Engineering-Prozesses.

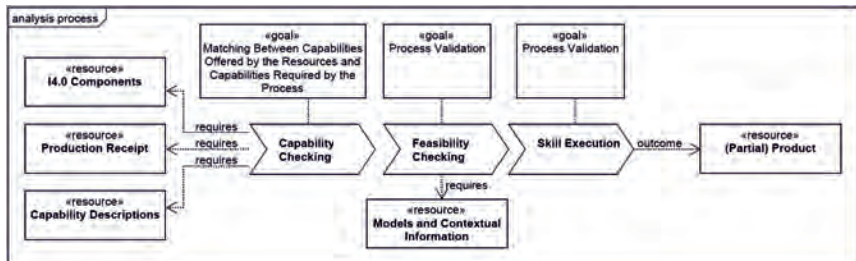


Bild 2: Phasen Engineering-Prozess - Details

Diese Phase umfasst daher Struktur- und Eigenschaftentests und prüft, das Vorhandensein benötigter Merkmale oder Eigenschaften. Ähnlich einem digitalen Marktplatz können hier, je nach Ausgestaltung weitere Ressourcen hinzugebucht werden, indem sie sich am System anmelden und Ihre Merkmale oder Eigenschaften dem System bekanntgeben.

Nach einem erfolgreichen Abgleich der benötigten und angebotenen Ressourcen findet die zweite Phase, der Feasibility Check statt.

Um zuverlässige und robuste Prüfprozesse zu erzeugen, wird in dieser Phase das Programmierparadigma des „design by contract“ eingeführt. Software ist dann zuverlässig, wenn sie ihre Arbeit wie spezifiziert ausführen kann, und sie ist robust, wenn sie abnormale Situationen beherrscht [GR05].

Für den Prüfprozess werden dazu formale Verträge zur Verwendung von Schnittstellen verhandelt. Das „design by contract“ Paradigma gewährleistet, dass jede Prüfroutine Konditionen als Bestandteile von Vertragsspezifikationen aufweist [AM18]. Diese Konditionen können sowohl statische Definitionen, als auch Anforderungen an dynamisches Verhalten beinhalten.

Bei den Konditionen wird zwischen drei verschiedenen Typen unterschieden, nämlich den Vorbedingungen, den Nachbedingungen und den Invarianten. Die Vorbedingungen enthalten Zusicherungen, die seitens der zu verwendenden Ressource z.B. einer Verwaltungsschale einzuhalten sind. Die Nachbedingungen enthalten Zusicherungen, welche der Initiator des Prüfprozesses einzuhalten hat.

Invarianten sind logische Aussagen, die für alle Prüfzenarien über den gesamten Objektlebenszyklus hinweg gelten können. Sie können sowohl vom Prüfprozess als auch von den Ressourcen festgelegt werden. Zunächst findet die Überprüfung der Vorbedingungen

²<https://github.com/admin-shell-io/submodel-templates/tree/main/development/Simulation/1/0>

statt. Dadurch wird sichergestellt, dass die Prüfumgebung für die Ausführung der entsprechenden Fähigkeiten vorbereitet ist.

Als Beispiel dient die Bewegung eines Zielobjektes von Position A nach Position B mithilfe eines Roboterarms. Eine gültige Vorbedingung wäre, dass sich das Zielobjekt vor der Ausführung der Fähigkeit auf Position A befindet, siehe [Ba22, S.6] und vgl. mit nachfolgender Abbildung. Da diese Konditionen in der Regel bestimmte Kontextinformationen beinhalten, z.B. die Position des Zielobjekts, werden als Eingangsinformationen für die Durchführung des sogenannten Feasibility Checks Beschreibungsformen der Anlage (Modelle) benötigt [Ba22, S.6]. Nach der Validierung der Vorbedingungen überprüft die Simulation die Invarianten. Als Invarianten gelten alle Subprozesse oder Eigenschaften, die bei veränderten Bedingungen unverändert bleiben.

In Zusammenhang mit dem Beispiel wäre als Invariante das Halten des Objekts auf dem Weg nach der Zielposition zu definieren [Ba22, S.6]. Am Ende des Feasibility Checks werden die Nachbedingungen verifiziert. Durch diese wird die erfolgreiche Durchführung der entsprechenden Fähigkeit simulationstechnisch bestätigt. In Abhängigkeit vom Produktionsrezept kann darauf basierend die Simulation eines weiteren Dienstes beginnen. Hierfür gilt als Voraussetzung, dass die Vorbedingung erfüllt wurde. Diese Voraussetzung könnte man daher als Nachfolgebedingung für den weiteren Prüfablauf definieren.

Nach dem Feasibility Check findet die Ausführungsphase in der obigen Abbildung „skill execution“ genannt, statt. Diese erfordert die Parametrisierung bzw. Kalibrierung der Ressourcen und das Aufrufen der festgelegten Fähigkeiten anhand des Produktionsrezeptes. Im Rahmen der Prozesssequenz werden die Bedingungen ebenfalls überprüft. Dieser Ansatz führt dazu, dass die für jeden Schritt notwendigen Ressourcen zur Laufzeit und in Abhängigkeit von den Prozessanforderungen, ihrer Verfügbarkeit und angebotenen Diensten festgelegt, gebucht, konfiguriert bzw. parametrisiert und orchestriert werden können.

Das von der I4.0 Plattform vorgestellte Engineering-Konzept ermöglicht eine dynamische Reorganisation, also die Wandelbarkeit eines Produktionsprozesses. Während die Umsetzungsstrategie nicht vollständig definiert ist und sich aufgrund ihrer Allgemeingültigkeit durch eine zeitintensivere Umsetzung charakterisiert, lässt sich diese auf Basis der eingeführten Abstraktionen als zukunftsicher und vom konkreten Produkt, Prozess oder Ressource unabhängig bezeichnen.

Die folgende Abbildung schlägt unter Verwendung der UML- Package Notation eine mögliche Softwarestruktur für den automatisierten Prüfprozess vor.

Dabei finden Verwaltungsschalen als Repräsentanten ihrer Assets in großem Umfang Verwendung.

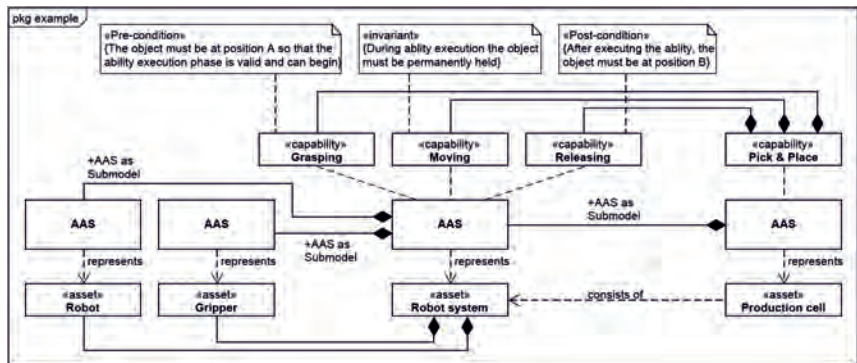


Bild 3: Beispiel Pick&Place – allg. Struktur (angelehnt an [Ba22])

Die Abbildung bildet die Struktur der im Vorhaben näher untersuchten Prüfumgebung ab. Dabei orientiert sich der Prüfprozess an den im Labor vorhandenen Ressourcen. Der Ansatz ist allerdings so allgemein gewählt, dass sich dieser mit sehr geringem Aufwand auf alternative Produktionsumgebungen übertragen lässt.

3. Fähigkeiten

Da dem Begriff Fähigkeit im Rahmen der Wandlungsfähigen Produktion eine besondere Bedeutung zukommt, soll dieser im Zusammenhang mit der Verwaltungsschale näher erläutert werden. Der Begriff "Fähigkeit" kann im Englischen mit Capability, Skill und Service übersetzt werden.

Im Allgemeinen wird unter einem Skill eine konkrete u.a. vom Asset abhängige Realisierung einer Aktivität bzw. einer gekapselten Funktionalität eines Produktionssystems auf dem Shop-Floor verstanden [Ma22]. Zu beachten ist, dass der Skill in einer direkten Korrelation zum konkreten Asset steht bzw. eine plattformspezifische Umsetzung einer Fähigkeit darstellt. Da dieselbe Aktivität auf unterschiedliche Arten umgesetzt bzw. implementiert werden kann, lässt sich ergibt sich, dass einem Skill mehrere Realisierungen zugeordnet werden können. Als Realisierungsvarianten werden in [Ba22, S.15] Zustands- und Auslösevariablen, Operationen bzw. Funktionen bei den prozeduralen und Methoden in der OOP sowie Funktionsblöcke und semantische Protokolle ausgewiesen. Das Dokument definiert den Funktionsblock als eine mögliche Option für die Umsetzung eines komplexen Skills.

²<https://github.com/admin-shell-io/submodel-templates/tree/main/development/Simulation/1/0>

In [Ba22, S.4] definiert die Plattform Industrie 4.0 als Fähigkeit im Sinne der „Capability“ die implementierungsunabhängige Beschreibung der Funktion einer Ressource, eine bestimmte Wirkung in der physischen oder virtuellen Welt zu erzielen.

Eine zusammengesetzte Fähigkeit lässt sich in mehreren atomaren Capabilities zerlegen.

Da der Engineering-Prozess Industrie 4.0 fähigkeitsorientiert ist, stellt die Verwaltungsschale ein geeignetes Werkzeug dar, um oben beschriebene Abstraktion entsprechend umsetzen. Auch wenn derzeit kein anlagen- und domänenübergreifendes Metamodell für die Erstellung von Fähigkeiten und derer Ontologien definiert ist, lassen sich für die Umsetzung der Skills oder auch Capabilities innerhalb der Verwaltungsschale bestimmte Anforderungen formulieren [Ma22]. Im Kontext der Verwaltungsschalen dienen Fähigkeiten, Merkmale, Operationen und Ereignisse in Form von Submodells oder Submodell-Elementen als Bestandteile des Software-Vertrags.

4. Testumgebung – Einordnung und Strategien

Während das RAMI 4.0 Modell zu einer Komplexitätsreduktion führt und die wichtigsten Aspekte der Industrie 4.0 Vision in der Form einer Schichtenarchitektur veranschaulicht, beschreibt es keine Methoden zur Validierung der korrekten Funktionsweise der bestehenden Anlagenressourcen [Ho18].

Nur mit einem ganzheitlichen und durchgängigen Ansatz der Qualitätssicherung in der Softwareerstellung und Testung können Fehlfunktionen vermindert bzw. vermieden werden.

Um bspw. zeitnah auf Systemänderungen reagieren zu können, müssen Teilsysteme du Komponenten möglichst automatisiert und ohne die Einbeziehung der Assets die Testszenarien durchlaufen. Hierfür ist die variable Abstimmung des Testszenarios mit den gestellten Anforderungen ein geeigneter Ansatz. Wurde beispielsweise in der Phase der Anforderungsermittlung nur ein funktionsbezogener Test vereinbart, so muss die Testumgebung in Form eines zugeordneten digitalen Zwillings oder Simulationsmodells auch nur dies ermöglichen.

Damit eine automatisierte Testdurchführung ohne zusätzliches Engineering erfolgt sind weitere Anforderungen an die Umgebung zu erfüllen. Eine Anforderung ist die lose Kopplung der Verwaltungsschalen untereinander unter Verwendung definierter Schnittstellen. Eine weitere Anforderung ist die automatisierte Erzeugung von Verwaltungsschalen als Repräsentanten von Assets.

Im Rahmen des BMBF geförderten Projektes „OpenBasys 4.0“ wurde beispielsweise ein Verwaltungsschalengenerator für speicherprogrammierbare Steuerungen realisiert [Sc21a]. Eine Verwaltungsschale kann nicht nur ein Asset, sondern darüber hinaus auch einen

Prüfablauf-Koordinator, welcher die Testdurchführung koordiniert und protokolliert, repräsentieren. Mit diesen Erweiterungen lassen sich automatisierte Tests bei jeder Änderung mit geringem Aufwand durchführen, so dass ein niederschwelliger Einstieg in die Prüfung möglich ist. Darüber hinaus die Prüfungsumgebung nachvollziehbare Ergebnisse mit dokumentierter Qualität.

5. Automatisiertes Testen mit Verwaltungsschalen

Um zu verifizieren, ob die gestellten Anforderungen an die Wandelbare Umgebung vollständig und widerspruchsfrei erfüllt werden, sind entsprechende Testszenarien notwendig. Diese lassen sich aber mit überschaubarem Aufwand nur semi- oder besser vollautomatisiert durchführen. Für das semi- oder vollautomatisierte Testen ist ein geeigneter Engineering-Prozess erforderlich, siehe Abb. 5.

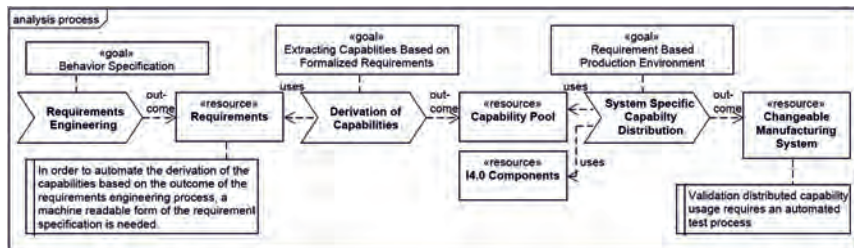


Bild 5: Kontext – Engineering Prozess

Dieser beginnt mit einer Anforderungsanalyse. Das Ziel des Requirement Engineerings liegt in der Spezifikation des konkreten Verhaltens der Anlagekomponenten in Zusammenhang mit dem benötigten Anwendungsfall. Die zum Verhalten korrespondierenden Voraussetzungen wie Pre- und Postkonditionen sowie Invarianten und Ressourcencharakteristiken müssen ebenfalls spezifiziert und quantisiert werden, damit sie messbar sind. Der erste Schritt liefert als Ergebnis eine Anforderungsspezifikation, welche das konkrete Verhalten der zu integrierenden Ressourcen detailliert beschreibt.

Zur Automatisierung des Engineering-Prozesses und als Voraussetzung für die zweite Phase ist eine maschinenlesbare Form der Anforderungsliste erforderlich. Ein beispielhaftes Verfahren für eine solche Formalisierung auf Basis von UML und einer Requirement Description Language ist in [MR16] vorgestellt.

Aus den formalisierten Anforderungen werden mithilfe eines Engineering-Werkzeugs die entsprechenden Fähigkeiten abgeleitet. Die abgeleiteten Anforderungen werden aus

²<https://github.com/admin-shell-io/submodel-templates/tree/main/development/Simulation/1/0>

Wiederverwendbarkeitsgründen in einer Bibliothek abgelegt. Da der hier beschriebene Engineering-Prozess dient als Konzept für die Realisierung wandelbarer Industrieanlagen dient, werden im Folgenden drei typische Szenarien betrachtet. Der allgemeine Zusammenhang ist in Abb. 6 exemplarisch dargestellt. Im ersten Szenario wird sowohl das Asset als auch die zugehörige Verwaltungsschale vom Lieferanten aufbereitet, validiert und bereitgestellt. In diesem Szenario ist der Lieferant für die Qualität der Verwaltungsschale und die Übereinstimmung des Produktes mit der entsprechenden Leistungsbeschreibung verantwortlich. Die Bereitstellung der Ressource zusammen, die als Basis für die Verifikation des Gesamtproduktes verwendet wurde, kann sich an dieser Stelle als sinnvoll erweisen.

Die Plattform Industrie 4.0 unterscheidet zwischen passiven ohne Kommunikationsschnittstelle vom Typ 1, reaktiven Verwaltungsschalen mit passiver Kommunikationsschnittstelle vom Typ 2 und proaktiven Verwaltungsschalen mit aktiver Kommunikationsschnittstelle Typ 3) [MI21].

Handelt es sich um eine passive Verwaltungsschale für ein aktives Asset, dann erfordert die vollständige Integration der Ressource nicht nur die semantische Beschreibung der vom Asset angebotenen Dienste, sondern auch die Ergänzung der Verwaltungsschale um eine Kommunikationsschnittstelle, die einen Zugriff zu den entsprechenden Funktionalitäten gewährleistet. In diesem Zusammenhang wird auch über eine vom Anwender vorzunehmende Typ 1 zu Typ 2 Transformation gesprochen, welche fachspezifische Kompetenzen seitens des Nutzers erfordert und sowohl zeit- als auch kostenineffizient ist. Um solche Entwicklungs- bzw. Anpassungsphasen nach der Lieferung zu vermeiden, kann der Anwender dem Lieferanten mitteilen, welcher Typ der Verwaltungsschale für den

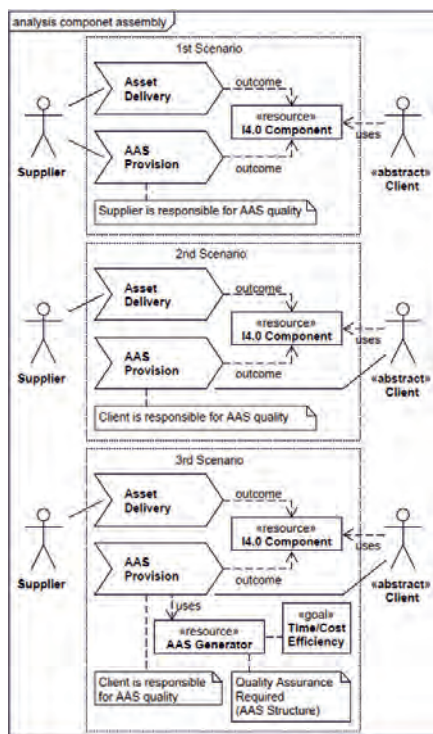


Bild 6: Drei typische Szenarien der Verwaltungsschalengenerierung

entsprechenden Anwendungsfall benötigt wird.

Bei dem zweiten Szenario wird vom Lieferanten nur das Asset zur Verfügung gestellt. Die Realisierung der Verwaltungsschale liegt in der Verantwortung des Anwenders und verlangt anwenderseitige Fach- und Engineering-Aufwand, sowie Ressourcen im Rahmen des Erstellungsprozesses. Der Anwender muss auf Basis der vom Lieferanten bereitgestellten technischen Beschreibung eine Verwaltungsschale Typ 1 erstellen und, wenn eine Kommunikationsschnittstelle erforderlich ist, daraus eine Typ 2 Verwaltungsschale erzeugen. Hierfür werden standardisierte Prozesse und eine entsprechende Toolchain benötigt. Dieses Szenario basiert auf der manuellen Spezifikation der Verwaltungsschalenstruktur. Als Werkzeug für die manuelle Erstellung von Typ 1 Verwaltungsschalen in der Form einer JSON- oder XML-Datei dient unter anderem der AASX Package Explorer vergl. hierzu [AA22].

Nach der Erstellung der Verwaltungsschale ist eine Validierung der syntaktischen Korrektheit, Vollständigkeit und der Übereinstimmung der Verwaltungsschale mit den Charakteristiken des Assets erforderlich. Die Qualität der Verwaltungsschale-Struktur lässt sich mittels Strukturtests nachweisen. Die Spezifikation und die Dokumentation der Testszenarien gehören ebenfalls zu den Aufgaben des Integrators, wobei zum Zweck der Qualitätssicherung eine Beschreibung der Testanforderungen erforderlich ist. Aufgrund des hohen Engineering-Aufwands im Rahmen des manuellen Verwaltungsschale-Erstellungsprozesses stellt zweite Szenario ebenfalls eine ressourcenineffiziente Alternative dar.

Das dritte Szenario basiert auf dem zweiten Szenario und unterscheidet sich in der automatisierten Erstellung aktiver Verwaltungsschalen unter Einsatz eines Verwaltungsschalengenerators. Als Basis des Generierungsvorgangs dient die Existenz einer passiven Verwaltungsschale, die entweder vom Lieferanten bereitgestellt oder vom Anwender realisiert werden kann. Die Transformation der passiven in einer aktiven Verwaltungsschale erfolgt dynamisch, wobei dafür bestimmte Softwareressourcen erforderlich sind. Während dieses Szenario als eine zeit- und langfristig kosteneffiziente Alternative betrachtet werden kann, besteht hier die Herausforderung einen vertrauenswürdigen Generierungsvorgang zu gewährleisten, dessen Ergebnisse das Asset aus Struktursicht fehlerfrei abbilden.

6. Automatisiertes Testen mit Verwaltungsschale

Im Folgenden wird eine generische Methode zum Testen von Anlagenkomponenten anhand eines beispielhaften Anwendungsszenario vorgeschlagen. Die Methode setzt den Einsatz eines Verwaltungsschale-Generators voraus, umfasst die Durchführung von Struktur- und Verhaltenstests und kombiniert traditionelle Softwaretestmethoden mit der Verwendung von Verwaltungsschalen als digitale Repräsentanten ihrer Assets.

²<https://github.com/admin-shell-io/submodel-templates/tree/main/development/Simulation/1/0>

Schritt 1 – Generierungsvorgang

Für den Generierungsvorgang wird zunächst eine passive Verwaltungsschale benötigt, die u.a. Strukturinformationen beinhaltet. Als Strukturinformationen dienen sowohl die allgemeine Struktur der Verwaltungsschale mit ihren Submodellen und die Anordnung der einzelnen Elemente jedes Submodells. Jedes Element, unabhängig davon, ob es sich um ein Submodell, ein Property oder eine Operation handelt, gilt im Kontext des Generierungsvorgangs als Strukturinformation.

Des Weiteren kann der passive Vertreter auch eine Beschreibung des Kommunikationskanals und einen Zustandsmaschinen-Deskriptor beinhalten. Die Beschreibung spezifiziert alle zum Aufbauen der Kommunikationsbeziehung zwischen der aktiven Verwaltungsschale und dem Asset notwendigen Informationen. Als Beispiele dienen das erforderliche Kommunikationsprotokoll, die Identifikationsadresse des Assets und das Zeitfenster für den Verbindungsaufbau. Die Existenz des Deskriptors entscheidet den Ablauf des Testprozesses. Ist eine Zustandsmaschine beispielsweise Führungskomponente mit dem Asset assoziierbar, muss diese im Rahmen des Testvorgangs validiert werden.

Der Verwaltungsschalengenerator ist derzeit nur für Assets geeignet, die als OPC UA Server fungieren. In Abhängigkeit von der Existenz des Zustandsmaschine-Deskriptors können im Adressraum des Servers die Schnittstellen bzw. die Variablenliste der Zustandsmaschine zusammen mit weiteren allgemeinen und ergänzenden Variablen instanziiert werden. Zum Beginn des Generierungsprozesses wird die Kommunikationsbeziehung zum Asset aufgebaut. Im nächsten Schritt analysiert der Generator den Adressraum des OPC UA Servers. Als Grundlage dienen die vorher instanziierten Schnittstellen der SPS-Applikation in Form von Submodellen. Darauf folgend wird mittels der BaSyx.NET SDK eine aktive Verwaltungsschale mit einem HTTP-Kommunikationskanal auf Basis der Strukturinformationen erstellt und die Submodelle innerhalb der Verwaltungsschale instanziiert. Abb. 7 veranschaulicht den Generierungsvorgang.

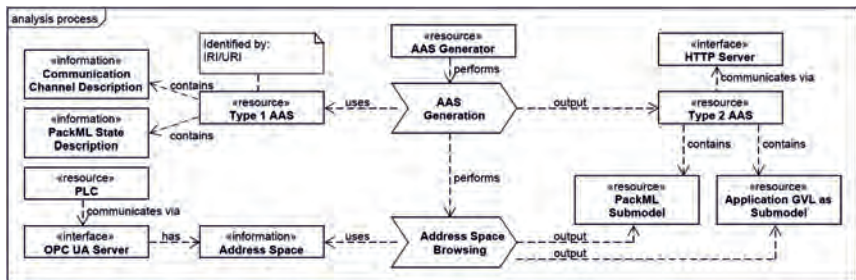


Bild 7: Prozesse Generierungsvorgang

Schritt 2 – Strukturtests

Nach dem Generierungsvorgang erfolgt die Durchführung von Strukturtests. Der Fokus auf Strukturtests lässt sich darüber motivieren, dass die potenzielle Verifikation eines Verwaltungsschale- oder Submodell-Elements, z.B. ein Sub-modell oder eine Operation, das Zugreifen bzw. die Existenz der übergeordneten Informationseinheit erfordert. Ist diese nicht vorhanden, so ist das Testobjekt nicht erreichbar. Dann gilt der Strukturtest als nicht bestanden. Ziel dieses Testes ist, dass die generierte Verwaltungsschale mit der Spezifikation des Assets übereinstimmen.

Im Rahmen der Testdurchführung wird die Struktur der passiven Verwaltungsschale mit der Struktur der aktiven Verwaltungsschale vom Testkoordinator verglichen. Wird eine Übereinstimmung erkannt, so lässt sich die Gültigkeit des Generatorergebnisses mittels einer Kennzeichnung bestätigen. Alle gekennzeichneten Verwaltungsschalen werden zur nächsten Testphase zugelassen.

Zur Vereinfachung des Softwareentwicklungsprozesses und als Voraussetzung für die Testdurchführung lässt sich eine Methode zur Erstellung von Templates von Steuerungsprogrammen konzipieren. Die Anwendung der Methode erfordert die Nutzung einer passiven Verwaltungsschale in der Form einer JSON oder XML-Datei. Aus der Verwaltungsschale müssen durch geeignete Algorithmen allgemeine Softwareeinheiten wie z.B. eine globale Variablenliste oder ein Funktionsblock- bzw. Funktionssignaturen als strukturierter Text gemäß IEC 61131-3 abgeleitet werden. Basierend auf den globalen Variablen und den generierten Funktionsschnittstellen kann der Programmierer die konkrete Implementierung vornehmen. Für den nachfolgenden Einsatz des Generators ist die Existenz einer Funktionsblocksignatur erforderlich, welche für die Identifizierung der Kommunikationspartner verantwortlich und mit einem Anmeldeprozess assoziierbar ist.

²<https://github.com/admin-shell-io/submodel-templates/tree/main/development/Simulation/1/0>

In [Sc21b] ist das exemplarische Design eines Verwaltungsschale-Generierungsvorgangs vorgestellt worden, wobei der eingesetzte Generator durch eine proaktive Verwaltungsschale repräsentiert wird und aktive digitale Vertreter produziert. Zum Zweck der Realisierung der Kommunikationsbeziehungen findet die Generierung nach einer Anmeldung des Controllers bzw. Assets bei dem Generator statt. Die SPS initiiert den Generierungsvorgang, indem sie einen RESTful-Aufruf mit den entsprechenden Parametern durchführt. Zu diesen zählt die ID der zu erstellen-den Verwaltungsschale, die IP-Adresse der Steuerung und die Portnummer ihres zugeordneten OPC UA Servers. Die Realisierung der Kommunikationsbeziehungen Abb.10 bildet die Grundlage für die Integration eines Testvorgangs.

Nach dem Kommunikationsaufbau lässt sich einen Vergleich zwischen dem Informationsmodell des OPC Servers und der Typ 1 Verwaltungsschale durchführen, welcher als Strukturtest während der Laufzeit des Generierungsprozesses dient. Stimmt die Struktur der passiven Verwaltungsschale mit dem Informationsmodell des Servers überein, so gilt der Strukturtest als bestanden. Wird ein Unterschied zwischen den Strukturen identifiziert, so kann die automatisierte Erstellung entsprechend abgebrochen werden. Auf diese Weise lässt sich eine automatisierte Strukturvalidierung in Programmlaufzeit durchführen. Voraussetzungen für die Verwaltungsschale-Generierung und die Testdurchführung sind dann, dass das Asset als Teilnehmer eines OPC UA Netzwerks gilt und vom Informationsmodell seines OPC UA Servers mit einer hohen Genauigkeit abgebildet wird. Nach dem Sicherstellen der Übereinstimmung des Informationsmodells mit der passiven Verwaltungsschale ist die Überprüfung der Identität des erstellten Template durch geeignete Verifizierungsmethoden erforderlich. Werden im Laufe der Generierung keine Strukturunterschiede zwischen dem Informationsmodell, dem Template und der passiven Verwaltungsschale softwaretechnisch festgestellt, so kann eine aktive Verwaltungsschale mit einer HTTP-Schnittstelle vom Generator produziert werden.

Nach einem erfolgreichen Generierungsvorgang gelten die integrierten Strukturtests als bestanden und die Funktionsschnittstellen als verifiziert. Im nächsten Schritt ist eine Validierung des Kommunikationskanals erforderlich. Als realistischer Ansatz dient hier die Durchführung von Stress- und Shutdown- bzw. Restart-Tests. Dies gilt als Voraussetzung für die Validierung der implementierten Funktionalitäten über die Verwaltungsschale. Geeignete Ansätze dafür sind die Testmethoden der Softwareentwicklung.

Schritt 3 – Testen der Zustandsmaschine

Wurde ein Deskriptor der Zustandsmaschine im Rahmen des Generierungsvorgangs erkannt, so muss ihre korrekte Funktionsweise verifiziert werden. Zum Zweck der Verifikation kann der

Koordinator jeder gekennzeichneten Verwaltungsschale, dessen Asset auf Basis einer Zustandsmaschine operiert, Kommandos senden, welche zu Transitionen zwischen den unterschiedlichen Zuständen führen. Anhand der Zustandswechsel, lassen sich die Anforderungen verifizieren.

Da mehrere Komponenten zustandsorientiert von einer Verwaltungsschale, die ebenfalls über eine Zustandsmaschine verfügt, betrieben werden können, ergibt sich die Frage wie der Testprozess aussieht. Als grundlegende Voraussetzung für die Testdurchführung wird ein hierarchisch priorisierter Zustandsautomat zugrunde gelegt. Darauf aufbauend kann jede einzelne Zustandsmaschine nach dem Bottom-Up Prinzip, sowie durch Trigger und Auswerten von Zustandsübergängen seitens des Testkoordinators getestet werden.

7. Fazit

Der Beitrag stellt ein standardisiertes Prüfverfahren für die automatisierte Validierung der funktionalen Eigenschaften der Anlagenkomponenten in wandelbaren Produktionsumgebungen vor. Das Verfahren ist unabhängig von der Simulationsumgebung. Automatisierte Tests lassen sich bei jeder Änderung mit geringem Aufwand durchführen. Sie liefern nachvollziehbare Ergebnisse mit dokumentierter Qualität. Automatisierte Tests ermöglichen darüber hinaus robuste und wiederverwendbare Anlagenkomponenten.

Auch wurde gezeigt, wie dieser systematische Ansatz für das Testen von Anlagen mit Hilfe eines Generierungsprozesses der Verwaltungsschale durchgeführt werden kann. Dies kann in einen ganzheitlichen Engineering-Ansatz integriert werden. Der Ansatz reduziert nicht nur den Engineering-Aufwand, sondern verringert auch die Fehlermöglichkeiten durch Einführung eines definierten Entwicklungsprozesses und die Prüfung der Verwaltungsschalenstruktur. Dieser Ansatz eröffnet eine Vielzahl weiterer Forschungsthemen, wie die formale Verifikation des Generierungsprozesses und die Zertifizierung von Verwaltungsschale auf Basis der durchgeführten Tests.

²<https://github.com/admin-shell-io/submodel-templates/tree/main/development/Simulation/1/0>

Literatur:

- [AA22] C# based viewer and editor for the Asset Administration Shell – GitHub, online unter: <https://github.com/admin-shell-io/aasx-package-explorer>, aufgerufen am 20.05.2022
- [AM18] A. Algarni and K. Magel, "Toward Design-by-Contract Based Generative Tool for Object-Oriented System," 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS), 2018, pp. 168-173, doi: 10.1109/ICSESS.2018.8663719
- [Ba22] Bayha, A.; Bock, J.; Boss, B.; Diedrich, C.; Malakuti, S.: Describing Capabilities of Industrie 4.0 Components, Joint White Paper between Plattform Industrie 4.0, VDI GMA 7.20, BaSys 4.2./, 2022.
- [Ba19] Bader, S. and Barnstedt, E. and Bedenbender, H. and Billman, M. and Boss, B. and Braunmandl, A.,
Details of the Asset Administration Shell Part 1--The Exchange of Information between Partners in the Value Chain of Industrie 4.0, Federal Ministry for Economic Affairs and Energy (BMWi), Berlin, 2019
- [Gl18] Glock, T.; Sillman, B.; Kobold, M.; Rebmann, S.; Sax, E.: Model-based validation and testing of industry 4.0 plants. In: 2018 Annual IEEE International Systems Conference (SysCon). KS-Energy-Verlag, S. 1– 8, 2018, isbn: 9783945622018, url: <https://books.google.de/books?id=1vcBrgEACAAJ>.
- [GR05] R. C. Guerrero, F. de la Rosa Troyano and S. del Pozo Hidalgo, "Diagnosis of Inconsistency in Contracts using Design by Contract," in IEEE Latin America Transactions, vol. 3, no. 1, pp. 137-142, March 2005, doi: 10.1109/TLA.2005.1468673
- [Gr13] Grünfelder, S.: Software-Test für Embedded Systems: Ein Praxishandbuch für Entwickler, Tester und technische Projektleiter. dpunkt.verl., Heidelberg, 2013, isbn: 3864900484.
- [Ho18] Homann: Testen vernetzter Systeme für Industrie 4.0./, 2018, url: <https://www.vdi.de/ueberuns/presse/publikationen/details/testenvernetzter-systeme-fuer-industrie-40>.
- [Ie21] "Iso/iec/ieee international standard - software and systems engineering - software testing -- part 2: test processes - redline," in ISO/IEC/IEEE 29119-2:2021(E) - Redline , vol., no., pp.1-64, 28 Oct. 2021.

- [Ma22] Marco Sprenger; Prof. Dr. Martin Ruskowski; Magnus Volkmann; Jörn Peschke: Skill based Production – Wie sehen zukünftig Produktionsstrukturen aus: Podcast der Technologie-Initiative SmartFac-tory KL e.V. 2022, url: <https://smartfactory.de/podcast/skill-based-production-die-produktionsarchitektur-der-zukunft/>, Stand: 14. 04. 2022.
- [MR16] Magnus, S.; Russ, T.; Krause, J.: Anforderungs- und modellbasierte Testfallgenerierung und Testdurchführung unter Nutzung von Methoden zur Netzwerkanalyse. In: Mai 2016.
- [MI21] Miny, T. Functional view of the asset administration shell in an industrie 4.0 system environment, [Online].
Available: <https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/Functional-View.html>, 2021
- [PSB13] Pfrommer, J.; Schleipen, M.; Beyerer, J.: Fähigkeiten adaptiver Produktionsanlagen. atp Edition 55/, S. 42–49, Nov. 2013.
- [Sc21a] Schäfer, S.; Schöttke, D.; Kämpfe, T.; Ralinovski, K.; Klempnow, H.: Nutzung der BaSys 4.2 Middleware-Lösung in der Umgebung von Industriesteuerungen./, 2021, issn: 0083-5560.
- [Sc21b] Schäfer, S.; Schöttke, D.; Kämpfe, T.; Ralinovski, K.; Tauber, B.; Lehmann, R.: Design and Deployment of Digital Twins for Programmable Logic Controllers in Existing Plants. In (Panetto, H.; Macchi, M.; Madani, K., Hrsg.): Proceedings of the 2nd International Conference on Innovative Intelligent Industrial Production and Logistics, IN4PL 2021, October 25-27, 2021. SCITEPRESS, S. 145–150, 2021, url: <https://doi.org/10.5220/0010711000003062>.
- [VP21] Diskussionspapier, "Verwaltungsschale in der Praxis", online unter: <https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/2020-verwaltungsschale-in-der-praxis.html>, Plattform Industrie 4.0, 2021, aufgerufen am 20.05.2022
- [Zi22] Ziesche, C.: Industrie neu gedacht - ein Tech-Podcast von Bosch Rexroth: Was leistet die Verwaltungsschale?, 2022, url: <https://industrie.podigee.io/10-verwaltungsschale>, Stand: 14. 04. 2022.

²<https://github.com/admin-shell-io/submodel-templates/tree/main/development/Simulation/1/0>

Dienste-Design für modulare Fertigungsanlagen mit individuellen Produkten – Variantenvielfalt der Dienste

P. Habiger, G. Hildebrandt, Prof. Dr.-Ing. R. Drath,
Hochschule Pforzheim;
Prof. Dr.-Ing. A. Fay, Helmut-Schmidt-Universität, Hamburg

Kurzfassung

Modulare Produktionsanlagen werden nicht mehr über Signale, sondern über Dienste mit ihren einheitlichen Schnittstellen und hinterlegten Zustandsautomaten angesteuert. Orchestrierungssysteme stehen trotz einheitlicher Schnittstellen vor der Herausforderung, die verfügbaren Dienste richtig zu "verstehen" und zu erkennen, wie und wann diese Dienste angesteuert werden sollen. Daher werden in diesem Beitrag Diensteseigenschaften untersucht, anhand derer die Dienste in Varianten kategorisiert werden. Die Kombination der Varianten ermöglicht eine bedarfsorientierte Produktion, bei der viele Produktionskonfigurationen auch gleichzeitig die Module und damit die Dienstkombinationen passieren können. Unter Zuhilfenahme eines Informationsmodells kann ein Orchestrierungssystem die Informationen über Diensteseigenschaften und -varianten autonom erkennen, interpretieren und dementsprechend koordinieren, sie werden erkundbar. Die Orchestrierungssysteme werden damit in die Lage versetzt, selbständig explizit mit unterschiedlichen Diensten umgehen zu können.

Abstract

Modular production plants are no longer controlled via signals, but via services with their uniform interfaces and stored state machines. Despite uniform interfaces, orchestration systems are faced with the challenge of correctly "understanding" the available services and recognizing how and when these services should be controlled. Therefore, in this paper, service properties are examined, based on which the services are categorized into variants. The combination of the variants enables a demand-oriented production in which many production configurations can also pass through the modules and thus the service combinations at the same time. With the help of an information model, an orchestration system can independently recognize, interpret, and accordingly coordinate the information about service properties and variants; they become explorable. The orchestration systems are thus put in a position to be able to deal explicitly with different services on their own.

1 Warum ein Dienste-Design notwendig ist

Die zunehmende Individualisierung von Produkten erhöht die Produktvarianz und damit die erforderliche Komplexität und Flexibilität, mit der Produktionsanlagen umgehen müssen [1]. Gefordert sind künftig wandelbare Anlagen, die zügig und spezifisch auf sich ändernde Produktionsanforderungen angepasst werden können. Eine Lösung hierfür bilden modulare Anlagen, wie sie bspw. in der Prozesstechnik eingesetzt werden [2]: Module (*Stationen* bzw. *Fertigungseinheiten* [3]) werden im Optimalfall nach dem Baukastenprinzip flexibel und herstellerübergreifend zusammengesetzt. Dabei werden die Funktionalitäten der Module als zustandsbasierte Dienste (*Services*) abstrahiert. Für die Herstellung von Produkten werden die Dienste gemäß einer Prozessabfolge ausgeführt. Die Koordination der Produktion übernimmt in der Regel ein zentrales und auftragsorientiertes Orchestrierungssystem (OS).

Auch wenn die Steuerungsschnittstelle eines Dienstes wohldefiniert ist, sind Dienste nicht gleich Dienste – das beweisen die vielfältigen Einsatzvarianten [4-8]. Das ist insbesondere von Bedeutung, wenn Module verschiedener Hersteller zum Einsatz kommen, die ihre Dienste anderen Systemen nach unterschiedlichen Prinzipien bereitstellen. Beispielsweise können die Ausführungsparameter bei manchen Diensten nur zum Dienststart, bei anderen Diensten in einem pausierenden Zustand oder aber jederzeit verändert werden. Selbst solche scheinbar kleinen Unterschiede haben jedoch großen Einfluss auf ihre Einsetz- und Kombinierbarkeit. Die Vielfalt der Dienste ermöglicht es Modulherstellern, individuelle Funktionalitäten zu implementieren. Um aber diese Vielfalt in Orchestrierungssystemen verarbeiten zu können, ist es notwendig, die zur Ansteuerung relevanten Eigenschaften von Diensten explizit erkundbar zu machen. Darum werden diese Eigenschaften in diesem Beitrag hergeleitet, kategorisiert, als Informationsmodell konzeptioniert und in einem Anwendungsfall näher beleuchtet.

Neben dem einheitlichen Verständnis eines Dienstes existiert in der bedarfsorientierten diskreten Fertigung die Herausforderung, dass unterschiedliche Produkte in diversen Produktzuständen und veränderlicher Reihenfolge zur gleichen Zeit auf demselben Modul gefertigt werden sollen. Sofern ein Dienst (bspw. *Transportieren*) zu jedem Zeitpunkt lediglich ein einzelnes Produkt beeinflusst, ist die Dienstauführung unkritisch. Bei langen Bearbeitungsprozessen oder Transportwegen können Dienste für längere Zeit blockiert sein, sodass die nachlaufenden Produkte nicht weiter gefertigt werden können - die Durchlaufzeiten und die Produktionseffizienz sinken. Das Ziel soll es sein, dass sich mehrere Produkte gleichzeitig im Einflussbereich eines Dienstes oder mehrerer ineinandergreifender Dienste befinden können und dabei gemäß ihren Produktionsanforderungen gefertigt werden, ohne sich gegenseitig zu beeinflussen. Daher widmet sich dieser Beitrag der Fragestellung: Wie können Dienste in modularen fertigungstechnischen Anlagen gewährleisten, dass die Bearbeitung und der notwendige

Materialfluss von mehreren unterschiedlichen diskreten Produkten, mit variablen Produktkonfigurationen oder Produkten in diversen Produktzuständen effizient und einheitlich umgesetzt werden kann?

Im kommenden Abschnitt werden der derzeitige Einsatz und die Ausprägungen von Diensten in der Industrie und Forschung vorgestellt. Darauf basierend werden in Abschnitt 3 Dienstleistungen und deren Ausprägungen aufgezeigt, die speziell für die Ansteuerung von Diensten notwendig sind. Anschließend werden Kombinationen der Dienstleistungen für eine bedarfsorientierte modulare Anlage ausgearbeitet, die u.a. einen parallelen Transport und Bearbeitung von Produkten ermöglichen können. Für eine der identifizierten Kombinationen wird in Abschnitt 5 eine prototypische Implementierung mitsamt einem beschreibenden Informationsmodell aufgezeigt. Das Modell dient für die Orchestrierungssysteme zur automatisierten Auswertung, wie es den Dienst anzusteuern hat. Der Beitrag endet mit einer Zusammenfassung und einem Ausblick.

2 Einsatz von Diensten

Auch wenn die grundlegenden Eigenschaften von Diensten festgelegt sind, ist deren Anwendung für diverse Domänen und Anwendungsfälle unterschiedlich. Zum Beispiel kommt in der Prozesstechnik das *Module Type Package* mit den darin definierten *Services* zum Einsatz. Das MTP definiert in Blatt 4 der VDI 2658-4 [4] die Schnittstelle eines Dienstes und damit dessen Einstellmöglichkeiten (*parameters*) wie bspw. die Ausführungsstrategien (*procedures*) und das Ausführungsende (*self-completing*). Ein Dienst ist so generisch formuliert, dass dieser bspw. für einfache Regelungen bis hin zu komplexen Mischprozessen eingesetzt werden kann. Das MTP ist für die Handhabung von formlosen und homogenen Gütern ausgelegt und damit vorrangig für die Prozesstechnik gedacht. Eine Anwendung in der Fertigungstechnik mit ihrer produktspezifischen sowie bedarfsorientierten Ausführung der Produktionsschritte seitens der Module ist derzeit nicht berücksichtigt.

Bei der Übertragung des MTPs auf die Intralogistik mit geometrisch bestimmten Festkörpern wurden bereits einige Ausführungsvarianten von Diensten vorgestellt und damit die Anwendbarkeit des MTPs für die Intralogistik untersucht [5]. Blumenstein et. al zeigen in [5], dass Dienste für jedes eintreffende Produkt einmalig mit individueller Parametrierung, zyklisch mit gleichbleibender Parametrierung (*cyclic execution service*) oder mehrmalig im pausierenden Zustand mit einer individuellen Parametrierung (*single execution service*) durchlaufen werden können. Damit werden Produkte konkret in die Ausführungsweise der Dienste integriert sowie Parametrierungszeitpunkte berücksichtigt. Wenn sich jedoch mehrere Produkte gleichzeitig auf einem Modul befinden, stoßen solche Dienste an ihre Grenzen.

In [6] werden Dienste in Form von *Skills* mit einem OPC-UA-Zustandsautomaten, einer Informationsschnittstelle und einer Parameterschnittstelle abgebildet. Für jedes Produkt wird in einem agentenbasierten System eine Prozessabfolge als *Rezept* zur Laufzeit auf einen *Device Adapter* übertragen. Das Rezept zeigt für jede Produkt-ID auf, welcher *Skill* mit welchem Parametersatz auszuführen ist. Damit werden die ausführungsspezifischen Daten asynchron auf die Module übertragen, sodass bei einem Ausfall des orchestrierenden Systems die Module weiterhin produktiv bleiben können. Die Umsetzung auf klassischen IEC 61131-3 Steuerungen stellt zwar aufgrund der hohen Anforderungen an die Konnektivität, der dynamischen Datenübertragung und der Schnittstelle zu einem überlagerten *Manufacturing Service Bus* an den *Device Adapter* eine große Herausforderung dar, bietet aber erhebliches Potenzial für die Handhabung mehrerer Produkte.

Im Projekt BaSys 4.0 / 4.2 [7] werden Konzepte für modulare und dienstbasierte Systeme entwickelt. Ein Modul kann dabei als *Control* oder *Group Component* abstrahiert werden, dessen Schnittstelle eindeutig definiert ist. Jede Komponente realisiert nur einen Dienst in Form einer *Operation* mit Betriebsarten und dazugehörigen Zustandsautomaten. Mehrere Dienste, die potenziell von einem Modul ausgeführt werden können, müssen auf mehrere *Control Components* verteilt werden. Die Ausführung der Dienste ist derzeit auftragsorientiert und wird seitens eines OS zentral gesteuert.

3 Anwendungsfall

In diesem Beitrag werden die erarbeiteten Konzepte an einem Anwendungsfall, der das CP Factory System [8] umfasst, beschrieben. Es ist ein modulares und flexibles Ausbildungssystem, das von einem zentralen OS angesteuert, orchestriert und beobachtet wird. Mit der Identifizierung eines Produktes erfragen die Module bei dem OS speziell für dieses Produkt, ob eine bearbeitende oder materialflussändernde Tätigkeit ausgeführt werden soll. Die dafür passenden Parameter werden an die Module übermittelt. Der Materialfluss ist dauerhaft gewährleistet, sodass sich mehrere Produkte gleichzeitig im Arbeitsbereich des Moduls befinden. Ein Beispielszenario für den Prozessablauf auf der CP Factory ist in Bild 1 skizziert. Dieses umfasst nur die Transport- und Bearbeitungsschritte, aber keine Kommunikation zum OS. Ein Produkt A wird vom Eingang des Moduls II zur Bearbeitungsstelle transportiert (1). Dort wird es gemäß den hinterlegten Informationen bearbeitet (2). Während des Bearbeitungsprozesses trifft ein weiteres Produkt B an Modul II ein und wird zur Bearbeitungsstelle transportiert (3). Zeitgleich wird das bearbeitete Produkt A auf das Modul I übergeben (4) und dort gemäß der produktspezifischen Transportroute (5) ausgeschleust.

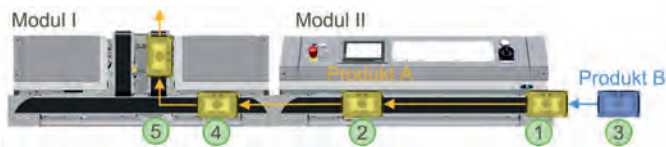


Bild 1: Beispielszenario für den Transport und die Bearbeitung zweier Produkte

Die Kommunikation der Module mit dem OS findet über ein proprietäres Protokoll statt. Der Aufbau der Datenpakete ist sowohl in dem OS als auch im Modul bekannt, sodass über einzelne Signale die Module angesteuert werden. Eine dienstorientierte Ansteuerung der Module existiert bislang nicht. Im Rahmen des Forschungsprojektes METHODS (13FH511KX9) wird die Anlage für ein Plug & Produce im heterogenen modularen Umfeld befähigt. Die Module sollen über Dienste angesteuert werden, was zu den Fragen führt: Welche Dienste gibt es und welche Diensteeigenschaften sollen diese besitzen?

4 Eigenschaften, Varianten und Varianten-Kombinationen von Diensten

Im Folgenden werden Diensteeigenschaften erläutert und für den Anwendungsfall so kombiniert, dass eine bedarfsorientierte parallele Produktion ermöglicht wird.

4.1 Diensteeigenschaften

Die ansteuerungsrelevanten Eigenschaften von Diensten werden von den Autoren als **Diensteeigenschaften** (E) bezeichnet. Eine Auswahl davon ist folgend aufgelistet.

Produktabhängigkeit (E1): Dienste sind abhängig (E1.1) oder unabhängig (E1.2) von Produkten. Produktabhängige Dienste werden individuell für eine Produktkonfiguration ausgeführt, sofern die Produkte identifizierbar sind. Produktunabhängige Dienste agieren mit einer Parameterkonfiguration immer gleich, egal welches Produkt manipuliert wird.

Dienstende (E2): Ein Dienst ist selbstbeendend (E2.1) oder nicht-selbstbeendend (E2.2) (siehe [4]). Nicht-Selbstbeendende Dienste verharren im Gegensatz zu selbstbeendenden Diensten nach einer erfolgreichen Bearbeitung im ausführenden Zustand.

Abarbeitungshäufigkeit (E3): Ein Dienst bearbeitet im ausführenden Zustand ein Produkt und wird anschließend beendet (E3.1). Ein Dienst kann auch nach dem erfolgreichen Bearbeiten eines Produktes in einen wartenden Zustand versetzt (E3.2) werden, der bei erneuter Aufforderung in den ausführenden Zustand verlassen werden kann (siehe [5]). Damit kann ein Dienst mehrere Produkte produzieren, ohne in den Ausgangszustand versetzt zu werden. Außerdem kann ein Dienst so ausgelegt sein, dass der ausführende Zustand für jegliche Produkte nicht verlassen wird (E3.3).

Produkthandhabung pro Dienst (E4): Abhängig von der Abarbeitungshäufigkeit können mit einem Dienst nur ein Produkt (E4.1), mehrere Produkte hintereinander (E4.2) oder mehrere Produkte parallel (E4.3) bearbeitet werden. Letzteres ist ein Sonderfall, der dann eintritt, wenn der Dienst im ausführenden Zustand die Produkte individuell differenzieren und bearbeiten kann – bspw. zwei Produkte mit zwei Transportrouten innerhalb eines Moduls.

Granularität gegenüber der Orchestrierung (E5): Für die Granularität von Services gibt es bisher keine einheitliche Metrik. Daher unterscheiden die Autoren vereinfachend zwischen feingranularen (E5.1) und grobgranularen Diensten (E5.2). Feingranulare Dienste repräsentieren Funktionalitäten von Aktoren, kleinen Gruppen von Aktoren oder Bereichen auf Modulen, die z.B. für Materialflussänderungen relevant sind. Grobgranulare Dienste führen höherwertige Funktionalitäten aus, welche die innere Struktur des Moduls stärker abstrahieren. Ein grobgranularer Dienst kann starten, wenn ein Produkt dem Modul beitrifft und wird beendet, wenn es das Modul verlässt.

Zeitpunkt der Parameter-Konfiguration (E6): Parameter können entweder vor dem Start des Durchlaufs des Dienstes (E6.1), in einem pausierenden Zustand des Dienstes (E6.2) oder jederzeit (E6.3) manipuliert werden.

Art der Parameter-Konfigurationsübertragung (E7): Produktionsparameter können initial zum Ausführungsstart eines Dienstes (E7.1), auf Anfrage des Moduls an ein OS (E7.2) oder durch eine asynchrone Übertragung seitens eines OS (E7.3) konfiguriert werden. E7.1 ist mit einem *auftragsorientierten* Methodenaufruf mit zugehörigen Parametern vergleichbar. Die Anfrage eines Moduls bei E7.2 repräsentiert die *bedarfsorientierte* Implementierung. Ein Modul fragt bei einem OS für einen Parametersatz des identifizierten Produkts an und dieses antwortet passend. Die asynchrone Übertragung umfasst die *gemischte auftrags- und bedarfsorientierte* Übertragung der Parametersätze. Zu bestimmten Zeitpunkten wie z.B. bei einer neuen Produktkonfiguration werden alle Parametersätze zugehörig zu den Produkten an die Module übertragen und dort für den jeweiligen Dienst abgespeichert (siehe [6]).

Ausführungsauslöser (E8): Dienste können entweder von externen Systemen (E8.1) oder intern (E8.2) z.B. von anderen Diensten angesteuert werden. Dazu zählt u.a. der Start eines Dienstes, aber auch das Umschalten von Dienstzuständen.

Die Summe aller Dienstigenschaften bildet die Art und Weise der Ansteuerung des Dienstes ab. Dabei korrelieren die Dienstigenschaften, sodass sich diese ergänzen, verstärken oder gegenseitig ausschließen. Eine generische Korrelationsmatrix ist zu komplex und in vielen Fällen sind die Kombinationsmöglichkeiten spezifisch für den Anwendungsfall. Daher stützen sich die folgenden Betrachtungen auf den Anwendungsfall aus Abschnitt 3.

4.2 Dienstvarianten und -rollen

Die Kombination von Ausprägungen der Diensteseigenschaften ergibt eine **Dienstvariante** (V). Eine Dienstvariante kann eine bestimmte **Rolle** (R) einnehmen, eine Rolle kann über mehrere Dienstvarianten verfügen. Die Rolle beschreibt eine abstrakte Funktionalität eines Dienstes. Im Falle des Beispielszenarios aus Abschnitt 3 können folgende Rollen auftreten: materialfluss-sändernd (R1), materialflussausführend (R2), bearbeitend (R3) und identifizierend (R4). Ein Dienst für die Änderung des Materialflusses (bspw. Schritt 5 bei Bild 1) kann bspw. gemäß den in Bild 3 markierten Diensteseigenschaften in die Dienstvarianten R1.V1 und R1.V2 eingeteilt werden. Diese unterscheiden sich in der Art und dem Zeitpunkt der Parameterkonfiguration sowie der internen bzw. externen Auslösung des Dienstes. Die Dienstvarianten bilden für die verschiedenen Rollen im Zusammenspiel die Möglichkeit, wie die Produkte sowie der Prozessablauf auf den Modulen gehandhabt werden.

4.3 Dienstkombinationen

Dieses im vorherigen Abschnitt benannte Zusammenspiel von Dienstvarianten für die Rollen wird hier als **Dienstkombination** (K) betitelt. Es ist damit die Kombination der Dienstrollen mit genau einer konkreten ausgewählten Dienstvariante. Aus dem bedarfsorientierten Anwendungsfall heraus, der vorschreibt, dass sich mehrere Produkte gleichzeitig im Einflussbereich des Moduls befinden, ergeben sich im Wesentlichen drei Dienstkombinationen mit je vier ausgeprägten Dienstrollen (siehe Bild 2): Anfrage-Koordinierte feingranulare Dienste (K1), Asynchron-Koordinierte feingranulare Dienste (K2) und Asynchron-Koordinierte grobgranulare Dienste (K3).

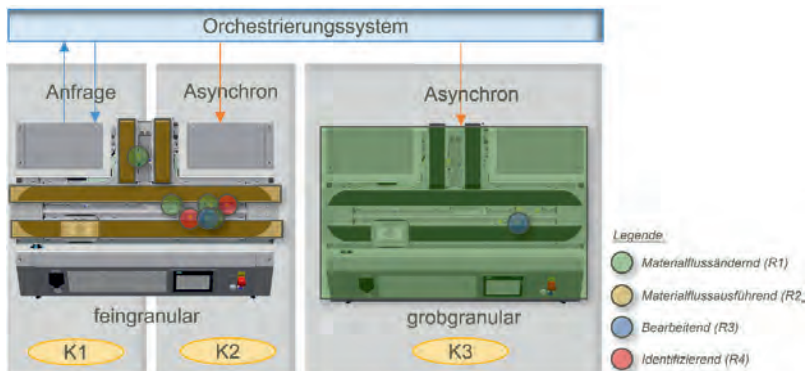


Bild 2: Dienstkombinationen (K) mit deren Rollen (R) für den Anwendungsfall im Vergleich

Die Dienstvarianten unterscheiden sich grundlegend in der Granularität (E5) und der Art der Parameter-Konfiguration (E7). Die Dienstkombinationen mit deren zugewiesenen Dienstvarianten und -eigenschaften sind in Bild 3 abgebildet. Natürlicherweise existieren weitere Dienstkombinationen, die aber aufgrund deren Komplexität nicht weiter betrachtet werden.

			Varianten, Rollen & Kombinationen							
			K1			K2			K3	
			R1/ R3	R2	R4	R1/ R3	R2	R4	R1/ R3	
Eigenschaft	Variation	#	V1	V1	V1	V2	V1	V1	V3	
Produktabhängigkeit	abhängig	E1.1								
	unabhängig	E1.2								
Dienstende	selbstbeendend	E2.1								
	nicht-selbstbeendend	E2.2								
Abarbeitungshäufigkeit	einmalig	E3.1								
	wartender Zustand	E3.2								
	ausführender Zustand	E3.3								
Produkthandhabung pro Dienst	ein Produkt	E4.1								
	hintereinander	E4.2								
	parallel	E4.3								
Granularität	feingranular	E5.1								
	grobgranular	E5.2								
Zeitpunkt Parameter-Konfig.	Start des Durchlaufs	E6.1								
	Pausierender Zustand	E6.2								
	Jederzeit	E6.3								
Art Parameter-Konfig.	Initial	E7.1								
	Anfrage	E7.2								
	Asynchrone Übertragung	E7.3								
Ausführungsauslöser	Extern	E8.1								
	Intern	E8.2								

Bild 3: Dienstkombinationen (K) mit deren Rollen (R) und Varianten (V) sowie gesetzten (blau) und ausgewählten (grün) Dienststeigenschaften (E) für den Anwendungsfall

4.3.1 Dienstkombination 1 (K1): Anfragekoordinierte feingranulare Dienste

Feingranulare Dienste (E5.1) zeichnen sich durch die Aufteilung der Funktionalitäten auf kleine Bereiche innerhalb eines Moduls aus. Die Bereiche müssen so gewählt sein, dass ein Dienst von nur einem Produkt belegt wird. Damit wird der Materialfluss und die Bearbeitung anderer Produkte weiterhin gewährleistet. In diesem Anwendungsfall betrifft dies genau die definierten Dienstvarianten R1.V1 bis R4.V1.

Materialflussausführende Dienste (R2.V1) werden unabhängig von konkreten Produkten dauerhaft intern ausgeführt, sodass der Materialfluss stetig gewährleistet ist und mehrere Produkte parallel transportiert werden können. Die Produktidentifizierungsdienste (R4.V1) finden, wie bei R2.V1, intern und produktunabhängig statt, aber der Dienst verharrt in einem wartenden Zustand, aktiviert sich selbst bei einer Objekterkennung und teilt der Orchestrierung das identifizierte Produkt wie z.B. die Produkt-ID mit. Das Modul stellt somit eine Anfrage an die Orchestrierung, was mit dem Produkt gemacht werden soll. Bei einer geplanten Aktivität

überträgt die Orchestrierung für die materialflussändernden (R1.V1) und bearbeitenden Diensten (R3.V1) den Parametersatz und startet den Dienst von extern.

Bei dem Beispielszenario aus Abschnitt 3 werden die materialflussausführenden und identifizierenden Dienste auf allen Modulen initial gestartet. Das Produkt A wird damit bei Modul II auf einem stetig aktiven Förderband (1) zu der Bearbeitungsstelle (2) transportiert, wo das Produkt zum Stillstand kommt. Der identifizierende Dienst erkennt in seinem pausierenden Zustand das Produkt und teilt die Produktinformationen der Orchestrierung mit. Die Orchestrierung prüft, ob und wie das Produkt bearbeitet werden soll und übermittelt den dazugehörigen Parametersatz an den bearbeitenden Dienst. Parallel läuft das Produkt B auf dem Förderband 1 nach. Wenn der bearbeitende Dienst beendet ist, wird dies der Orchestrierung rückgemeldet und der materialflussändernde Dienst wird für das Öffnen des Stoppers aktiviert. Produkt A geht auf das Modul I über, während Produkt B erneut am Stopper stehen bleibt. An der Abzweigung auf Modul I findet dieselbe Abfolge der Prozesse statt, nur dass hier keine Bearbeitung gestartet wird, sondern nur das Stellen der Weiche als materialflussändernder Dienst.

4.3.2 Dienstkombination 2 (K2): Asynchronkoordinierte feingranulare Dienste

Die Aufteilung der feingranularen Dienstvarianten gilt analog zu Abschnitt 4.3.1. Der Hauptunterschied liegt darin, dass die Parameterkonfigurationen für Produkte asynchron auf die Module übertragen werden, sodass bei Änderungen der Produktkonfigurationen die in den Modulen hinterlegten Parameter jederzeit aktualisiert werden. Bei der Erkennung von Produkten wird der dazu hinterlegte Parametersatz intern verwendet, um damit den materialflussändernden (R1.V2) oder bearbeitenden (R3.V2) Dienst auszuführen. Für die Übertragung der Parameter muss gewährleistet sein, dass sich der Dienst bestenfalls im pausierenden Zustand bzw. auch im ausführenden, aber darin sicheren Zustand befindet. Empfohlen wird, dass der wartende Zustand eingenommen wird, da so dem OS implizit mitgeteilt werden kann, wann Parameter übermittelt werden können. Die Orchestrierung ist dabei nur für die Konfiguration der Dienste zuständig, während das Modul selbstständig die Dienste ausführt. Bei R2 und R4 ändert sich gegenüber K1 nichts.

Der Ablauf des Beispielszenarios verläuft, wie bei K1 mit dem Unterschied, dass die Module vor der Ausführung von R1.V2 und R3.V2 nicht mehr bei dem OS nach dem passenden Parametersatz nachfragen, sondern dies intern bereits vorhanden ist. Das OS muss jedoch im Vorhinein die Konfigurationen mit den Modulen synchronisieren. Das OS ist im Bilde aller feingranularen Dienste, übernimmt jedoch lediglich eine konfigurierende und keine steuernde Funktion.

4.3.3 Dienstkombination 3 (K3): Asynchronkoordinierte grobgranulare Dienste

Bei grobgranularen Diensten werden die internen Materialflussrouten sowie die Identifikationsmechanismen durch die Transport- und Bearbeitungsdienste abstrahiert. R2 und R4 sind der Orchestrierung gänzlich unbekannt. Damit wird ein Großteil der Transport- und Bearbeitungslogik von den Modulen selbst übernommen. Die Produkte werden im ausführenden Zustand mit unterschiedlichen Produktionszielen von R1.V3 und R3.V3 gleichzeitig bearbeitet bzw. transportiert. Es existiert dabei im Steuerungsprogramm des Moduls nur eine Instanz eines Dienstes. Bei einer Mehrfachinstanziierung eines grobgranularen Dienstes für jedes Produkt gäbe es Zugriffsprobleme auf die Aktoren, Sensoren und Sub-Dienste. Mit Hilfe einer geschickten Parametrierung und ggfls. mehreren parallel geltenden Parametersätzen ist die gleichzeitige Handhabung mehrerer Produkte möglich. Das setzt auch voraus, dass der Dienst im ausführenden Zustand verharrt, die Parameter jederzeit unter der Voraussetzung, dass die bestehende Bearbeitung oder Transport nicht beeinflusst werden, geändert und wie bei K2 intern ausgeführt werden können.

Im Anwendungsbeispiel existieren nur Dienste für das Bearbeiten der Produkte sowie das Ändern des Materialflusses wie bspw. das Stellen der Weiche, die einmalig gestartet werden und dann im Pause-Zustand auf Produkte und auszuführende Prozesse warten. Für jedes Produkt werden die Parametersätze jedes Dienstes an die Module I und II übermittelt und dort intern wie bei K2 zu gegebener Zeit ausgeführt.

4.3.4 Vergleich der Dienstkombinationen

Alle Dienstkombinationen sind für den gegebenen Anwendungsfall geeignet, unterscheiden sich aber gemäß der Merkmale Kommunikationsaufwand, Speicherbedarf, Autarkiegrad und Logikverteilung (siehe Tabelle 1).

Tabelle 1: Differenzierungsmerkmale für Dienstkombinationen

Differenzierungsmerkmal	K1	K2	K3
Kommunikationsaufwand	Hoch	Gering	Gering
Speicherbedarf	Gering	Hoch	Hoch
Autarkiegrad	Gering	Mittel	Hoch
Logikverteilung	Ausgelagert	Integriert	Integriert

Bei der asynchronen Parameter-Konfigurationsübertragung (K2 und K3) werden nur zu bestimmten Zeitpunkten Parametersätze zugehörig zu den Produkten an die Module übertragen. Damit agieren die Module bei einem Ausfall der Kommunikation zum OS mit der dann

geltenden Parameterkonfiguration autark. Zusätzlich ist der Kommunikationsaufwand gering. Jedoch erfordert diese Variante einen hohen Speicherbedarf und ein aufwendiges Datenmanagement in den Modulen in Abstimmung mit dem OS. Datensätze müssen synchronisiert, übertragen, gespeichert, gelöscht und manipuliert werden, was für klassische SPS einen großen Implementierungs- und Betriebsaufwand birgt. In einem anfrageorientierten Ansatz (K1) ist der Kommunikationsaufwand größer, die Koordination der Nachrichten aufwendiger und die Module können nur bedingt autark agieren, denn sie sind immer auf die Antwort des OS angewiesen. Die Module benötigen jedoch kein Speichermanagement, da die Parameter immer passend zum Produktions- oder Transportstart übermittelt werden. Bei der Differenzierung von feingranular (K1 und K2) und grobgranular (K3) ist die Logikverteilung hervorzuheben. Bei feingranularen Strukturen wird sehr viel Know-How, was in den Modulen selbst geregelt werden könnte, für das OS offengelegt und vorausgesetzt, dass dieses damit umgehen kann. Grobgranulare Strukturen abstrahieren die Logik sehr gut, müssen aber in den Modulen so vorgedacht und implementiert sein, dass mehrere Produkte gleichzeitig gehandhabt werden können. Zusätzlich ist an dieser Stelle zu erwähnen, dass sich zukünftig Orchestrierungssysteme nicht nur auf diese drei Kombinationen fokussieren müssen. Bei den präsentierten Beispielen handelt es sich um Vorschläge basierend auf dem Anwendungsfall.

5 Informationsmodellierung

Damit die Dienstigenschaften und -rollen den Orchestrierungssystemen mitgeteilt werden können, sodass diese mit den diversen Dienstarten umgehen können, ist ein Informationsmodell bzw. Zuweisungen für bestehende Informationsmodelle notwendig.

Innerhalb einer erweiterbaren *RoleClass*-Bibliothek werden alle Dienstrollen mit den hinterlegten Dienstvarianten als *Attribute* modelliert. Jede Dienstvariante besteht aus einer Liste an Attributen, welche die Dienstigenschaften repräsentieren. Die Ausprägungen der Dienstigenschaften sind mittels *Constraints* vorgegeben. Bei der Zuweisung einer *RoleClass* zu einer *SystemUnitClass* oder *InternalElement* wird die jeweilige Dienstvariante der Klasse oder des Objekts vererbt. Die Anwendung der Bibliothek folgt im kommenden Abschnitt.

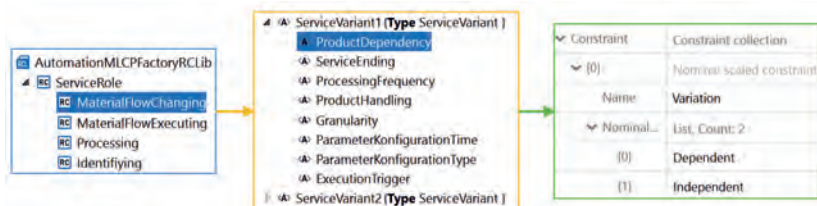


Bild 4 Informationsmodell als Rollenklassenbibliothek der Dienstvarianten und -rollen

6 Prototypische Implementierung und Informationsmodellierung

Da alle Kombinationen aus Abschnitt 4.2 Vor- und Nachteile und somit ihre Daseinsberechtigung haben, wird hier im ersten Schritt eine prototypische Implementierung für das Branch-Modul der CP Factory für die Dienstkombination K1 gemäß Abschnitt 4.3.1 vorgestellt. Für jede Dienstvariante aus K1 sind die entsprechenden Dienste auf dem Modul gemäß dem Module Type Package mit Hilfe des darauf spezialisierten PA-Toolkits [9] auf einer virtuellen CoDeSys-Steuerung implementiert (siehe Bild 5 oben). Der Prozessablauf für den Transport auf der unteren Hälfte des Moduls (grüner Pfeil) ist im unteren Bereich von Bild 5 aufgezeigt. Das Lesen der Produkt-ID findet über einen Rückgabeparameter eines separaten Dienstes für das RFID-Lesegerät statt. Eine weitere Möglichkeit ist das *Service-Interaction-Pattern* des MTPs zu nutzen, in dem ein Informationsaustausch aus Fragen und Antwortmöglichkeiten innerhalb der bearbeitenden und materialflussändernden Dienste ermöglicht wird.

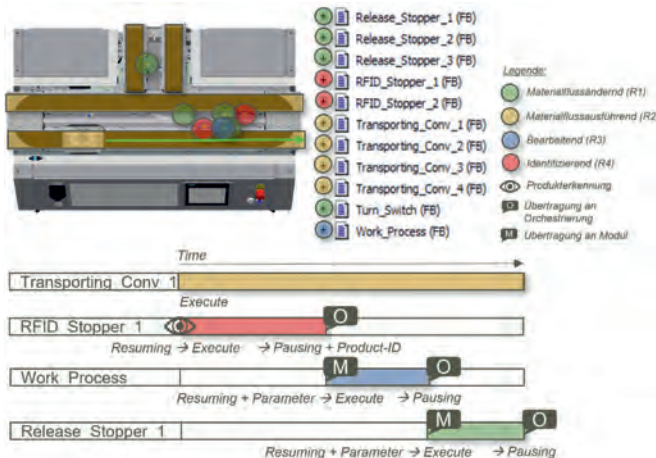


Bild 5: Implementierte Dienste und der Programmablauf für ein simples Szenario

Aus dem PA-Toolkit wird die dazugehörige MTP-Datei generiert. Diese Datei wird von einem erweiterten Informationsmodell angelehnt an [10], das unter anderem die Arbeitsbereiche, Geometriemodelle etc. enthält, referenziert. In diesem Modell ist jeder Dienst als *Skill* erneut aufgeführt, referenziert auf den MTP-Dienst. Dem *Skill* ist eine Dienstrolle mit einer ausgewählten Dienstvariante gemäß Abschnitt 5 zugewiesen. Das Informationsmodell für das Branch-Modul ist in Bild 6 dargestellt – Referenzen sind modelliert, aber nicht sichtbar.

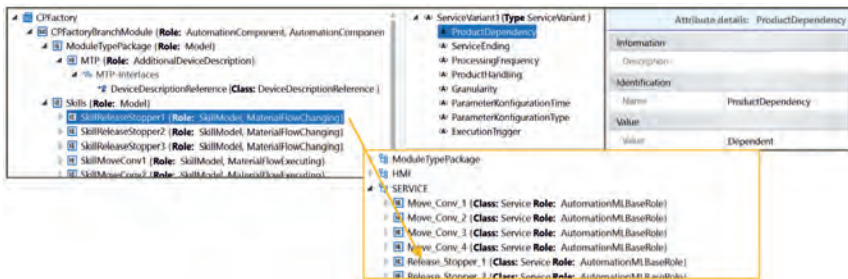


Bild 6: Informationsmodell der prototypischen Implementierung

Derzeit ist kein konkretes OS im Einsatz, das die Informationen automatisiert verarbeitet und gemäß den hier gestellten Anforderungen in Kooperation mit den Modulen ausführt. Mittels manueller Tätigkeiten werden die Dienste nach und nach gemäß dem Ablauf gestartet und pausiert. Die parallele Handhabung von Produkten ist damit ermöglicht.

7 Zusammenfassung und Ausblick

Dieser Beitrag beleuchtet, dass Dienste trotz Standardisierung vielfältig ausdifferenzierte Eigenschaften besitzen, die ihre Einsatzmöglichkeiten maßgeblich beeinflussen. Um ein Orchestrierungssystem zu befähigen, Dienste softwarebasiert erkunden, verstehen und somit zielgerichtet verwenden und für verschiedene Einsatzzwecke kombinieren zu können, werden sie mit einem Informationsmodell ergänzt. Am Beispiel einer parallelen Produktion mehrerer Produkte wurde mit Hilfe der in diesem Beitrag vorgestellten Kategorisierung sowie Implementierung ein Ansatz vorgestellt, um Orchestrierungssysteme automatisiert zu befähigen, ihnen bislang unbekannte Dienste zu integrieren. Dienste können demnach in bestimmte Eigenschaften mit konkreten Ausprägungen untergliedert werden, die sog. Diensteeigenschaften. Diese beschreiben bspw., ob der Dienst vom Modul selbst intern oder von einem anderen System extern angesteuert werden muss. Diese Information ist für das ansteuernde System notwendig, sodass dieses entsprechend den Prozessablauf der Produkte orchestrieren kann. Für verschiedene Prozessrollen wie bspw. bearbeitende Prozesse wurden Ausprägungen der Diensteeigenschaften ausgewählt und zu sog. Dienstrollen mit Dienstvarianten zugeordnet. Durch eine geschickte Kombination dieser Dienstvarianten entstehen sog. Dienstkombinationen. Eine Dienstkombination ist eine aggregierte Musterlösung, wie eine modulare Anlage seitens eines Orchestrierungssystem bedarfsorientiert angesteuert werden kann und dabei gewährleistet, dass ein Modul mit mehreren Produkten gleichzeitig umgehen kann. Aufgrund der expliziten Modellierung der Eigenschaften, Rollen und Varianten der Dienste in einem

Informationsmodell liegt dem Orchestrierungssystem jegliche Information vor, wie die Dienste der Module angesteuert werden können – sofern diese in der Lage sind, das Modell zu interpretieren. An einem Anwendungsbeispiel wurde aufgezeigt, wie mit feingranularen Diensten eine Ansteuerung der Module möglich ist. Um die feine Granularität der Module gröber zu gestalten, wären zukünftig sogenannte abstrakte Dienste möglich, die bereits in dem Informationsmodell aus der Kombination mehrerer feingranularer Dienste gebildet werden. Auf diese Weise sind wiederverwendbare Vorlagen für grobgranulare Dienste modellierbar. Diese Dienste können dann wiederum in der Orchestrierung auch mehrfach für die Produktion der Produkte instanziiert werden. Somit wäre eine abstrakte Modellierung von Dienstabläufen trotz feingranularer Dienste möglich.

Referenzen

- [1] C. Plehn, A method for analyzing the impact of changes and their propagation in manufacturing systems, Dissertation, München, 2017.
- [2] S. Lier, M. Grünewald, *Net Present Value Analysis of Modular Chemical Production Plants*, In: Chemical Engineering & Technology, Bd. 34, Nr. 5, 2011.
- [3] DIN SPEC 91345. Referenzarchitekturmodell Industrie 4.0 (RAMI4.0), 2016.
- [4] VDI 2658-4. Automatisierungstechnisches Engineering modularer Anlagen in der Prozessindustrie, 2022.
- [5] M. Blumenstein, A. Fay, K. Gryczycha, S. Lier, C. Kotsch, F. Borgmann, N. Austermann: *Designprinzipien für den Modul- und Serviceentwurf in modularen Logistikanlagen*, In: Automation 2021, S. 101–115, 2021.
- [6] K. Dorofeev, C.-H. Cheng, M. Guedes, P. Ferreira, S. Profanter, A. Zoitl: *Device adapter concept towards enabling plug&produce production environments*. In: ETFA 2017, S. 1–8, 2017.
- [7] J.A. Grothoff, C.A. Wagner, U. Epple: *BaSys 4.0: Metamodell der Komponenten und Ihres Aufbaus*, DOI: 10.18154/RWTH-2018-225880, 2018.
- [8] Festo Didactic SE: *CP Factory Die Cyber-Physical Factory*, www.festo-didactic.com, [Letzter Zugriff: 14.04.2022]
- [9] Festo: *PA-Toolkit*, www.festo.com/pa-toolkit, [Letzter Zugriff: 14.04.2022]
- [10] P. Habiger, G. Hildebrandt, R. Drath, M. Barth, A. Fay, A. Zor, M. Marseu: *Plug & Produce im real-virtuellen Kontext fertigungstechnischer heterogener Anlagen*. In: AALE 2022, S. 33-44, 09.03.-11.03.2022.

Typenbasiertes Engineering für modulare Anlagen: Fallstudie zur Umsetzung eines modularen Neutralisationsprozesses

Dipl.-Ing. **Julius Lorenz**, Dipl.-Ing. **Anselm Klose**,
Dipl.-Ing. **Isabell Viedt**, Dipl.-Ing. **Jonathan Mädler**,
Dr. rer. nat. **Valentin Khaydarov**, Prof. Dr.-Ing. **Leon Urbas**,
Technische Universität Dresden, Dresden

Kurzfassung

Modulare Anlagen versprechen hohe Flexibilität und Zeitgewinn sowohl im Engineering als auch Betrieb [1]. Die systematische Umsetzung eines typenbasierten Engineerings auf Basis des Module Type Package (MTP), wobei konkrete Instanzen eine vordefinierte, abgestimmte Typ-Spezifikation erfüllen, ist grundlegend für diese Ziele. Die klare Aufgabentrennung und Entkopplung der Entwicklungszyklen der Module durch die *Hersteller* und dem der modularen Anlage durch den *Betreiber* wird muss zudem unterstützt werden [2].

Um die hohen Anforderungen an Innovations-, Produkt-, und Prozessflexibilität in modularen Anlagen zu erfüllen, ist der Übergang zu einem typenbasierten Planungsprozess von Vorteil [3]. Eine Entwicklung im Rahmen konkreter Anwendungsfälle führt zudem tendenziell zur Entwicklung von Unikaten mit spezifischer, aufwendiger Beschreibung. Die Einigung auf Typen erhöht somit auch die Wahrscheinlichkeiten für Auswahl und Wiederverwendung.

Es wurde daher ein Lösungsansatz für eine typenbasierte Orchestrierung modularer Anlagen entwickelt. Das Vorgehen wurde im Process-to-Order Lab [4] am Beispiel eines Neutralisationsprozesses in Form einer Fallstudie sowohl aus *Hersteller-* als auch *Betreiberperspektive* und mit dem Process Orchestration Layer Polaris untersucht.

1. Motivation

In der Prozessindustrie sollen modulare Anlagen die Zeit von einer Produktidee bis zur Anlage in Betrieb stark verkürzen sowie die Auswahl, Rekonfiguration und Wiederverwendung von Anlagenteilen erlauben [5]. Die effiziente Umsetzung dieses Konzepts erfordert grundlegende Änderungen der Planungsprozesse, Anlagenstruktur und Schnittstellen. Damit die Konzepte flächendeckend umgesetzt werden können spielt die herstellerübergreifende Standardisierung eine Schlüsselrolle. VDI 2776 beschreibt Planung und Design modularer Anlagen aus der verfahrenstechnischen Perspektive. Eine modulare Anlage (engl. modular Plant - MP) besteht demnach aus modularen Prozesseinheiten (engl. Process Equipment Assemblies - PEA), der

Infrastruktur sowie einem Prozess Orchestration Layer (POL) [6]. Die Richtlinie bietet die erforderlichen Hilfestellungen für die Realisierung und den Betrieb der PEAs sowie einer MP für einen konkreten Prozess basierend auf den PEAs.

VDI/VE/NAMUR 2658 beschreibt das automatisierungstechnische Engineering modularer Anlagen in der Prozessindustrie mit dem Ziel eine herstellerunabhängige Ansteuerung einer PEA mittels einheitlicher Schnittstelle und Beschreibung, dem Module-Type-Package (MTP), durch die POL zu ermöglichen [7]. Eine VDI-Handlungsempfehlung [8] beschreibt das Zusammenspiel der beiden Disziplinen.

Process Equipment Assembly

Eine PEA realisiert in einer modularen Anlage eine verfahrenstechnische Grundoperation, welche eine oder mehrere prozesstechnische Funktionen erfordert. Eine verfahrenstechnische Grundoperation ist der einfachste Vorgang zur Stoffveränderung oder –Umwandlung bei der Durchführung eines Verfahrens [6]. Aus automatisierungstechnischer Sicht werden die prozesstechnischen Funktionen von einer PEA in Form von Diensten für die POL zur Verfügung gestellt. Eine direkte Interaktion mit der Feldebene und einzelnen Aktoren ist in diesem Fall nicht mehr erforderlich, vielmehr kann direkt ein Dienst zum „Erhitzen“ mit Parametern aufgerufen werden. Die Beschreibung im MTP stellt die einheitliche, herstellerunabhängige Beschreibung und die Integrationsfähigkeit in eine POL sicher. [7]

Process Orchestration Layer

Die POL ist die „PEA-übergreifende, automatisierungs- und informationstechnische Ebene für den Betrieb einer modularen Anlage“ [6]. In der POL erfolgt die Orchestrierung der modularen Anlage, was die Phasen Import, Konfiguration, Betrieb und Export mit den grundlegenden Ergebnissen aus dem Engineering der modularen Anlage und der PEAs verbindet [9,10]. Die konkrete Ausgestaltung der internen Funktionen und Abläufe sowie Architekturen innerhalb einer POL sind derzeit Gegenstand der Forschung [9,11]. Der Import der MTPs bildet den Ausgangspunkt für die Konfiguration und Koordination des gemeinsamen Betriebs der PEAs auf Anlagenebene z.B. Konfiguration von Interlocks zwischen den PEAs, Erstellung und Ausführung dienstebasierter Rezepte [12]. Da das MTP die wesentlichen Informationen für einen Großteil der Orchestrierung beinhaltet, könnte die Anlagenkonfiguration parallel zur Entwicklung der Instanzen begonnen werden, wenn das MTP schon als Teil der funktionalen Anforderungsspezifikation entsteht [3]. Außerdem wird das typenbasierte Konzept benötigt, um Inhalt und Struktur für flexible Einsatzzwecke und unabhängig von der Instanz abzustimmen.

Typen ermöglichen die Herstellung einer Relation zwischen einem konkreten Asset und einem oder mehreren Prozessschritten (Abbildung 1) [3].

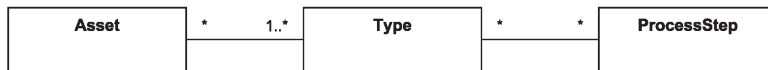


Bild 1: Zusammenhang zwischen Asset, Type und ProcessStep aus [3]

2. Engineering von modularen Anlagen

Der Planungsprozess modularer Anlagen beinhaltet die Entwicklung der modularen Anlage durch den *Betreiber* und die Entwicklung der erforderlichen PEAs durch die jeweiligen *Hersteller*, sofern der Betreiber keine PEAs zur Wiederverwendung besitzt [2]. Der *Betreiber* plant eine modulare Anlage indem die notwendigen Funktionen spezifiziert werden und passende PEAs ausgewählt werden. Diese können entweder vorhanden sein, oder bei einem Hersteller neu beauftragt werden. Im Worst-Case Szenario ist die Entwicklung komplett abhängig von der Spezifikation des Herstellers und eine neue PEA muss entworfen werden, deren MTP erst mit Fertigstellung der PEA ausgeliefert wird (siehe Abbildung 2). Verzögerungen an einer Stelle haben direkte Auswirkungen auf die gesamte Entwicklung und spät identifizierte Schwachstellen können zu einem iterativen Prozess führen.

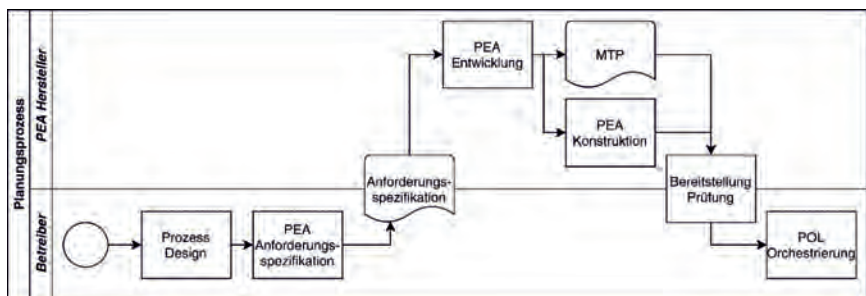


Bild 2: Engineering einer modularen Anlage mit den Rollen Betreiber und PEA Hersteller

Eine Beschleunigung kann erreicht werden, wenn das MTP frühzeitig als Teil der Anforderungsspezifikation erstellt wird. Das MTP kann dann für die POL-Orchestrierung parallel zur Entwicklung der PEAs genutzt werden (Abbildung 3). Notwendige Änderungen die während der Entwicklung einer PEA identifiziert werden, stellen ein Risiko dar, denn diese müssen dann bei der bereits begonnenen POL-Orchestrierung berücksichtigt werden.

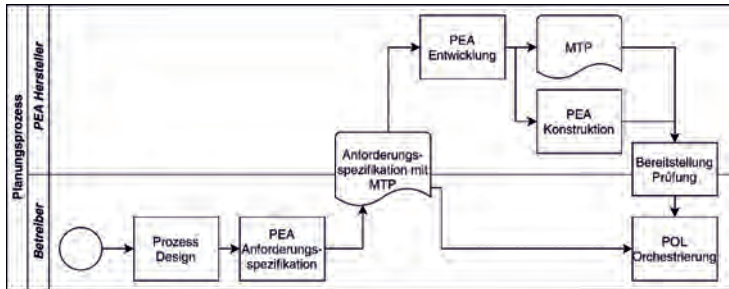


Bild 3: Engineering einer modularen Anlage mit früher MTP Spezifikation

Sind passende PEAs für die im Anlagenengineering identifizierten Funktionen sogar schon von einem Hersteller entwickelt worden, müssen diese nur ausgewählt und bestellt werden [13]. Im Anschluss erfolgt die Konstruktion durch den Hersteller oder die PEA kann sofort bereitgestellt werden, wenn die Konstruktion schon erfolgt ist und die PEA auf Lager ist (siehe Abbildung 4).

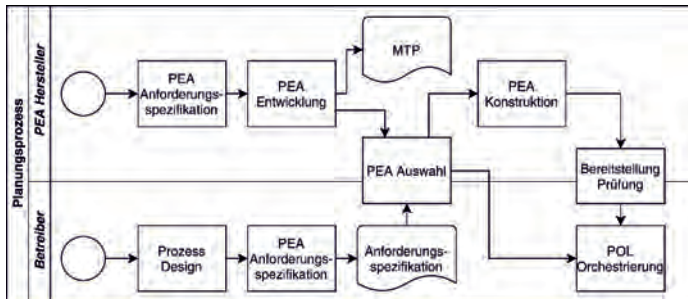


Bild 4: Engineering einer modularen Anlage mit initiativer Entwicklung einer PEA durch den Hersteller

Die daraus resultierende Parallelisierung und Entkopplung der Entwicklungszyklen im Planungsprozess einer modularen Anlage führen zur Verkürzung der Entwicklungszeit, wie bereits in der 50%-Idee beschrieben [14-18]. Es lässt sich aber auch der Bedarf einer einheitlichen Semantik zur Beschreibung des Verhaltens, der Eigenschaften und Attribute einer PEA ableiten [6]. Es müssen außerdem fest definierte Übergabepunkte vorhanden sein, welche beispielsweise durch das MTP als Schnittstelle zwischen Verfahrenstechnik und Automatisierung

nungstechnik realisiert werden können. Das typenbasierte Engineering kann helfen die Kommunikation zu strukturieren und Missverständnisse zu reduzieren z.B. bei Spezifikation und Auswahl einer PEA.

Typen im Engineering modularer Anlagen

Das Konzept eines Typs und abgeleiteten, konkreten Ausprägung (Instanz), welche die Spezifikation des Typs erfüllt, ist aus der Informatik und objektorientierten Programmierung bekannt und bietet auch für modularer Anlagen Vorteile. Ein Typ beschreibt eine durch charakteristische Merkmale gekennzeichnete Kategorie, die Instanz ist eine Ausprägung, die die Spezifikation eines oder mehrerer Typen erfüllt [19]. Dieses Konzept lässt sich auf die Bereiche des modularen Konzeptes übertragen bei denen Kategorien gebildet werden z.B. PEAs die die gleichen verfahrenstechnischen Grundoperationen bzw. prozesstechnischen Funktion(en) erfüllen oder wie bei den verschiedenen Grundfunktionen aus [8].

Ein bestimmter Typ repräsentiert eine wiederverwendbare Einheit ohne konkrete, instanzspezifische Informationen. Kennen im Planungsprozess sowohl *Betreiber* als auch *Hersteller* diesen Typ, kann die Kommunikation optimiert werden z.B.: *Benötigt wird eine PEA vom Typ Stirred Tank Unit gemäß der Spezifikation des BioPhorum* [20]. Vorteile ergeben sich auch bei der Spezifikation und Auswahl neuer PEAs. Realisiert die POL die Orchestrierung auf Basis von Typen, können PEAs die aus POL Sicht den gleichen Typ erfüllen leicht ausgetauscht, die Abhängigkeit der Orchestrierung von konkreten PEAs bzw. MTPs aufgehoben und somit die Abhängigkeit umgekehrt werden [3]. Die POL kennt die zu erwarteten Informationen aus dem MTP und kann diese als Platzhalter für die instanzspezifischen Informationen verwenden. In Zukunft könnten daraus auch Typen auf Anlagenebene als Blueprint für eine Anlagenkonfiguration definiert werden, ohne auf baugleiche PEAs angewiesen zu sein.

Die initiale Definition eines PEA-Typs ist eine Herausforderung, der Hersteller und Betreiber gemeinsam begegnen müssen. Bereits entwickelte MTPs können gesammelt, analysiert und kategorisiert werden, wobei ein gemeinsames Verständnis und eine gemeinsame Beschreibung wichtig ist. Die Definition der benötigten verfahrenstechnischen Grundoperationen respektive PEAs und die der Prozesstechnischen Funktionen muss aus der Perspektive der Betreiber und des Betriebs erfolgen. Der Betreiber definiert pro PEA ein Lastenheft mit MTP als standardisierte Beschreibung, woraus der Hersteller das Pflichtenheft ableitet und Verständlichkeit, Vollständigkeit, Konsistenz sowie Machbarkeit der Anforderungen überprüft. Im letzten Schritt muss der Betreiber prüfen, ob die PEA des Typs erfolgreich in der Orchestrierung genutzt werden kann.

Alle bereits entwickelten Typen können sowohl *Herstellern* als auch *Betreibern* zur Verfügung gestellt werden und jeweils als Grundlage für deren Planungsprozess dienen. Das Engineering einer neuen modularen Anlage kann dann auf Basis der Typen erfolgen (siehe Abbildung 5). Wählt ein *Hersteller* einen Typ aus der Menge verfügbaren Typen, kann der Typ als Kriterium bei der Auswahl dienen. Ein *Hersteller* muss zudem bei der Entwicklung einer neuen PEA (Abbildung 4) ein MTP nicht vollständig neu entwickeln, sondern kann auf ein bestehendes MTP referenzieren und dieses auch erweitern, sofern dies keine Rückwirkung auf den referenzierten Typ hat.

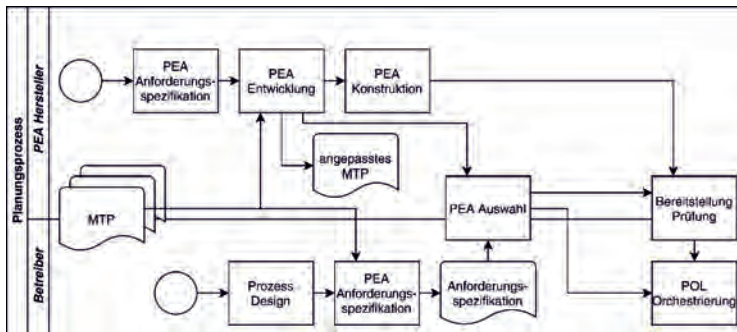


Bild 5: Engineering einer modularen Anlage auf Basis bestehender Typen

Beispiele für erste Typ Spezifikationen ist die *Stirred Tank Unit* des BioPhorum [19] oder die *Standard Dosing Unit* des Arbeitsausschusses Modulare Anlagen der Namur [8], welche von Herstellern und Betreibern entwickelt wurden.

Fallbeispiel Neutralisation

Im Process-to-Order Lab der TU Dresden werden aktuelle Forschungsthemen an industriennahen Demonstratoren untersucht. Ein Schwerpunkt ist die modulare Automation mit der Infrastruktur und POL für den Betrieb der derzeit 15 PEAs. Dies bildet die Ausgangssituation für die Untersuchung zum typenbasierten Engineering am Beispiel einer Neutralisation. Die Neutralisation ist ein bedeutender Prozess z.B. in der Abwasserbehandlung, da in der Reaktion die korrosive Wirkung einer Säure und einer Base aufgehoben werden kann [21]. Als Reaktionsprodukte entstehen Wasser und Salz. Idealerweise beträgt der resultierende pH-Wert 7, dazu müssen bei der Reaktion Konzentration und Menge der Säure und Basis im richtigen Verhältnis zu einander stehen.

Für die Umsetzung im Process-to-Order Lab sollen Lauge und Säure in einer bestimmten Konzentration erzeugt und bereitgestellt werden. Die zwei Stoffströme werden für die Neutralisation zusammengeführt. In der Neutralisation reagieren Essigsäure und Natriumhydrogencarbonat miteinander zu Natriumacetat, Kohlenstoffdioxid und Wasser.

Im Prozessdesign wurden die für eine modulare Neutralisationsanlage erforderlichen verfahrenstechnischen Grundoperationen für die Neutralisation identifiziert und von diesen die prozesstechnischen Funktionen nach VDI 2776 Blatt 2 und deren Anordnung abgeleitet (siehe Abbildung 5). Anschließend können die erforderlichen PEAs auf Basis der prozesstechnischen

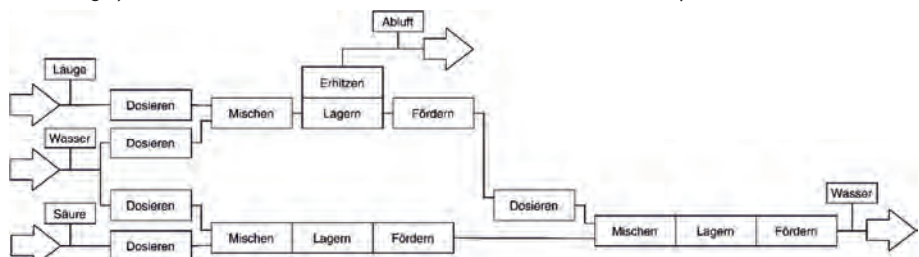


Bild 6: Identifizierte Grundfunktionen zur Umsetzung der modularen Neutralisation

Funktionen identifiziert werden [22], wobei auf bereits vorhandene Typen geprüft werden kann. Die im Process-to-Order Lab vorhandenen PEAs z.B. für Dosieren, Lagern oder Rührkesselreaktoren bilden eine mögliche Grundlage erster Typen. Ein Rührkesselreaktor im Process-to-Order Lab besitzt beispielsweise die Funktionen „Dosieren“ an den Eingängen, „Mischen“ mittels Rührwerk, „Lagern“ aufgrund des Volumens des Tanks sowie „Fördern“ und „Dosieren“ am Ausgang. Entsprechend der Funktionen besitzen diese PEAs Dienste, welche die Funktionen bereitstellen.

Im konkreten Fall wurden drei PEAs für die modulare Anlage identifiziert:

- Zwei Rührkesselreaktor („Dosieren“, „Dosieren“, „Mischen“, „Lagern“ und „Fördern“)
- Ein Rührkesselreaktor mit Heizfunktion („Dosieren“, „Mischen“, „Lagern“, „Erhitzen“ und „Fördern“)
- Eine Vorlage („Dosieren“, „Lagern“)

Der Rührkessel mit Heizfunktion könnte für einen anderen Einsatzzweck auch als Rührkesselreaktor genutzt werden, da die Heizfunktion als Ergänzung zu den grundlegenden Funktionen realisiert und nutzbar ist.

Bei der durchgeführten Untersuchung, ob eine PEA im Labor für den Einsatz geeignet ist, war die bestehende Beschreibung der PEAs und deren prozesstechnischen Funktionen hilfreich. Eine einheitliche Spezifikation hätte die Aufwände weiter reduziert, da eine Analyse bereits auf

funktionaler Ebene zum Ausschluss einer PEA geführt hätte. Unterschiedliche Bezeichnungen von Diensten und Parametern erforderten allerdings höhere Aufwände bei dem Vergleich im Rahmen der Bewertung zur Wiederverwendung.

Die Orchestrierung kann im Fallbeispiel für die Neutralisation unmittelbar beginnen, da die MTPs bereits verfügbar sind. Ein weiterer Vorteil bietet sich bei einheitlichen Typen und dem Ableiten von konkreten Instanzen von diesen in der Orchestrierung, wie in Abbildung 7 im Forschungs-POL Polaris dargestellt. Hier ist gezeigt, dass von einem Typen mehrere Instanzen abgeleitet werden, die im Anschluss für eine typenbasierte Orchestrierung zur Verfügung stehen.

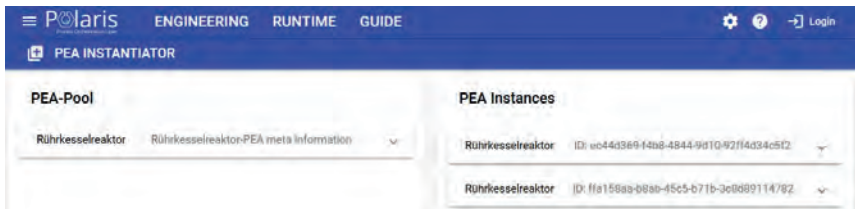


Bild 7: PEA Instanziierung auf Basis einer verfügbaren PEA Spezifikation im PEA-Pool in Polaris (POL)

Parallel zu der Herstellung der PEAs (Abbildung 3) kann in diesem Fall auch die Orchestrierung durchgeführt werden, wie in Abbildung 8 am Beispiel der typenbasierten Anlagentopologie zum Neutralisationsprozess im Process Orchestration Layer des Process-to-Order Lab zu sehen ist. Die Hexagone repräsentieren Platzhalter auf Basis der PEA-Typen und können zudem durch Attribute charakterisiert werden und somit das korrekte Platzieren von PEA-Instanzen unterstützen.

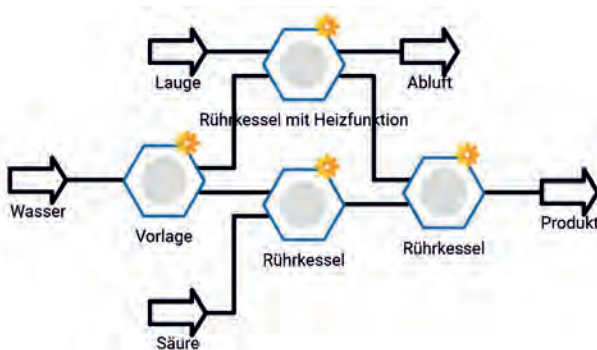


Bild 8: Anlagentopologie des Fallbeispiels auf Basis der Typen

Für den Fall, dass keine passenden PEAs verfügbar sind, kann ein MTP auf Basis bereits bekannter prozesstechnischer Funktionen oder bereits bestehender MTPs als Teil des Lastenheftes für einen Hersteller schnell erstellt werden.

Zusammenfassung und Ausblick

Das typenbasierte Vorgehensmodell hilft bestehenden Anforderungen und den damit verbundenen Herausforderungen zu begegnen. Die entwickelten Typen werden beim typenbasierten Engineering von Hersteller und Betreiber genutzt und tragen zur effektiven Entwicklung einer modularen Anlage bei, da bereits entwickelte Konzepte wiederverwendet werden können. Ein Typ stellt dabei eine Referenz für Hersteller und Betreiber dar. Mit der gemeinsamen Referenz können potentiell folgende Verbesserungen erreicht werden:

- Reduktion von Missverständnissen
- Etablierung einheitlicher Formate zur Spezifikation
- Transformation von der Spezifikation zur Auswahl von PEAs
- Orchestrierung auf Basis von Typen und deren MTPs in frühen Phasen

Es ergeben sich allerdings auch neue Herausforderungen z.B. bei der Definition der Typen und bei der Unterstützung und Etablierung des typenbasierten Planungsprozesses. Das MTP als Teil der funktionalen Anforderungsspezifikation ermöglicht auch die Parallelisierung der Beschaffung von PEAs und der ersten Schritte der Orchestrierung. Das Vorgehensmodell ermöglicht dadurch dem *Betreiber*, den Großteil der Orchestrierung einer modularen Anlage auf Basis von Typen durchzuführen. Die gewonnenen Erkenntnisse fließen in die Standardisierung und zukünftige Entwicklung von PEA-Typen im Process-to-Order Lab ein.

Danksagung

Die Autoren danken dem Bundesministerium für Wirtschaft und Energie sowie dem Projektträger Jülich für die finanzielle Unterstützung im Rahmen des Forschungsprojektes ProRegional (FKZ 03EWS001E).

Die Autoren danken dem Bundesministerium für Bildung und Forschung sowie dem Projektträger Jülich für die finanzielle Unterstützung im Rahmen des Forschungsprojektes eModule (FKZ 03HY116A) der Leitplattform H₂Giga.

Wir danken dem P2O-Lab für die Unterstützung bei der Durchführung der Fallstudie.

Quellen

- [1] Holm, T., 2016. Aufwandsbewertung im Engineering modularer Prozessanlagen: VDI Verlag, doi: 10.51202/9783186465207.
- [2] VDI 2776-2 Entwurf:2022, 'Verfahrenstechnische Anlagen - Modulare Anlagen - Design modularer Anlagen'. VDI, Apr. 2022.
- [3] Klose et al., Orchestration of modular plants Procedure and application for orchestration engineering, (in press).
- [4] Webseite Process-to-Order Lab, online, Zugriff am 15.05.2022, <https://tu-dresden.de/ing/forschung/bereichs-labs/p2o-lab>
- [5] ZVEI, „Modulbasierte Produktion in der Prozessindustrie – Auswirkungen auf die Automation im Umfeld von Industrie 4.0“. ZVEI, Februar 2015.
- [6] VDI 2776-1:2020, 'Verfahrenstechnische Anlagen - Modulare Anlagen - Grundlagen und Planung modularer Anlagen'. VDI, Nov. 2020.
- [7] VDI/VDE/Namur 2658: Automation engineering of modular systems in the process industry. Part 1: General Concept and Interfaces.
- [8] Stenger, F, Urbas, L. (Hrsg.). (2022). VDI-Handlungsempfehlung - Modulare Anlagen - Paradigmenwechsel im Anlagenbau: Zusammenspiel von Prozesstechnik und Automatisierungstechnik. VDI.
- [9] Stutz et al. "Der Process Orchestration Layer – Ein Überblick über die Gestaltungsmöglichkeiten für die prozessleittechnischen Systeme modularer Produktionsanlagen," in Automation 2020: Shaping Automation for our Future, 1st ed., VDI Wissensforum GmbH, Ed. Düsseldorf: VDI Verlag, 2020, pp. 823–842. doi: 10.51202/9783181023754-823.
- [10] Klose et al. (2019). "Anforderungen an die Orchestrierung für modulare Prozesseinheiten". 10.25532/OPARA-29.
- [11] Lorenz et al. "MicroFrontends as Opportunity for Process Orchestration Layer Architecture in Modular Process Plants," in 2021 26th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), Sep. 2021, pp. 01–04. doi: 10.1109/ETFA45728.2021.9613474.
- [12] Graube et al. (2019). Rezeptgetriebene Prozessführung in modularen Anlagen. Konzepte und Erfahrungen aus dem P2O-Lab Dresden. In: VDI Wissensforum GmbH (Ed.), Automation 2019. VDI Verlag, Düsseldorf, pp. 255–268. <https://doi.org/10.51202/9783181023518-255> (zuletzt abgerufen am 07. Januar 2022)

- [13] Schindel et al. (2021). General Approach for Technology and Process Equipment Assembly (PEA) Selection in Process Design. Chemical Engineering and Processing; Process Intensification 159, 108223 <https://doi.org/10.1016/j.cep.2020.108223> (abgerufen 07.01.2022)
- [14] Fleischer-Trebes et al. 2017. Planungsansatz für modulare Anlagen in der chemischen Industrie. Chemie Ingenieur Technik 89, 785–799. <https://doi.org/10.1002/cite.201600083>
- [15] Bramsiepe et al. 2012. Die 50 %-Idee: Modularisierung im Planungsprozess. Chemie Ingenieur Technik 84, 581–587. <https://doi.org/10.1002/cite.201100250>
- [16] Sievers et al. 2017. Lead time estimation for modular production plants. Chemical Engineering Research and Design 128, 96–106. <https://doi.org/10.1016/j.cherd.2017.10.003>
- [17] Radatz et al. 2017. Design of equipment modules for flexibility. Chemical Engineering Science 168, 271–288. <https://doi.org/10.1016/j.ces.2017.04.021>
- [18] Hohmann et al. 2017. Modules in process industry – A life cycle definition. Chemical Engineering and Processing: Process Intensification 111, 115–126. <https://doi.org/10.1016/j.cep.2016.09.017>
- [19] J. Rahm et al., 'Efficient Automation Engineering of Modular Process Equipment Assemblies Using the Digital Twin', *Chemie Ingenieur Technik*, vol. 93, no. 12, pp. 2081–2091, Dec. 2021, doi: 10.1002/cite.202100100.
- [20] Polley et al., "Automated Facility Plug-And-Play - Stirred Tank Unit Interface Specification." BioPhorum Operations Group Ltd, Feb. 2021. [Online]. Available: <https://www.biophorum.com/download/automated-facility-stirred-tank-unit-interfacespecification/>
- [21] Alvarez et al. "ph neutralization process as a benchmark for testing nonlinear controllers", *Industrial & engineering chemistry research* 40 (11) (2001) 2467–2473.
- [22] Radatz et al. 2019. Selection of equipment modules for a flexible modular production plant by a multi-objective evolutionary algorithm. Computers & Chemical Engineering 123, 196–221. <https://doi.org/10.1016/j.compchemeng.2018.12.009>

Datenbasierte Überwachung von reproduzierbaren, chemischen Batchprozessen für modellbasierte Sicherheitsfunktionen

B.Sc. **Philipp Langlitz**, Dipl.-Ing. **Felix Lenhart**,
Dr.-Ing. **Martin Roser**, BASF SE, Ludwigshafen

Kurzfassung

In komplexen und modellbasierten PLT-Sicherheitsfunktionen werden unsicherheitsbehaftete Prozessmesswerte zur Berechnung von nicht direkt beobachtbaren Abschaltwerten verwendet, was bei Nutzung der sicherheitstechnischen Messgenauigkeitsangaben der Hersteller und üblichen Fehlerfortpflanzungsmethoden zu hohen Sicherheitsabschlägen führt.

Dieser Beitrag resümiert fortschrittliche Online-Diagnosefunktionalitäten für PLT-Sicherheitsfunktionen und skizziert einen Ansatz zur datenbasierten Überwachung von reproduzierbaren, chemischen Batchprozessen. Das Potenzial dieses Verfahrens zur Erhöhung des Diagnosegrads und Erkennung unbekannter Anomalien im Produktionsprozess wird anhand von historischen Prozessdaten aufgezeigt und diskutiert.

Zukünftig könnte dieser Ansatz die Nutzung von geringeren Sicherheitsabschlägen auf Basis der Messwiederholbarkeit rechtfertigen, was sich äußerst positiv auf die Betreibbarkeit von komplexen und modellbasierten PLT-Sicherheitsfunktionen auswirken würde.

1. Motivation

In PLT-Sicherheitsfunktionen müssen Zeitverzögerungen, Prozessdynamik und die Genauigkeit der Messketten zum Auslösepunkt als Sicherheitszuschlag berücksichtigt werden. Aussagen zu absoluten Messgenauigkeiten bedürfen Annahmen zu Fertigungsqualität und Qualitätssicherung beim Gerätehersteller, Einsatzbedingungen beim Anwender sowie Kenntnisse über die Signalübertragung in der kompletten Messkette. Da ist es nicht verwunderlich, dass Herstellerangaben zu sicherheitstechnischen Messgenauigkeiten – sofern sie überhaupt in Sicherheitshandbüchern zu finden sind – recht üppig ausfallen. Bei eigenen Betrachtungen des Anwenders werden üblicherweise Worst-Case-Szenarien über die komplette Messkette betrachtet, was häufig in großen, zu berücksichtigenden Messgenauigkeiten von 2-5% (bezogen auf den Messspannenendwert) resultiert.

Bei einfachen PLT-Sicherheitsfunktionen (TZ+, PZ+, ...) werden die Sicherheitsmargen in aller Regel maßgeblich durch verfahrenstechnische Betrachtungen bestimmt und sind meist groß gegenüber den Messgenauigkeiten der Messkette. In komplexen und modellbasierten Sicherheitsfunktionen werden hingegen mehrere unsicherheitsbehaftete Prozessmesswerte zur

Berechnung eines nicht direkt beobachtbaren Abschaltwertes verwendet. Übliche Fehlerfortpflanzungsmethoden liefern dabei äußerst hohe Werte für einen zusätzlichen Sicherheitsabschlag, was die Betriebbarkeit signifikant einschränkt.

Auf der anderen Seite werden in der Prozessindustrie qualitativ hochwertige, betriebsbewährte Geräte eingesetzt, die viele Funktionen zur Selbstüberwachung mitbringen und eine hohe Messwiederholbarkeit aufweisen. Eine Kombination aus Online-Sensordiagnose und datenbasierter Prozessüberwachung eröffnet somit eine Möglichkeit zur Rechtfertigung von höheren Messgenauigkeitsannahmen, insbesondere für reproduzierbare Batchprozesse.

2. Online-Diagnosefunktionen

Das Versagen einer PLT-Sicherheitsfunktion kann prinzipiell auf systematische oder zufällige Fehler zurückgeführt werden. Für die Gesamtausfallrate der zufälligen, gefährlichen, unerkannten Fehler (λ_{DU}) wird bereits bei der Planung der PLT-Sicherheitsfunktion mit Hilfe der Herstellerangaben aller PLT-Komponenten der Nachweis geführt, dass sie innerhalb des geforderten Sicherheitsintegritätslevels (SIL) liegt. Systematische Fehler werden durch organisatorische Maßnahmen des Betreibers im Sicherheitslebenszyklus beherrscht.

Zu welcher Fehlerart gehört nun die Messgenauigkeit? Wird bei Kalibration eines Sensors unter Laborbedingungen eine Messabweichung außerhalb der Genauigkeitsangaben des Herstellers festgestellt, liegt ein zufälliger Sensorfehler vor. Dieser ist in der vom Hersteller angegebenen Ausfallrate λ_{DU} des Gerätes berücksichtigt.

Die Identifikation aller weiteren Effekte, die Einfluss auf die Messgenauigkeit der Messkette unter Betriebsbedingungen üben, unterliegen dem Betreiber und müssen mit einem hinreichend großen Sicherheitszuschlag in der Auslegung der PLT-Sicherheitsfunktion bedacht werden. Sie sind klar den systematischen Fehlern zuzuordnen. Dem Betreiber obliegt somit die Aufgabe, mit Sicherheitszuschlägen und organisatorischen Maßnahmen sicherzustellen, dass Messabweichungen nicht zu einem gefährlichen, unerkannten Fehler führen.

Typische Maßnahmen für redundante Messketten sind Online-Diagnosefunktionen auf Basis eines Messwertevergleichs der redundanten Kanäle [1,2]:

- Plötzliche (unerkannte) Geräteausfälle äußern sich durch Einfrieren von Messwerten oder Bereitstellung eines vordefinierten Ersatzwertes. Beides führt zu falschen Messwerten des betroffenen Sensors. In redundanten Messketten können diese Fehler durch einen A/B-Kanalvergleich mit Abweichungsüberwachung zwischen beiden Kanälen erkannt werden.
- Manche systematischen Fehler bei Auslegung, Inbetriebnahme und Betrieb sind zwar an den stationären Betriebspunkten der PLT-Sicherheitsfunktion nicht beobachtbar, führen

aber teilweise in transienten Prozessphasen zu zeitlich begrenzten, hohen Abweichungen innerhalb der redundanten Messwerte einer Messkette.

- Weitaus herausfordernder sind Abweichungen aufgrund von Alterung von Hardwarekomponenten (Drifteffekte). Diese entwickeln sich langsam während der Lebensdauer des Sicherheitssystems. In frühen Stadien – wenn die Abweichung zwischen den redundanten Sensoren noch im akzeptablen Bereich liegt – sind sie selbst bei periodischen Inspektionen kaum zu identifizieren. Eine Online-Diagnose von Drifteffekten liefert einen wertvollen Beitrag zur Einhaltung von hohen Messgenauigkeiten für relevante Messketten, sie unterliegt jedoch der Herausforderung, die relevanten Driftinformationen aus dem Prozesssignal und anderen Verzerrungen zu isolieren.

Diese Maßnahmen sind wichtige Online-Diagnosemöglichkeiten für redundante Messketten. Nicht-redundante Messketten sowie zeitgleich auftretende Fehler gemeinsamer Ursache können durch diese Vergleichsmethoden jedoch prinzipbedingt nicht erkannt werden. Zudem hängt der Produktionsprozess von vielen weiteren Faktoren ab, wie beispielsweise Aktoren, Regler, Einsatzstoffe oder Umgebungsbedingungen. Hier wäre der Schritt von einer Genauigkeit bezogen auf Messfehler hin zu einer Genauigkeit bezogen auf die Reproduzierbarkeit des Produktionsprozesses wünschenswert, da dies eine umfassendere Möglichkeit zur Erkennung von Anomalien bietet und das Ergebnis vieler Einflüsse bewertet, die weit über den Sensorvergleich von redundanten Messketten hinaus gehen.

3. Messgenauigkeit, Wiederholgenauigkeit und Reproduzierbarkeit eines Prozesses

In diesem Beitrag wird ein Ansatz zur datenbasierten Überwachung von reproduzierbaren Prozessen vorgestellt. Dies bedarf zunächst die Abgrenzung der Begriffe Messgenauigkeit und Wiederholgenauigkeit, wie in Abb. 1 illustriert. Während die Messgenauigkeit ein Maß für die Nähe der Messungen zum wahren Prozesswert darstellt (Abb. 1a), beschreibt die Wiederholgenauigkeit die Fähigkeit eines Messinstruments, bei sich wiederholenden Messaufgaben konstante Messwerte zu liefern.

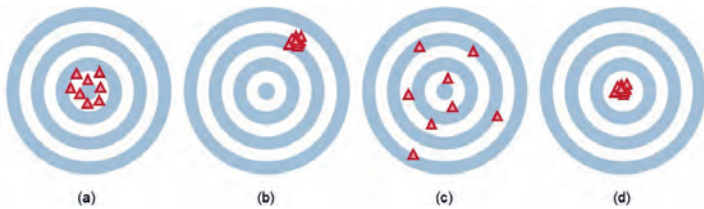


Bild 1: Verdeutlichung von Mess- und Wiederholgenauigkeit. Der wahre Prozesswert befindet sich in allen Fällen im Zentrum der Zielscheibe. Die Messungen weisen eine (a) hohe Messgenauigkeit und geringe Wiederholgenauigkeit, (b) geringe Messgenauigkeit und hohe Wiederholgenauigkeit, (c) geringe Messgenauigkeit und geringe Wiederholgenauigkeit, bzw. (d) hohe Messgenauigkeit und hohe Wiederholgenauigkeit auf.

Aufeinanderfolgende Messungen liegen also nahe beieinander, was nicht notwendigerweise dem wahren Prozesswert entspricht (Abb. 1b). Sofern eine hohe Wiederholgenauigkeit vorliegt, kann die Genauigkeit des Messgeräts durch Kalibrierung und Justage verbessert werden (Abb. 1d). Zur Erhöhung der Messgenauigkeit in einer Prozessanlage würde dies jedoch eine regelmäßige, aufwändige in situ Kalibrierung und Justage der Messkette am Arbeits- bzw. Auslösepunkt der PLT-Sicherheitsfunktion bedeuten.

Im Gegensatz zu den beiden Genauigkeitsdefinitionen wird in diesem Beitrag die Reproduzierbarkeit als Wiederholbarkeit von Ergebnissen verstanden. Die Reproduzierbarkeit eines chemischen Prozesses bedeutet, ein Produkt wiederholbar (innerhalb der geforderten Produktspezifikation) herzustellen, beispielsweise trotz bedienerbedingter Abweichungen.

Abb. 2a verdeutlicht den üblichen Umgang mit sicherheitstechnischen Messgenauigkeiten: Für die unsicherheitsbehafteten Messgrößen X_A , X_B , X_C werden maximale, sicherheitstechnische Fehlermargen EB_A , EB_B , EB_C definiert und mittels Methoden der Fehlerfortpflanzung eine konservativen Sicherheitsmarge EB_Y für die Prozessausgangsgröße Y der PLT-Sicherheitsfunktion berechnet. Je nach Anzahl der Messketten und mathematischer Zusammenhänge führt dies zu unrealistisch hohen Sicherheitsabschlägen (Abb. 2a, blau).

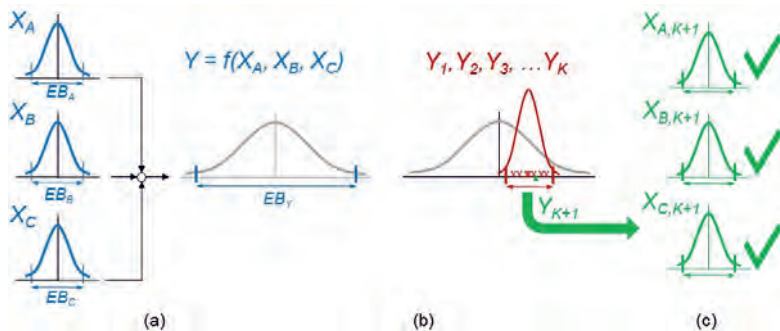


Bild 2: (a) Die Verwendung vieler unsicherheitsbehafteter Sensorwerte X_A , X_B , X_C führt zu großen Messgenauigkeitsmargen für die Ausgangsgröße Y (blau). (b) Bei reproduzierbaren Batchverläufen

$Y_1, Y_2, Y_3, \dots, Y_K$ spielt die absolute Messgenauigkeit eine untergeordnete Rolle, sondern vielmehr die Wiederholgenauigkeit von Y_{K+1} in Bezug auf die vorherigen Ergebnisse $Y_1, Y_2, \dots, Y_K$ (rot). (c) Überwachung der Wiederholgenauigkeit Y_{K+1} lässt auf die Fehlerfreiheit aller Eingangssignale sowie einen validen Batchverlauf schließen.

Zur gleichen Zeit weisen die überwachten Prozesse aufgrund von automatisierten Fahrweisen, hochwertigen Regelungen und hohen Qualitätsanforderungen an das Produkt eine äußerst hohe Reproduzierbarkeit über die verschiedenen Chargen 1, 2, 3, ...K auf (Abb. 2b, rot). Unter der Voraussetzung eines fehlerfreien Prozesses mit hoher Reproduzierbarkeit und fehlerfreien Messgeräten mit hoher Wiederholgenauigkeit weist die Prozessausgangsgröße Y_{K+1} von Charge K+1 ebenfalls eine hohe Wiederholgenauigkeit hinsichtlich der Prozessausgangsgrößen $Y_1, Y_2, Y_3, \dots, Y_K$ der historischen Chargen auf.

Im Umkehrschluss deutet bei einer möglichst hohen Online-Diagnoseabdeckung der Messketten (siehe Kapitel 2) und validierten, historischen Chargen 1, 2, 3, ...K eine erwartungstreue Abbildung der Prozessausgangsgröße Y_{K+1} mit hoher Wiederholgenauigkeit auf einen fehlerfreien Betrieb des Gesamtsystems hin (Abb. 2c).

Dieser Ansatz bietet eine deutlich umfassendere Möglichkeit zur Erkennung von Anomalien und schließt viele Einflüsse mit ein, die weit über den Sensorvergleich von redundanten Messketten hinaus gehen. Dazu gehören beispielsweise zeitgleich auftretende Drifteffekte in redundanten Messketten, Einsatzstoffe außerhalb der Spezifikation, falsche Rezeptfahrweisen, Verunreinigungen und Fehler von Akteuren.

4. Grundlagen zur datenbasierten Überwachung

Bei datenbasierten Überwachungsmethoden wird im Allgemeinen aus historischen Daten ein empirisches Referenzmodell entworfen, welches eine möglichst niedrigdimensionale Nachbildung des charakteristischen Produktionsverfahrens darstellt. Bei der Art der Modellierung können merkmalsorientierte und zeitaufgelöste sowie lineare und nichtlineare Ansätze unterschieden werden.

In diesem Beitrag wurden Modelle der multivariaten statistischen Prozesslenkung (engl. Multivariate Statistical Process Control, MSPC) verwendet. Dies sind lineare, zeitaufgelöste Modelle, die eine effiziente und umfangreiche Beschreibung der Chargen unter Berücksichtigung von zeitvarianten Effekten ermöglichen. Somit können die konkreten Zeitpunkte von Prozessabweichungen und Anomalien im reproduzierbaren Batchprozess präzise bestimmt werden [3].

Zu den gängigen MPSC Methoden gehören die Hauptkomponentenanalyse (engl. Principle Component Analysis, PCA) und die Regression der partiellen kleinsten Quadrate (engl. Partial Least Square, PLS). Sie wirken beide dem Dimensionalitätsproblem entgegen, indem sie umfangreiche Datensätze von Prozess- und Qualitätsvariablen $\{X, Y\}$ in wenige, möglichst aussagekräftige Linearkombinationen TP^T bzw. UQ^T unter Minimierung der quadratischen Fehler E bzw. F zerlegen,

$$X = TP^T + E \quad (1)$$

$$Y = UQ^T + F \quad (2)$$

mit den sogenannten „Score“-Matrizen T, U und „Loadings“-Matrizen P, Q .

Während die PCA ausschließlich Prozessvariablen X gemäß Gl. (1) betrachtet, werden bei der PLS Regression auch die Prozessausgangs- bzw. Qualitätsvariablen Y nach Gl. (2) berücksichtigt. Es besteht ein linearer Zusammenhang zwischen den Prozess- und Qualitätsvariablen, der mit Hilfe der Regressionsmatrix R wie folgt beschrieben werden kann [4]:

$$R = W(P^TW)^{-1}Q^T \quad (3)$$

Die PLS Regression wird bevorzugt verwendet, wenn das Ziel auch darin besteht, Auswirkungen von Änderungen der Prozessvariablen auf die Ausgangsvariablen zu bestimmen. Durch die Einbeziehung der Qualitätsvariablen ist die PLS Regression robust gegenüber Abweichungen in den Prozessvariablen, die keine Auswirkungen auf die Produktqualität haben. Bei der PCA würden solche Abweichungen hingegen stets als Anomalie erkannt werden [4].

Die zur Modellerstellung benötigten Prozessdaten $x_{i,j,k}$ liegen in einem dreidimensionalen Zusammenhang vor, wobei $i \in [1, \dots, I]$ die verwendeten Chargen, $j \in [1, \dots, J]$ die ausgewählten Prozessvariablen und $k \in [1, \dots, K]$ die konkreten Zeitpunkte innerhalb einer Charge definieren. Zur Verwendung in Gl. (1) müssen diese Daten entsprechend Abb. 3 entfaltet werden. Hierbei kann zwischen dem sogenannten Batch-Wise-Unfolding und Observation-Wise-Unfolding unterschieden werden.

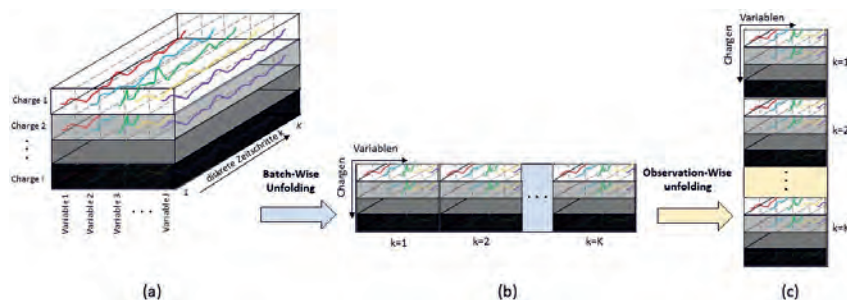


Bild 3: Daten Vorverarbeitungs Techniken. (a) Transformation mehrdimensionaler Prozessdaten in eine 2D Matrix mittels (b) Batch-Wise Unfolding, (c) Observation-Wise Unfolding.

Im Observation-Wise-Unfolding (Abb. 3c) bilden die Zeilen der zweidimensionalen Matrix X die Variablen $j \in [1, \dots, J]$ für jede Charge und zu jedem Zeitpunkt ab. Die Hauptkomponenten modellieren also die Unterschiede in den Messungen über die Zeit. Dieses Modell unterliegt der Annahme, dass sich die Struktur der verwendeten Variablen j über die Zeit nicht verändert und der statistische Durchschnitt zur Beschreibung des Prozesses ausreicht. Aufgrund der Verwendung des statistischen Durchschnitts der Prozessvariablen, kann schon mit wenigen Batchdaten ein aussagekräftiges Modell erstellt werden. Allerdings wird der Dynamik und dem nichtlinearen, zeitlich variierendem Verhalten von Batchprozessen nicht hinreichend Rechnung getragen [4].

Beim Batch-Wise-Unfolding (Abb. 3b) hingegen, bildet jede Zeile die Variablen $j \in [1, \dots, J]$ für alle Zeitpunkte $k \in [1, \dots, K]$ einer Charge ab. Durch die zeitabhängige Strukturierung der Variablen wird die Dynamik eines chemischen Prozesses wesentlich genauer modelliert [4]. Allerdings müssen die Zeitverläufe der unterschiedlichen Chargen synchronisiert werden [5]. Abb. 4 illustriert übliche Verfahren zur Datensynchronisation. Das einfachste Verfahren besteht in der Beschneidung aller Batchdaten unterschiedlicher Länge unter Beibehaltung der Zeitachse

(Abb. 4a). Zeitliche Variationen innerhalb der Chargen, beispielsweise markante Zeitpunkte eines Rezepts, werden dabei nur unzureichend zugeordnet.

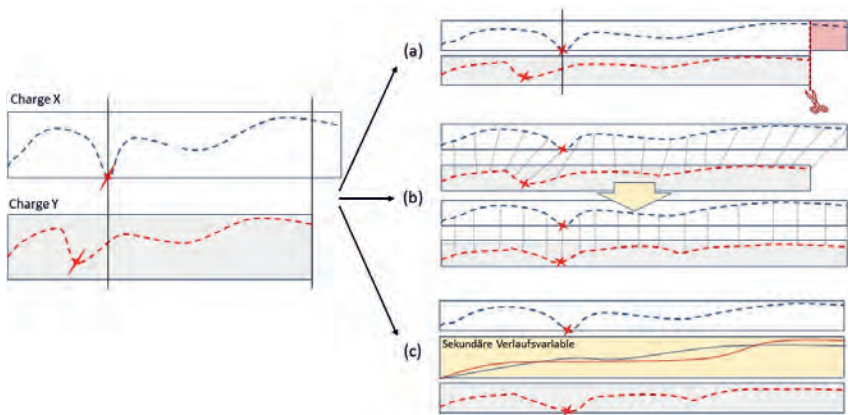


Bild 4: Methoden zur Synchronisation von Prozessdaten.

- (a) Beschneidung, (b) Dynamic Time Warping,
- (c) Normierung mittels Sekundärvariablen.

Abb. 4b zeigt hingegen ein Verfahren der dynamischen Zeitnormierung (engl. Dynamic Time Warping, DTW). Dies ermöglicht eine genaue Synchronisation der Daten, erweist sich jedoch für Echtzeitanwendungen als äußerst aufwändig und rechenintensiv. Bei automatisierten Batchprozessen bietet sich meist eine Normierung auf Basis von aussagekräftigen Sekundärvariablen an (Abb. 4c), da dadurch eine schnelle und ausreichend genaue Synchronisation auf wichtige Stellen im Rezeptverlauf erfolgen kann.

5. Modellierung

Für die Modellbildung wurden insgesamt 75 Chargen unterschiedlicher Rezepte durch eine manuelle Qualitätskontrolle in drei Qualitätsklassen unterteilt. Klasse 1 (48 Chargen) kennzeichneten sich durch eine ideale Fahrweise und einwandfreiem Ergebnis aus. Die 14 Chargen der Klasse 2 wiesen zwar Besonderheiten in der Fahrweise auf, die Qualität des Endprodukts befand sich jedoch innerhalb der Spezifikation. Für Klasse 3 (13 Chargen) wurden Batchverläufe ausgewählt, bei denen entweder eine unzureichende Produktqualität festgestellt wurde, Fehler in den Produktionsabläufen dokumentiert wurden, oder es wurden Chargen der Klasse 1 und 2 verfremdet, um Anomalien zu simulieren. Die ersten beiden Klassen wurden

in der Trainingsphase zur Modellbildung genutzt. Chargen der Klasse 3 stellten Testdatensätze zur Erkennung von Anomalien dar.

Als Prozessvariablen wurden Zulauftemperaturen, Kühlwasserdurchfluss, Kühlwassertemperaturen sowie die Reaktortemperatur über die Zulaufmenge (Sekundärvariable) gemäß Abb. 5 verwendet, die mittels Batch-Wise-Unfolding aufbereitet und für das Anlernen eines PLS Regressionsmodells mit bis zu fünf Hauptkomponenten verwendet wurden. Als Qualitätsvariable fungierte der prozentuale Umsatz.



Bild 5: Schematischer Aufbau des Batchreaktors und den verwendeten Messungen.

Die Bewertung des Modells erfolgte durch eine Kreuzvalidierung mit fünf Teilmengen, wobei zur korrekten Erfassung der zeitlichen Dynamik die Prozessdaten der einzelnen Chargen jeweils komplett einer Teilmenge zugeordnet wurden.

Abb. 6 zeigt den mittleren quadratischen Modellfehler (engl. Mean-Squared-Error, MSE) für die ersten fünf Hauptkomponenten. Der MSE bei drei Hauptkomponenten ist unwesentlich höher als bei Verwendung von vier Hauptkomponenten und weist eine deutlich geringere Streuung der MSE-Werte auf. Aus diesem Grund wurde das PLS Regressionsmodell für die folgenden Ergebnisse mit drei Hauptkomponenten angelernt.

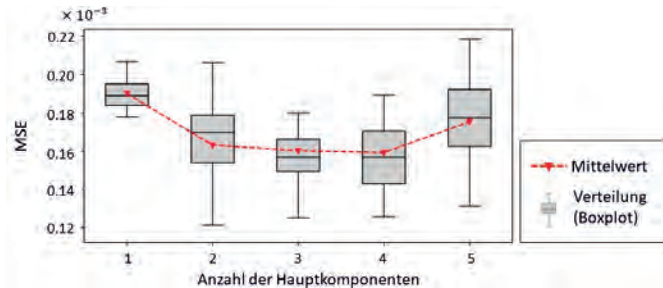


Bild 6: Ergebnis der Kreuzvalidierung einer PLS-Regression mit bis zu fünf Hauptkomponenten.

Mittelwert (rot) und Verteilung (schwarz) des mittleren, quadrierten Fehlers (MSE).

Um die Parametrierung der PLS Regression zu diskutieren, werden in Abb. 7 die Loadings-Matrizen P und Q illustriert. Jeder Punkt stellt die Gewichtung der jeweiligen Variablen zu einem gewissen Zeitpunkt im Prozessfortschritt dar. Der Abstand der Punkte zum Mittelpunkt zeigt den Einfluss der Variablen auf die jeweilige Hauptkomponente. Die Gewichtung von Zulauftemperatur (rot), Durchflussmenge des Kühlwassers (lila) und prozentualem Umsatz (hellblau) bleiben nahezu konstant und können als „statische Variablen“ angesehen werden. Die Einflüsse von Reaktor- (blau) und Kühlwassertemperaturen (orange/grün) weisen hingegen eine ausgeprägte Zeitvarianz auf (dynamische Variablen), was die Wahl des Batch-Wise-Unfoldings legitimiert. Zudem ist im Zentrum von Abb. 7 ein unstrukturierter Bereich mit Punkten verschiedener Prozessvariablen zu erkennen, an dem sich die große Volatilität des Prozesses zu Beginn zeigt. Dieser Bereich wird im Folgenden vernachlässigt.

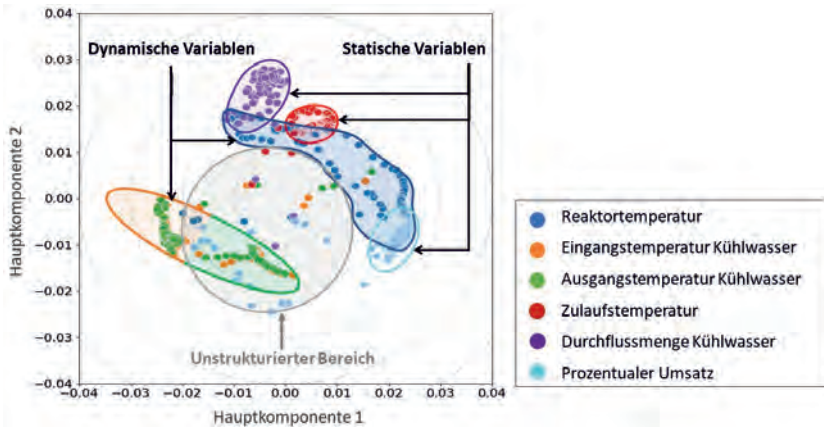


Bild 7: Unterteilung der Loadings der Prozessvariablen und Qualitätsvariable in dynamische, statische und unstrukturierte Bereiche.

6. Überwachung von reproduzierbaren Batchprozessen

Für die nähere Untersuchung der Modelleignung zur Überwachung von reproduzierbaren Batchprozessen wurden in Abb. 8 Chargen der Klasse 3 verwendet, bei denen nach etwa 30 Prozent des Prozesses ein starker Anstieg der Reaktortemperatur durch unzureichende Kühlmöglichkeiten (Charge 1-3) bzw. aufgrund eines feststehenden Regelventils (Charge 4) simuliert wurde. Abb. 9 zeigt zu diesen Chargen die Scores über die ersten beiden Hauptkomponenten zum Zeitpunkt der Anomalie (Abb. 9a) sowie zum Ende der Charge (Abb. 9b).

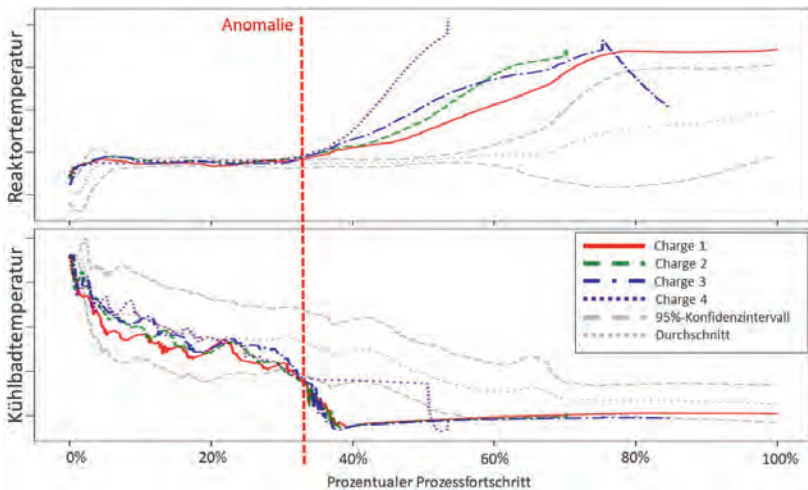


Bild 8: Verlauf von Reaktortemperatur und Eingangstemperatur des Kühlwassers von vier Chargen mit einer Anomalie bei etwa 30 Prozent des Prozessfortschritts

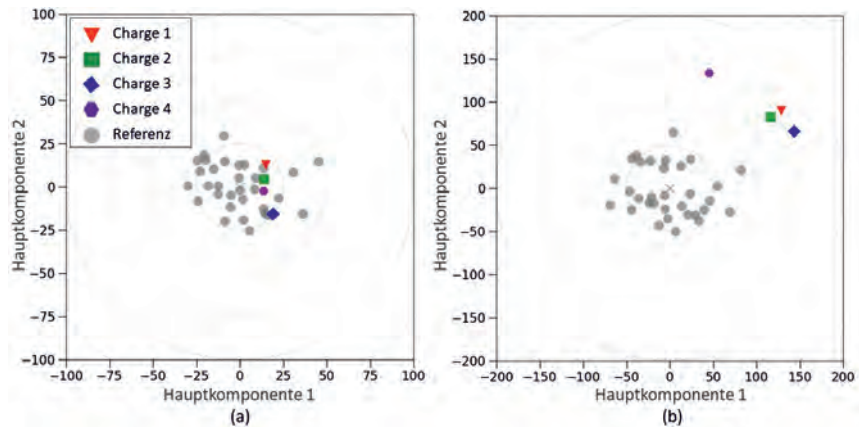


Bild 9: (a) Scores der Anomalie-Chargen zum Zeitpunkt der Anomalie und (b) am Ende der Chargen.

Zum Vergleich werden die Referenz-Scores der Trainingsdaten (grau) mit abgebildet. Es zeigt sich deutlich, dass zum Zeitpunkt des Auftretens der Anomalie die Scores der fehlerhaften Chargen nicht von den Referenz-Scores unterscheidbar sind (Abb. 9a). Zur Online-Diagnose

sind die Scores daher nur bedingt nutzbar. Zum Ende der Chargen liegen die Scores jedoch weit außerhalb des Referenzbereichs (Abb. 9b). Dies ermöglicht eine Validierung Chargen bzw. eine Klassifikation der Anomalie auf Basis der Verschiebung der Score-Werte.

Eine elegante Möglichkeit zur Online-Diagnose bietet sich durch Betrachtung des Modellfehlers. Das PLS Regressionsmodell wurde anhand von Chargen idealer bis akzeptabler Fahrweise sowie guter Endqualität (Klasse 1 und 2) angelern und zeigt, dass es Testdaten dieser Klassen mit geringem Modellfehler abbilden kann (Abb. 6). Tritt nun eine Anomalie in einer Charge auf, passt das angelernte Modell nicht mehr zu den vorliegenden Prozessdaten. Wie in Abb. 10 dargestellt, steigt der Modellfehler signifikant an. An dieser Stelle sei herausgestellt, dass bereits vor dem Zeitpunkt der Anomalie der Modellfehler die Konfidenzintervalle der Trainingsdaten deutlich übersteigt. Somit ist eine prinzipielle Eignung zur Online-Diagnose von Anomalien vorhanden.

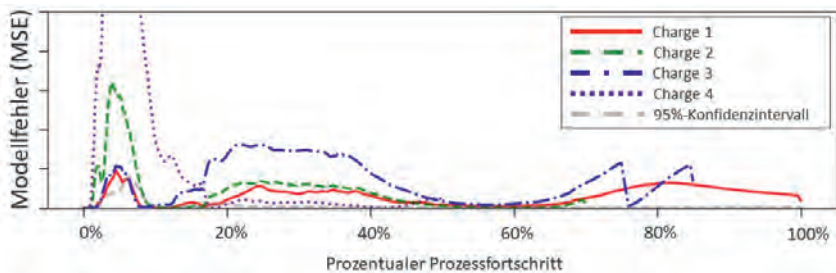


Bild 10: Der Modellfehler der vier Anomalie-Chargen über den Prozessfortschritt.

Abb. 11 zeigt die Verteilung des mittleren Modellfehlers für alle untersuchten Chargen der Qualitätsklassen 1 bis 3. Der mittlere Modellfehler für Chargen mit unzureichender Produktqualität oder Anomalien (Klasse 3) liegt um eine Zehnerpotenz über dem Modellfehler der übrigen Klassen. Dies ermöglicht eine Prozessvalidierung anhand des Modellfehlers.

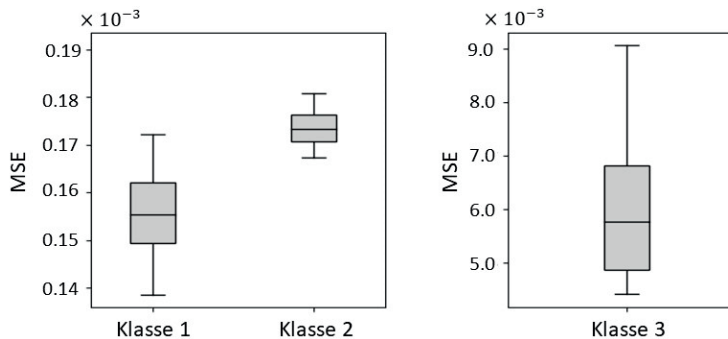


Bild 11: Modellfehler von Chargen unterschiedlicher Qualitätsklassen.

7. Zusammenfassung und Ausblick

Die datenbasierte Überwachung von reproduzierbaren, chemischen Batchprozessen mit Hilfe von Verfahren der multivariaten statistischen Prozesslenkung (MSPC) ergänzt etablierte On-line-Diagnosefunktionalitäten für redundante Messketten. Sie besitzt das Potenzial, den Diagnosegrad signifikant zu erhöhen und sicherheitstechnische sowie qualitätsrelevante Anomalien jenseits der typischen Sensoreinzelfehler aufzudecken, beispielsweise zeitgleich auftretende Drifteffekte in redundanten Messketten, Verwendung von Einsatzstoffen außerhalb der Spezifikation, falsche Rezeptfahrweisen, Verschmutzung oder Stellgerätedefekte.

Eingebettet in den Gesamtkontext eines Anlagensicherheitskonzepts kann dieser Ansatz die Nutzung von geringeren Sicherheitsabschlägen auf Basis der Messwiederholbarkeit rechtfertigen, was zu einer deutlichen Erweiterung des Betriebsfensters von komplexen und modellbasierten PLT-Sicherheitsfunktionen führt.

8. Literaturangaben

- [1] Hauff et al.; Genauigkeit von Messgeräten überwachen; atp edition vol. 55 2013; DI Verlag
- [2] Lenhart, F., Roser, M., Kleinert, T.; Dealing with Measurement Uncertainties for Safety Applications in Process Industry; 12. Internationales TÜV Rheinland Symposium, Funktionale Sicherheit und Security in industriellen Anwendungen; 2016
- [3] Rendall, R., Chiang, L.H., Reis, M. S.; Data-driven methods for batch data analysis – A critical overview and mapping on the complexity scale; Elsevier Ltd.; 2019
- [4] Zhiqiang G., Zhihuan S.; Multivariate Statistical Process Control, Springer-Verlag, 2013.
- [5] Wan, J., Marjanovic, O., Lennox, B.; Uneven batch data alignment with application to the control of batch end-product quality; Elsevier Ltd.; 2013

Security-Entscheidungen „by Design“ in das Engineering prozesstechnischer Anlagen integrieren

Konzept der „Automation Security by Design Decisions“

Integrating security decisions “by design” into the engineering of process plants

Introducing the “Automation Security by Design Decisions” concept

S. Fluchs, admeritia GmbH, Langenfeld,
Helmut-Schmidt-Universität/Universität der Bundeswehr, Hamburg;
E. Tastan, Prof. Dr.-Ing. **R. Drath**, Hochschule Pforzheim;
Dr. **M. Mertens**, INEOS Manufacturing Deutschland GmbH, Köln;
J. Ritter, Dr. **A. Horch**, HIMA Paul Hildebrandt GmbH, Brühl;
Prof. Dr.-Ing. **A. Fay**, Helmut-Schmidt-Universität/Universität der
Bundeswehr, Hamburg

1 Kurzfassung

Die intensivere Vernetzung automatisierungstechnischer Anlagen erhöht die Wichtigkeit wirk-samer Security-Maßnahmen. Sinnvolle Maßnahmen sind selten klar abgrenzbare, dedizierte Security-Komponenten, sondern bestehen aus vielen größeren und kleineren Security-Entscheidungen für operative Komponenten. Doch um das Design der Automatisierungssysteme noch beeinflussen zu können, müssen diese Entscheidungen früh genug während des Automation-Engineering-Workflows getroffen werden („Security by Design“). *Security by Design* scheitert bis heute oft daran, dass die beteiligten Ingenieure gar nicht wissen, welche Security-Entscheidungen sie wann treffen müssten – beziehungsweise welche Security-Entscheidungen sie mit ihren Design-Festlegungen implizit treffen.

Der vorliegende Beitrag klassifiziert existierende „Security by Design“-Ansätze hinsichtlich Basis-Workflows und Security-Integrationsmechanismen und ergänzt sie durch ein „Security by Design“-Konzept, das auf Security-Entscheidungen als zentralem Element aufbaut. Ein Security-Entscheidungspunkt bündelt eine in sich abgeschlossene Menge von Security-Parametern und betrifft eine klar abgrenzbare Menge operativer Funktionen. Die Entscheidungspunkte bevorzugen Security-Maßnahmen, die durch die Nutzung systemimmanenter Eigenschaften umsetzbar sind und keine zusätzlichen, dedizierten Security-Lösungen benötigen.

Gemeinsam mit einer Anleitung für das systematische Treffen der Security-Entscheidungen befähigt dies die Ingenieure der am Automation-Engineering-Workflow beteiligten Gewerke, „ihre“ Security-Entscheidungen auch ohne dediziertes Security-Knowhow selbst zu treffen.

2 Abstract

The increasing networking of automation systems boosts the importance of effective security measures. Reasonable measures are only rarely clearly delineated, dedicated security components, but are instead composed of a hodgepodge of larger and smaller security decisions for operational components. However, in order to still be able to influence the design of automation systems, these decisions must be made early enough during the automation engineering process (*"security by design"*). *Security by design* often fails because the engineers involved do not know which security decisions they have to make and when - or which security decisions they implicitly make with their design specifications.

This paper classifies existing security-by-design approaches with respect to their base processes and security integration mechanisms and complements them with a security-by-design concept based on security decisions as a central element. A security decision point bundles a self-contained set of security parameters and affects a clearly defined set of operational functions. The decision points favor security measures that can be implemented by leveraging inherent system characteristics and do not require additional, dedicated security solutions. Together with a guideline for a systematic decision-making process, this enables the engineers of the trades involved in the automation engineering process to make "their" security decisions themselves, even without dedicated security know-how.

3 Warum brauchen wir *Security by Design*?

Aufgrund der intensiveren Vernetzung automatisierungstechnischer Anlagen unter zunehmender Nutzung von Internettechnologien und gängigen IT-Produkten nehmen die möglichen und leicht erreichbaren Kommunikationsschnittstellen zu Automatisierungssystemen zu. Dies macht sie anfälliger und lohnenswerter für kriminelle Cyberangriffe und erhöht die Wichtigkeit wirksamer Security-Maßnahmen zur Verhinderung solcher Angriffe. Wirksame Security-Maßnahmen manifestieren sich jedoch weniger im Einbau einer dedizierten „Security-Komponente“ wie einer Firewall, sondern vielmehr in vielen größeren und kleineren Entscheidungen, die die grundlegende Architektur, Anforderungen an die Auswahl von Komponenten, Konfigurationen bereits festgelegter Komponenten oder auch Anforderungen an die operative Nutzung der Komponenten betreffen können.

Heute werden Security-Entscheidungen oft gar nicht bewusst oder erst sehr spät im Engineering-Workflow getroffen. Dann sind jedoch häufig nur noch solche Security-Maßnahmen möglich, die die Automatisierungssysteme komplexer, unpraktischer und teurer machen – etwa der Einbau zusätzlicher Komponenten, die Nutzung zusätzlicher Benutzerkonten und Passwörter oder komplizierte Regeln für Fernzugriffsverbindungen. Damit auch die grundlegenden, praktikableren Security-Maßnahmen geplant werden können (etwa Netzwerkarchitekturmaßnahmen oder die Auswahl von Komponenten, die „by Default“ sichere Authentifizierungsmechanismen mitbringen), muss die Entscheidung dafür bereits früh im Engineering-Workflow fallen. Genau diese Bestrebungen werden gerne unter dem Schlagwort *Security by Design* zusammengefasst.

Security by Design ist also die ausreichend frühe Integration von Security-Entscheidungen in einen existierenden Engineering-Workflow, sodass wichtige Design-Entscheidungen noch beeinflusst werden können.

4 Klassifizierung existierender Methoden für *Security by Design* in der Automatisierungstechnik

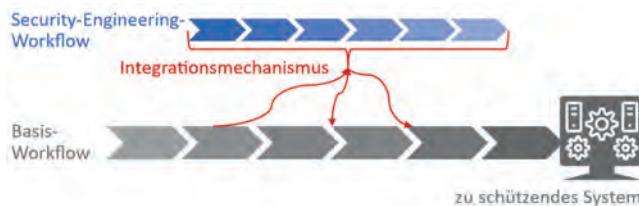


Bild 1: *Security by Design*-Grundbegriffe

Bild 1 zeigt die Grundbegriffe von *Security by Design*, wie es in diesem Beitrag verstanden wird. Ein „*Security by Design*“-Konzept setzt stets einen bereits

existierenden Engineering-Workflow voraus, der dem Design des zu schützenden Systems dient. Dieser existierende Workflow wird im Folgenden als Basis-Workflow bezeichnet. Wenn Security-Engineering in diesen Basis-Workflow integriert werden soll, muss darüber hinaus definiert werden, welcher Security-Engineering-Workflow und welcher Mechanismus für die Integration verwendet werden.

Im Folgenden wird ein Überblick über vorhandene Methoden für *Security by Design* gegeben. Der Scope dieses Beitrags umfasst *Security by Design* für Automatisierungssysteme. Übersetzt in die eingeführten Grundbegriffe heißt das: das zu schützende System ist die automatisierte Anlage, der Basis-Workflow ist der Workflow zum Engineering von Automatisierungssystemen (Automation-Engineering). Dabei werden auch Grenzbereiche wie das Engineering von Cyber-Physical Systems (CPS) und das Engineering von kritischen Infrastrukturen, die

Automatisierungssysteme beinhalten, in diesem Beitrag dem Automation-Engineering zugeordnet. Einen umfassenderen Überblick über „Security by Design“-Konzepte auch aus dem Software-Engineering und Systems-Engineering gibt [1].

Der Security-Engineering-Workflow und der Integrationsmechanismus in den Basis-Workflow hingegen sind in den betrachteten Methoden durchaus verschieden. Reine Security-Engineering-Methoden für Automatisierungssysteme, die keinerlei Bezug zum Basis-Workflow aufweisen, werden nicht betrachtet. Einen Überblick über solche Methoden gibt [2].

Im Folgenden werden existierende „Security by Design“-Methoden nach dem verwendeten Integrationsmechanismus kategorisiert. Eine ausführlichere Beschreibung der einzelnen Methoden ist an dieser Stelle aus Platzgründen nicht möglich, kann jedoch in den referenzierten Literaturquellen nachgelesen werden.

4.1 Integration durch Modifikation des Basis-Workflows („Verschmelzung“)

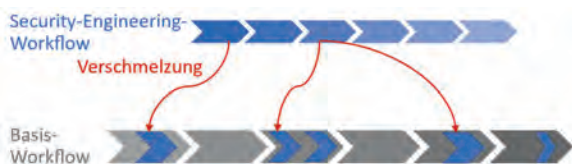


Bild 2: Integrationsmechanismus "Verschmelzung"

Der naheliegendste Integrationsmechanismus ist die Modifikation des Basis-Workflows, sodass er das Security-Engineering enthält – die beiden Workflows „verschmelzen“ also (Bild 2).

Dieser Ansatz wird in mehreren Publikationen verfolgt, setzt eine genaue Definition des Basis-Workflows voraus und muss für jeden Basis-Workflow erneut durchgeführt werden. Das Kernproblem dieses Integrationsmechanismus ist, dass sich die Basis-Workflows je nach Branche, Organisationstyp (z.B. Hersteller oder Betreiber), Region oder Größe des Engineering-Projektes stark unterscheiden [3], was die Integration von Security-Engineering durch Veränderung des Basis-Workflows schwierig verallgemeinerbar macht.

Beispiele für diesen Integrationsmechanismus ist der Development Lifecycle for Cyber-Physical Production Systems (SDL-CPPS) von **Eckhart et al.** [4]. Dort werden vierzehn Security-Aktivitäten identifiziert, die an verschiedenen Stellen in bestimmten Phasen eines definierten Basis-Workflows durchgeführt werden sollen – etwa im Electrical Engineering oder dem Instrumentation & Control (I&C) Engineering. Die Vorgehensweise ähnelt dem im Software-Engineering etablierten Konzept „DevSecOps“ [5], bei dem Security-Aktivitäten in die DevOps-Methode eingeflochten werden.

Die Integration direkt in den Basis-Workflow wurde auch mehrfach für das **Engineering der funktionalen Sicherheit (Safety-Engineering)** vorgeschlagen (z.B. [6], [7]). Das liegt nahe,

weil es dafür mit dem Safety Lifecycle aus der IEC 61508 (und davon abgeleiteten Standards) einen wohldefinierten, in sich abgeschlossenen Basis-Workflow gibt.

4.2 Integration durch Verwendung von Ergebnissen aus dem Basis-Workflow an mehreren Stellen („Kopplung“)

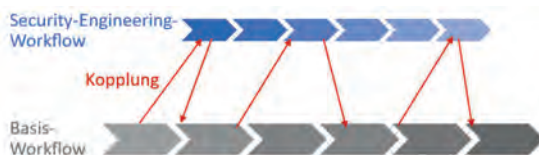


Bild 3: Integrationsmechanismus "Kopplung"

Ein weiterer Integrationsmechanismus ist die Verwendung von Ergebnissen aus dem Basis-Workflow im Security-Engineering-Workflow. Findet die Verwendung an mehreren „An-

knüpfungspunkten" im Basis-Workflow statt und fließen die Ergebnisse aus dem Security-Engineering-Workflow wiederum zurück in den Basis-Workflow, kann das Resultat als Kopplung der beiden Workflows beschrieben werden (Bild 3). Dies birgt gegenüber der Verschmelzung den Vorteil, dass beide Workflows unabhängig voneinander verändert werden können. Auch bei diesem Integrationsmechanismus ist eine starke Abhängigkeit zum Basis-Workflow gegeben.

Der Integrationsmechanismus wird in der Methode von **Tebbe et al.** verwendet, die im Rahmen des Forschungsprojekts INSA [8], [9] und der daraus entstandenen Dissertation [10] entwickelt wurde. Die Security-Analyse wird iterativ an mehreren Stellen des Workflows durchgeführt, an sogenannten „Anknüpfungspunkten“, zu denen jeweils die zu diesem Zeitpunkt vorhandenen Informationen einfließen. Ein weiteres Beispiel aus dem **Safety-Engineering** ist der **ISA-TR84.00.09 Ed3**, der sich gerade in der Entwurfsphase befindet [11]. Dort wird für jede Phase und jeden Prozessschritt des Safety-Lifecycles gemäß IEC 61511 analysiert, welche Ergebnisse security-relevant sind und eine Anleitung für die Verarbeitung dieser „Inputs“ im Security-Engineering und die Erstellung von Security-Outputs gegeben. Damit entspricht der Ansatz dem im **IEC TR 63069:2019** definierten, aber nicht weiter ausgeführten „Co-Engineering“ [12].

4.3 Integration durch Verwendung von Ergebnissen aus dem Basis-Workflow an einer Stelle („Auslösung“)



Bild 4: Integrationsmechanismus "Auslösung"

Wenn es nur einen Anknüpfungspunkt an den Basis-Workflow gibt, der als Startpunkt des Security-Engineering-Workflows dient, kann von einer „Auslösung“ des

Security-Engineering-Workflows durch ein Ergebnis des Basis-Workflows gesprochen werden (Bild 4). Es ist möglich, aber keine Voraussetzung, dass die Ergebnisse nach dem Abschluss des Security-Engineerings in den Basis-Workflow zurückgespielt werden.

Dieser Integrationsmechanismus ist unabhängiger vom Basis-Workflow – nur die Information, die den Security-Prozess auslöst, muss auf jeden Fall vorkommen. Es ist damit allerdings schwieriger, ein rechtzeitiges Einwirken auf security-relevante Entscheidungen im Basis-Workflow zu erreichen, weil die Ergebnisse des Security-Engineerings gar nicht oder nur an einer – meist nicht genau spezifizierten – Stelle in den Basis-Workflow zurückgespielt werden. Ein Beispiel für den Integrationsmechanismus „Auslösung“ ist die Methode von **Eckhart et al.** [13], um Security-Risiken, dargestellt als Angriffsgraphen, automatisiert zu identifizieren, nachdem Engineering-Daten aus dem Automation-Engineering als AutomationML (AML [14])-Artefakte eingelesen und mit Security-Informationen teils automatisiert, teils manuell durch Ingenieure angereichert wurden. Ein weiteres Beispiel ist der von **Dedousis et al.** beschriebene Ansatz für das Engineering kritischer Infrastrukturen [15]. Ausgangspunkt ist die Modellierung einer Anlage als Material Flow Network Graph. Dieser Graph wird dann automatisiert auf Abhängigkeiten und Security-Risiken hin analysiert. **Bochman und Freeman** beschreiben in [16] das „Consequence-driven, Cyber-Informed Engineering“ (CCE). Ein Kernelement der Methode ist die Verwendung „nicht-digitaler“ Maßnahmen, die üblicherweise im Rahmen des Automation-Engineering-Workflows festgelegt werden. Daher ist der CCE-Workflow auf Ergebnisse und Expertise aus dem Automation-Engineering angewiesen. **Marszal und McGlone** erklären eine Methode mit Namen „Security PHA Review“ in [17]. Die Idee ist, Ergebnisse einer im Rahmen des Safety-Engineerings durchgeführten Process Hazard Analysis (PHA) als Einstiegspunkt für eine Security-Analyse zu verwenden.

4.4 Integrierbarkeit durch Verwendung von Konzepten des Basis-Workflows im Security-Engineering-Workflow („Harmonisierung“)

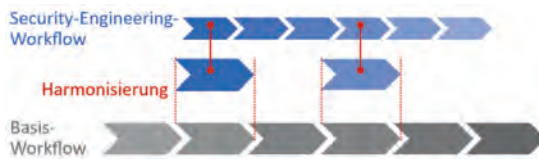


Bild 5: Integrationsmechanismus "Harmonisierung"

Bei diesem Integrationsmechanismus wird Security gar nicht explizit in den Basis-Workflow integriert. Stattdessen wird der Security-Engineering-Workflow so angepasst,

dass er dem Basis-Workflow möglichst ähnlich ist; beide Workflows werden also „harmonisiert“ (Bild 5). Damit wird eine einfache Integrierbarkeit dieses Security-Engineering-Workflows angenommen, ohne dass dafür explizit ein Prozessmodell oder eine Methode definiert wird. Der Integrationsmechanismus ist wiederum nur bei methodischer Festlegung auf einen Basis-Workflow geeignet; zusätzlich engt er auch den methodischen Rahmen für den Security-Engineering-Workflow stark ein.

Ein Beispiel für die Integrierbarkeit durch Harmonisierung stellen **Mažeika und Butleris** vor [18]. Die Integrierbarkeit soll durch die Verwendung von Konzepten und Methoden aus dem Model-Based Systems-Engineering (MBSE) erfolgen, insbesondere durch ein UML-basiertes Security-Profil für cyber-physische Systeme (CPS). Einen ähnlichen Ansatz verfolgen **Apvrille und Roudier** [19] mit dem SysML-Profil SysML-Sec für eingebettete Systeme innerhalb des Model-Driven Systems-Engineering und **Eckhart et al.** [13] mit der AutomationML-Bibliothek AMLsec. Ähnlich gehen eine Reihe weiterer Autoren vor, die Security in **Systems-Engineering** ([20]–[22]), **Requirements Engineering** [23] oder **Software Engineering** [24]–[27]) als Basis-Workflow integrieren. Diese Ansätze haben jedoch keinen Fokus auf Automatisierungssysteme. Ein Sonderfall ist die **IEC 62443-4-1:2018** [28]. Hier werden in Anlehnung an Microsofts Secure Development Lifecycle [29] und andere etablierte Konzepte aus dem Software-Engineering auf sehr hoher Abstraktionsebene Anforderungen an einen sicheren Produktentwicklungsprozess beschrieben.

5 Anforderungen an die Integration von Security in das Automation-Engineering

Die Analyse existierender Ansätze ergibt die folgenden Anforderungen an die Beschreibung von *Security by Design* für industrielles Automation-Engineering:

- **A1: Unabhängigkeit vom Basis-Workflow.** Die Analyse der existierenden Ansätze zeigt, dass der Integrationsmechanismus von Security-Engineering in den Basis-Workflow möglichst losgelöst vom Basis-Workflow beschrieben werden muss.

Begründung: Eine Methode für *Security by Design*, die auf die Konformität mit einem

fest definierten Engineering-Workflow ausgerichtet ist, bliebe auf einen zu kleinen praktischen Anwendungsraum beschränkt.

- **A2: Durchführbarkeit für die Gewerke des Basis-Workflows:** Das „*Security by Design*“-Konzept sollte ermöglichen, dass jedes Gewerk das Security-Engineering für seinen Einflussbereich durchführt.

Begründung: Security-Engineering setzt ein hohes Maß an Verständnis der Systeme des Basis-Workflows voraus. Da aber bereits das Automation-Engineering mit seinen vielen beteiligten Gewerken interdisziplinär ist, ist es unwahrscheinlich, dass eine Person ausreichendes Systemwissen über all diese Gewerke hinweg hat.

- **A3: Sicherstellung von rechtzeitigen Entscheidungen.** Security-Entscheidungen müssen früh genug getroffen werden, um wichtige Designentscheidungen noch beeinflussen zu können. Dafür ist allerdings die Definition eines genauen Zeitpunkts weniger wichtig als die Definition eines spätestmöglichen Zeitpunkts für eine Security-Entscheidung.

Begründung: Auch wenn der chronologische Ablauf nicht die zentrale Rolle für die Beschreibung des „*Security by Design*“-Ansatzes spielt, so ist die rechtzeitige Entscheidungsfindung doch ein zentrales Ziel von *Security by Design*.

- **A4: Nutzung systemimmanenter Eigenschaften für Security.** Das „*Security by Design*“-Konzept soll es ermöglichen, die in den sowieso verbauten Systemkomponenten immanenten Eigenschaften für Security zu nutzen und die zusätzlichen Lösungen nur nach Ausschöpfung dieser Möglichkeiten in Erwägung ziehen.

Begründung: Viele der bisherigen Ansätze versuchen, zusätzliche, dedizierte Security-Lösungen in Automatisierungssystemen unterzubringen. Solche zusätzlichen Lösungen sind jedoch tendenziell aufwendiger und bergen die Gefahr, ihrerseits die Angriffsfläche zu erhöhen. (Diese Forderung findet sich auch in den Anforderungen an *Security by Design* der VDI/VDE 2182-4 [30].)

- **A5: Methodische Anleitung für das Security-Engineering:** Es soll nicht nur definiert werden, was wann im Basis-Workflow getan oder entschieden werden soll, sondern auch methodische Hilfestellung für das „Wie“ gegeben werden.

Begründung: Diese Anforderung speist sich aus einer weiteren Beobachtung bei der Analyse der existierenden „*Security by Design*“-Methoden: Die methodische Anleitung fehlt oft. Gerade weil Automatisierungsingenieure in der Regel wenig Security-Know-how haben, wäre diese aber wichtig.

6 Konzept "Automation Security by Design Decisions"

Die Autoren schlagen ein „Security by Design“-Konzept vor, das auf Security-Entscheidungen (Security Decisions) als zentralem Element aufbaut – es präzisiert also den Begriff *Security by Design* zu *Security by Design Decisions*. Entscheidungen haben im Gegensatz zu Phasen oder Aktivitäten zwei Vorteile: Erstens erfordern sie keine Festlegung eines chronologischen Ablaufs, zweitens stellen sie das Ergebnis einer Security-Aktivität – eben das Treffen der Entscheidung – in den Vordergrund.

Im Rahmen des Automation-Engineering-Workflows müssen eine Reihe von Security-Entscheidungen getroffen werden. Nicht immer sind das zusätzliche Entscheidungen. Manchmal reicht es, Security-Entscheidungen explizit zu machen, die andernfalls durch funktionale Design-Festlegungen unbewusst getroffen werden.

6.1 Security-Entscheidungspunkte (Decision Points)

Um das Treffen von Security-Entscheidungen zu systematisieren, werden Security-Entscheidungspunkte definiert. Ein Security-Entscheidungspunkt bündelt eine in sich abgeschlossene Menge von Security-Parametern und betrifft eine klar abgrenzbare Menge operativer Funktionen.

Hinsichtlich der zuvor beschriebenen Integrationsmechanismen ordnet sich das Konzept in den Mechanismus „Kopplung“ ein, jedoch mit einem wichtigen Unterschied: Statt einer vollendeten Kopplung an einen feststehenden Basis-Workflow werden die Entscheidungspunkt als „Koppelpunkte“ beschrieben, die eine möglichst flexible Kopplung an verschiedene Basis-Workflows ermöglichen. Bild 6 verdeutlicht das Prinzip: Jeder Security-Entscheidungspunkt kann als einer der dreieckigen Koppelpunkte verstanden werden. Dies ermöglicht die Unabhängigkeit vom Basis-Workflow (**Anforderung A1**).

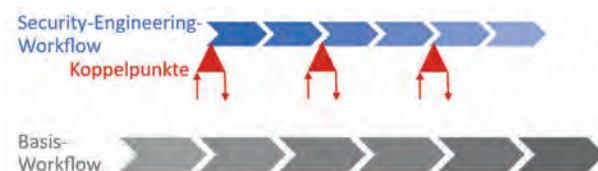


Bild 6: Integrationsmechanismus "Koppelpunkte"

Ein **Security-Entscheidungspunkt** hat die folgenden Inhalte:

- **Titel und Entscheidung:** Die Beschreibung der zu treffenden Entscheidung.

- **Betroffene operative Funktion(en):** Die operativen Funktionen fassen die funktionalen Anforderungen aus dem Basis-Workflow zusammen, die von der Security-Entscheidung (potenziell) betroffen sind. Sie geben damit den security-relevanten Informationen einen operativen Kontext.
(Beispiel: Das Forcen oder Brücken von Steuerungsausgängen.)
- **Security-relevante Informationen:** Informationen aus dem Basis-Workflow, die für die Security-Entscheidung relevant sind. Sie werden weiterhin danach unterteilt, wie sie im Rahmen der Security-Entscheidung verwertet werden können:
 - o **Indikatoren:** Informationen aus dem Basis-Workflow, die durch die Security-Entscheidung zwar nicht veränderbar sind, aber die Entscheidungsfindung beeinflussen. Meist geben sie die Hinweise auf den Schutzbedarf von Funktionen oder Komponenten oder mögliche Gefährdungsszenarien (Beispiel: die sicheren Betriebszustände eines Aktors oder einer Maschine).
 - o **Parameter:** Informationen aus dem Basis-Workflow, die potenziell durch die Security-Entscheidung verändert werden können (dies bedeutet meist ein „Tuning“ einer oder mehrerer operativer Funktionen). Ob ein Parameter tatsächlich verändert werden darf, wird im Rahmen der Entscheidungsfindung festgelegt.
(Beispiel: Die Wahl eines Kommunikationsprotokolls)

Security-Entscheidungspunkte sind zeitlich flexibel im Basis-Workflow verteilbar. Die enthaltenen Entscheidungen können weitgehend isoliert voneinander und auch parallel getroffen werden. Damit ist es auch möglich, dass unterschiedliche Gewerke sich mit unterschiedlichen Entscheidungspunkten befassen (**Anforderung A2**). Die gelisteten operativen Funktionen eines Entscheidungspunktes helfen, die für die Entscheidungspunkte relevanten Personen zu identifizieren. Sollten die Entscheidungspunkte im Falle des betrachteten Basis-Workflows doch mehrere Gewerke betreffen, können sie durch Auswahl nur einer Teilmenge der operativen Funktionen weiter unterteilt werden.

Darüber hinaus dienen die operativen Funktionen als Hilfestellung für die Verankerung der Entscheidungspunkte im Basis-Workflow: Spätestens, wenn die vom Entscheidungspunkt betroffenen operativen Funktionen entworfen werden, muss auch die Security-Entscheidung getroffen werden – dies stellt sicher, dass Security-Entscheidungen rechtzeitig getroffen werden (**Anforderung A3**).

Die Unterteilung der security-relevanten Informationen aus dem Basis-Workflow in Indikatoren und Parameter ergänzt die Arbeit von Tebbe [10] um die Hilfestellung, wie diese Informationen

im Rahmen des Security-Engineering-Workflows zu nutzen sind (**Anforderung A5**). Die Identifikation sowohl der Security-Parameter als auch der operativen Funktionen ermöglicht dabei die Nutzung systemimmanenter Eigenschaften für die Security (**Anforderung A4**) - siehe Abschnitt 6.2 für die methodische Anleitung zum Treffen der Security-Entscheidungen auf Basis der Entscheidungspunkte.

Die Autoren haben die vorliegenden Security-Entscheidungspunkte in Engineering-Workflows von Herstellern und Betreibern von Automatisierungslösungen identifiziert. Es wurden dabei sowohl in der Literatur ([31]–[35]) beschriebene als auch in realen Unternehmen (INEOS als Betreiber und HIMA als Hersteller) gelebte Engineering-Workflows zugrunde gelegt.

Auch sind Security-Entscheidungen, die in existierenden „*Security by Design*“-Methoden ([4], [10], [11], [16], [21]) implizit oder explizit definiert werden, bei der Definition der Entscheidungspunkte berücksichtigt worden. Tabelle 2 gibt einen Überblick über die vorgeschlagenen Entscheidungspunkte.

Bild 3 zeigt ihre beispielhafte Verankerung in einem kombinierten Engineering-Workflow eines Betreibers und eines Steuerungsherstellers. Zur Orientierung sind Phasen in Anlehnung an das NAMUR-Arbeitsblatt NA35 [31] mit eingezeichnet.

Tabelle 2: Überblick über Security-Entscheidungspunkte (Decision Points, DP)

ID	Titel	Entscheidungen
DP1	Prozessnahe Komponenten	Welche Security-Fähigkeiten sollen die Feldgeräte und Steuerungen haben? Wie soll Security bei ihrer Nutzung berücksichtigt werden?
DP2	Leitsystemkomponenten	Welche Security-Fähigkeiten sollen die Komponenten zum Bedienen, Beobachten und Engineeren haben? Wie soll Security bei ihrer Nutzung berücksichtigt werden?
DP3	IT-Dienste	Welche Komponenten / Dienste soll es außer den direkt für die Automatisierung notwendigen Komponenten geben und welche Security-Fähigkeiten sollen sie haben?
DP4	Zugriffsarten	Wie soll der Zugriff auf / die Kommunikation mit Komponenten erfolgen?
DP5	Beschaffung	Welche Security-Anforderungen sollen in welcher Form als Beschaffungskriterien definiert werden?
DP6	Anzeigen und Alarme	Wie sollen Operator über Security informiert werden und darauf reagieren können?
DP7	Logik	Wie soll die Steuerungs- und Leitsystemlogik zu den Security-Eigenschaften der Komponenten beitragen?
DP8	Prüfungen	Wie soll Security in Abnahmetests oder später im Betrieb geprüft werden?
DP9	Betrieb	Welche (regelmäßigen) Betriebsprozesse sollen die Aufrechterhaltung der Security gewährleisten?

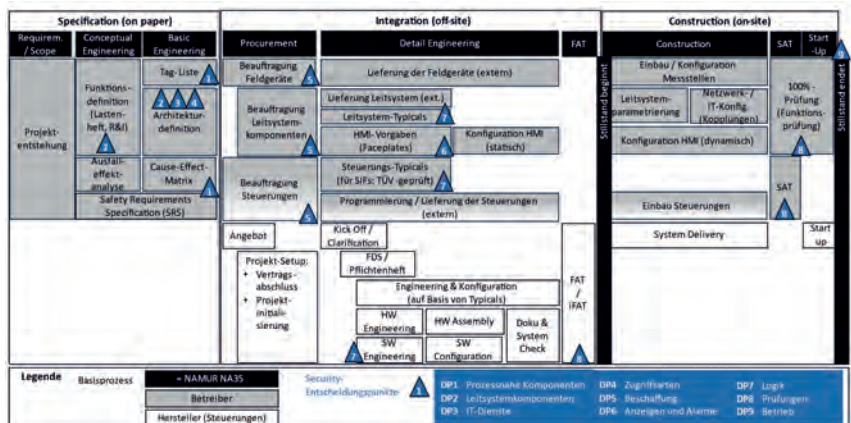


Bild 7: Beispielhafte Integration der Entscheidungsprozesse in den kombinierten Engineering-Workflow eines Betreibers und eines Steuerungsherstellers

6.2 Systematische Security-Entscheidungsfindung mit Entscheidungspunkten

Wenn Automatisierungsingenieure wissen, welche Security-Entscheidungen sie treffen müssen, bedeutet das noch nicht, dass sie diese auch treffen können. Um sie dazu zu befähigen, wird eine Systematik für die Security-Entscheidungsfindung auf Basis der Entscheidungspunkte beschrieben (**Anforderung A5**).

Beim Treffen der Entscheidung wird der Lösungsraum Schritt für Schritt verkleinert. Die erste Einschränkung ist bereits mit der Auswahl eines Security-Entscheidungspunktes, also einem Entscheidungszweck, geschehen: Damit werden nur noch ein Teil der security-relevanten Informationen und -Funktionen betrachtet.

Das weitere Vorgehen innerhalb eines Entscheidungspunktes zeigt Bild 4. Zunächst werden die im Entscheidungspunkt gelisteten operativen Funktionen daraufhin analysiert, ob sie für das individuelle Engineering-Projekt notwendig sind, um die funktionalen Anforderungen zu erfüllen. Es ist auch möglich, operative Funktionen hinzuzufügen; die gegebenen Funktionen des Entscheidungspunktes dienen lediglich der Arbeitserleichterung und des besseren Verständnisses des Entscheidungspunktes. Nicht benötigte operative Funktionen werden direkt entfernt. Auch diese Eliminierung ist bereits eine Security-Anforderung; denn die Reduzierung von Funktionalität bedeutet auch eine Reduzierung der Angriffsfläche (Härtung). Gleichzeitig ist es die grundlegendste Art, systemimmanente Eigenschaften für Security zu nutzen (**Anforderung A4**).

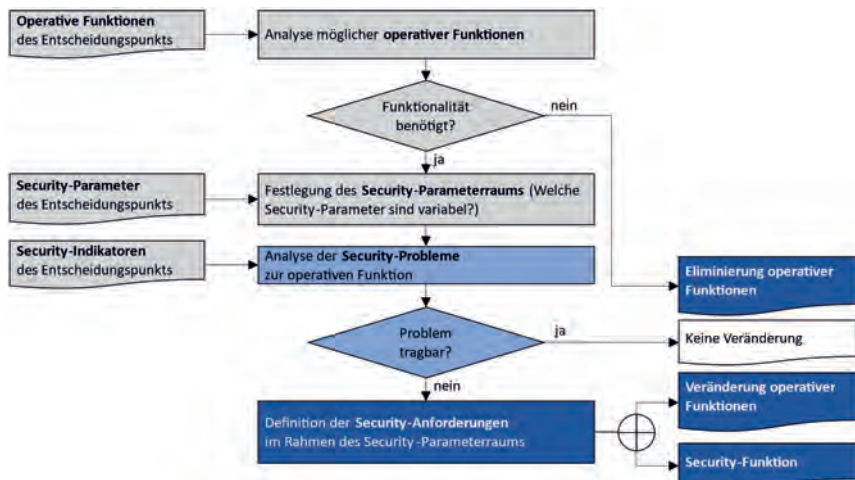


Bild 8: Systematische Security-Entscheidungsfindung mit Hilfe eines Entscheidungspunktes

Im nächsten Schritt wird basierend auf dem Wissen über notwendige operative Funktionen der Security-Parameterraum bestimmt. Dabei wird für das individuelle Engineering-Projekt festgelegt, welche Parameter des Entscheidungspunktes fix oder variabel sind, also ob eine funktionale Anforderung die Fixierung des Parameters erfordert oder nicht. Die Parameter beeinflussen die Art der Ausführung einer oder mehrerer operativer Funktionen.

Die Festlegung der benötigten operativen Funktionen und des Security-Parameterraums beschränkt einerseits den Security-Lösungsraum so, dass die funktionalen Anforderungen sichergestellt bleiben; andererseits ermöglicht es die Nutzung der verbleibenden systemimmanenten Parameter für die Security-Entscheidungen (**Anforderung A4**).

Im dritten Schritt werden für die verbliebenen operativen Funktionen damit verbundene Security-Probleme identifiziert. Probleme können aus einer Risikobetrachtung resultieren; ein Security-Problem kann aber auch die fehlende Compliance mit internen oder externen regulatorischen Security-Anforderungen sein. Für tragbare Probleme ist keine weitere Aktion erforderlich.

Ist ein Security-Problem nicht tragbar, müssen innerhalb des zuvor festgelegten Parameter-raums Security-Anforderungen definiert werden. Neben der Eliminierung operativer Funktion ist auch die Veränderung („Security-Anreicherung“) operativer Funktionen möglich. Wenn diese beiden Möglichkeiten ausgeschöpft sind, ist die Nutzung systemimmanenter Eigenschaften für Security nicht möglich und eine dedizierte Security-Funktion muss definiert werden.

7 Diskussion und Ausblick: Die Entscheidungsbasis

Mit dem Konzept der Entscheidungspunkte und der Anleitung zum Treffen der Entscheidungen ist der methodische Grundstein gelegt, um *Security by Design* in den Automation-Engineering-Workflow zu integrieren. Eine Kernanforderung des Konzepts ist die Befähigung der am Automation-Engineering beteiligten Gewerke, „ihre“ Security-Entscheidungen selbst zu treffen. Die methodische Befähigung allein löst aber noch nicht das Problem, dass *Security by Design* für jedes dieser Gewerke ein erheblicher Zusatzaufwand ist. Diesen Mehraufwand beabsichtigen die Autoren zu verringern, in dem zu jedem der Entscheidungspunkte eine Entscheidungsbasis mitgeliefert wird, die voraussichtlich folgende Komponenten umfasst:

- Menschenlesbare Security-Diagramme (als Artefakte einer visuellen Modellierungssprache), die die operativen Funktionen und die Vielzahl an Security-Informationen für Menschen schnell erfassbar visualisieren.
- Ein während des Security-Engineerings befülltes elektronisches Datenmodell, das die effiziente Pflege, Ergänzung, Erzeugung sowie das Gewerke-übergreifende Austauschen dieser Security-Diagramme ermöglicht. Damit wird gleichzeitig notwendige Dokumentation für Zertifizierungen sowie eine maschinenlesbare Schnittstelle für die

Implementierung der Security-Entscheidungen im weiteren Verlauf des Engineerings, aber auch das schnelle Auffinden von Informationen in der Betriebsphase ermöglicht.

- Lösungsmuster, die bewährte Implementierungen für die Security-Anforderungen zusammenfassen – idealerweise in für Automatisierungssingenieure vertrauten Formaten wie Typicals oder Faceplates.
- Security-Bibliotheken für die operativen Funktionen, Security-Informationen, Security-Probleme, Security-Anforderungen (in Form von security-angereicherten operativen Funktionen oder Security-Funktionen) sowie Lösungsmuster. Diese Bibliotheken sollen das Security-Wissen für die einzelnen Entscheidungspunkte kapseln und wiederverwendbar machen.

Überdies steht die Validierung des Konzeptes sowie der definierten Entscheidungspunkte noch aus. Da die Entscheidungspunkte keinen Anspruch auf Vollständigkeit erheben, ist dabei eine Ergänzung der Entscheidungspunkte mitsamt ihren operativen Funktionen und Security-Informationen zu erwarten.

8 Danksagung

Diese Arbeit wird unterstützt vom Bundesministerium für Bildung und Forschung (BMBF) im Rahmen des Projektes „IDEAS“ (Integrated Data Models for the Engineering of Automation Security, Vorhaben 16KIS1269K) [36]. Anwendungspartner des Projektes sind INEOS und HIMA. Die oben umrissene Security-Entscheidungsbasis wird ein zukünftiges Arbeitsergebnis des Projektes. Das Datenmodell wird in Zusammenarbeit mit dem NAMUR AK 1.3 (Informationsmodelle) entwickelt und dient gleichzeitig als das von der Industrial Digital Twin Association (IDTA) anerkannte Teilmodell für Security-Engineering in der Verwaltungsschale.

9 Literatur

- [1] S. Fluchs, R. Drath, und A. Fay, „How to Prepare Security by Design Decisions for Industrial Control Systems: Analysis of Relevant Concepts from Automation Engineering, Software Engineering, and Systems Engineering“, gehalten auf der Entwurf komplexer Automatisierungssysteme (EKA), Magdeburg, Germany, to be published 2022.
- [2] S. Fluchs und H. Rudolph, „Wie OT-Security-Engineering eine Ingenieurwissenschaft wird - Ein Denkmodell und ein Datenmodell [Making OT Security Engineering an Engineering Discipline - A Thought Model and a Data Model]“, *atp*, Nr. 08–2019, S. 74–86, Aug. 2019, [Online]. Verfügbar unter: <https://www.atpinfo.de/aktuell/aus-der-praxis/20-11-2019-hauptbeitrag-ot-security-engineering/>

- [3] P. Ghobadi-Bigvand, „An adaptive, context-sensitive, workflow support system for process and automation engineering of production plants“, Helmut Schmidt Universität / Universität der Bundeswehr Hamburg, Hamburg, 2018. [Online]. Verfügbar unter: https://edoc.sub.uni-hamburg.de/hsu/volltexte/2018/3218/pdf/PhD_Dissertation_Pou-ria_Bigvand.pdf
- [4] M. Eckhart, A. Ekelhart, A. Luder, S. Biffl, und E. Weippl, „Security Development Lifecycle for Cyber-Physical Production Systems“, in *IECON 2019 - 45th Annual Conference of the IEEE Industrial Electronics Society*, Lisbon, Portugal, Okt. 2019, S. 3004–3011. doi: 10.1109/IECON.2019.8927590.
- [5] US Department of Defense, „DoD Enterprise DevSecOps Fundamentals, Version 2.0“. Mai 2021. [Online]. Verfügbar unter: <https://dl.dod.cyber.mil/wp-content/uploads/devsecops/pdf/DoDEnterpriseDevSecOpsFundamentals.pdf>
- [6] R. Bramberger, H. Martin, B. Gallina, und C. Schmittner, „Co-engineering of Safety and Security Life Cycles for Engineering of Automotive Systems“, *Ada Lett.*, Bd. 39, Nr. 2, S. 41–48, 2019, doi: 10.1145/3394514.3394519.
- [7] C. Schmittner, Z. Ma, und E. Schoitsch, „Combined safety and security development lifecycle“, in *2015 IEEE 13th International Conference on Industrial Informatics (INDIN)*, Cambridge, United Kingdom, Juli 2015, S. 1408–1415. doi: 10.1109/INDIN.2015.7281940.
- [8] C. Tebbe, M. Glawe, A. Scholz, K.-H. Niemann, A. Fay, und J. Dittgen, „Wissensbasierte Sicherheitsanalyse in der Automation. Aufwandsreduzierung am Beispiel der IT-Sicherheit“, *atp edition*, Nr. 4/2015, 2015.
- [9] C. Tebbe, K.-H. Niemann, und A. J. Fay, „Ontology and life cycle of knowledge for ICS security assessments“, gehalten auf der 4th International Symposium for ICS & SCADA Cyber Security Research 2016, Okt. 2016. doi: 10.14236/ewic/ICS2016.5.
- [10] C. Tebbe, „Durchgängiges Wissensmanagement von OT-Security-Wissen im Lebensweg von Produktionsanlagen“, Helmut Schmidt Universität / Universität der Bundeswehr Hamburg, Hamburg, 2021.
- [11] ISA-TR84.00.09 Ed.3, „Cyber Security Related to the Safety Lifecycle (Draft 2022)“.
- [12] IEC TR 63069:2019, „Industrial-process measurement, control and automation - Framework for functional safety and security“. 2019.
- [13] M. Eckhart, A. Ekelhart, und E. R. Weippl, „Automated Security Risk Identification Using AutomationML-based Engineering Data“, *IEEE Trans. Dependable and Secure Comput.*, S. 1–1, 2020, doi: 10.1109/TDSC.2020.3033150.

- [14] IEC 62714 series, „Engineering data exchange format for use in industrial automation systems engineering - Automation Markup Language“.
- [15] P. Dedousis, G. Stergiopoulos, G. Arampatzis, und D. Gritzalis, „Towards Integrating Security in Industrial Engineering Design Practices“; in *Proceedings of the 18th International Conference on Security and Cryptography*, Online Streaming, --- Select a Country ---, 2021, S. 161–172. doi: 10.5220/0010544001610172.
- [16] A. A. Bochman und S. G. Freeman, *Countering Cyber Sabotage. Introducing Consequence-Driven, Cyber-Informed Engineering (CCE)*, First edition Published 2021. CRC Press, Taylor & Francis Group, 2021.
- [17] E. Marszal und J. McGlone, *Security PHA Review for Consequence-Based Cybersecurity*. ISA, 2019.
- [18] D. Mažeika und R. Butleris, „Integrating Security Requirements Engineering into MBSE: Profile and Guidelines“, *Security and Communication Networks*, Bd. 2020, S. 1–12, März 2020, doi: 10.1155/2020/5137625.
- [19] L. Aprville und Y. Roudier, „SysML-Sec: A SysML environment for the design and development of secure embedded systems“, *APCOSEC, Asia-Pacific Council on Systems Engineering*, S. 8–11, 2013.
- [20] J. Fran. Ruiz, A. Maña, und C. Rudolph, „An Integrated Security and Systems Engineering Process and Modelling Framework“, *The Computer Journal*, Bd. 58, Nr. 10, S. 2328–2350, Okt. 2015, doi: 10.1093/comjnl/bxu152.
- [21] D. D. Walden, G. J. Roedler, K. Forsberg, R. D. Hamelin, und T. M. Shortell, *INCOSE Systems engineering handbook: a guide for system life cycle processes and activities*, 4th edition. Hoboken, New Jersey: Wiley, 2015.
- [22] R. Ross, M. McEvilley, und J. C. Oren, „NIST SP 800-160, Volume 1: Systems security engineering: considerations for a multidisciplinary approach in the engineering of trustworthy secure systems“, National Institute of Standards and Technology, Gaithersburg, MD, NIST SP 800-160v1, Nov. 2016. doi: 10.6028/NIST.SP.800-160v1.
- [23] S. Rehman und V. Gruhn, „An Effective Security Requirements Engineering Framework for Cyber-Physical Systems“, *Technologies*, Bd. 6, Nr. 3, S. 65, Juli 2018, doi: 10.3390/technologies6030065.
- [24] J. L. Vivas, J. A. Montenegro, und J. López, „Towards a Business Process-Driven Framework for Security Engineering with the UML“, in *Information Security*, Bd. 2851, C. Boyd und W. Mao, Hrsg. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, S. 381–395. doi: 10.1007/10958513_29.

- [25] U. Evertsson, U. Örtberg, und L. Yngström, „Integrating Security Into Systems Development“, in *Security and Privacy in the Age of Uncertainty*, D. Gritzalis, S. Capitani di Vimercati, P. Samarati, und S. Katsikas, Hrsg. Boston, MA: Springer US, 2003, S. 157–168. doi: 10.1007/978-0-387-35691-4_14.
- [26] M. Schumacher, E. Fernandez-Buglioni, D. Hybertson, F. Buschmann, und P. Sommerlad, *Security patterns: integrating security and systems engineering*. Chichester, England ; Hoboken, NJ: John Wiley & Sons, 2006.
- [27] D. Hatebur, M. Heisel, und H. Schmidt, „A Security Engineering Process based on Patterns“, in *18th International Conference on Database and Expert Systems Applications (DEXA 2007)*, Regensburg, Germany, Sep. 2007, S. 734–738. doi: 10.1109/DEXA.2007.36.
- [28] IEC 62443-4-1:2018, „Security for industrial automation and control systems - Part 4-1: Secure product development lifecycle requirements“. 2018.
- [29] S. Lipner und M. Howard, *The Security Development Lifecycle. SDL: A Process for Developing Demonstrably More Secure Software*. 2006.
- [30] VDI/VDE 2182-4, „IT security for industrial automation - Recommendations for the implementation of security properties for components, systems, and equipment [Informationssicherheit in der industriellen Automatisierung - Empfehlungen zur Umsetzung von Security-Eigenschaften für Komponenten, Systeme und Anlagen]“. 2020.
- [31] NAMUR NA 35, „Engineering and execution of PCT projects in process industry [PLT-Planung und -Abwicklung in der Prozessindustrie]“. 2019.
- [32] M. Hollender, *Collaborative process automation systems*. Research Triangle Park, NC: ISA, 2010.
- [33] G. Rathwell, „PERA Control and Information Systems Lead Engineer's Guide - Industry Sector: Oil & Gas“. 2018.
- [34] ISA-TR106.00.02-2017, „Procedure Automation for Continuous Process Operations - Work Processes“. 2017.
- [35] VDI/VDE 3695-1, „Engineering of industrial plants - Evaluation and optimisation of the engineering - Fundamentals and procedure [Engineering von Anlagen - Evaluieren und Optimieren des Engineerings - Grundlagen und Vorgehensweise] Draft 2020“. 2020.
- [36] German Federal Ministry of Education and Research (BMBF), „IDEAS - Integrated Data Models for the Engineering of Automation Security. Project overview (in German)“. 2021. [Online]. Verfügbar unter: <https://www.forschung-it-sicherheit-kommunikationssysteme.de/projekte/ideas>

AutomationML-basierte Modellierungsansätze für ein Security-Engineering-Informationsmodell

AutomationML-based modeling approaches for a security engineering information model

Emre Tastan, Prof. Dr.-Ing. **Rainer Drath**,
Hochschule Pforzheim;
Sarah Fluchs, admeritia GmbH, Langenfeld,
Helmut-Schmidt-Universität/Universität der Bundeswehr, Hamburg

Kurzfassung

Bei der industriellen Digitalisierung und der Einführung von Internettechnologien in die Produktion stellt die Security eine der größten Herausforderungen dar. Mit zunehmender Vernetzung ihrer Komponenten in der Industrie 4.0-Umgebung sind diese zu attraktiven Angriffszielen für IT-Angriffe und Malware geworden. Während die funktionale Sicherheit tief in die Entwicklung von Produkten oder Prozessen integriert ist, ist dies bei der Security nicht der Fall. Security-Engineering muss sich analog zur funktionalen Sicherheit in den bestehenden und sich gerade stark verändernden Automatisierungs-Engineering-Prozess eingliedern, vor allem muss es aber für Automatisierungingenieure effizient durchführbar sein. Dieser Beitrag begründet den Bedarf an einem Security-Engineering-Modell, berichtet über die laufenden Arbeiten zu den Modellierungsansätzen mit AutomationML und stellt diese in Vergleich.

Abstract

Industrial digitization and the introduction of Internet technologies pose a major challenge for security. With increasing networking of their components in the Industry 4.0 environment, these have become attractive targets for IT attacks and malware. While functional safety is deeply integrated into the development of products or processes, this is not the case with security. Security engineering, like functional safety, must integrate into existing automation engineering processes, which is currently undergoing significant changes, but most importantly, it must be efficient for automation engineers to perform. This paper justifies the need for a security engineering model, reports on the ongoing work on modeling approaches with AutomationML and compares them.

1. Motivation und Zielstellung

Die vierte industrielle Revolution (Industrie 4.0) verfolgt unter anderem die durchgängige Digitalisierung und intelligente Vernetzung von Maschinen, Anlagen, Prozessen sowie von Menschen. Systeme der Prozess- und Fertigungsautomation werden durch Internettechnologien miteinander verbunden. Zeitgleich werden diese Anlagen jedoch auch zu attraktiven Zielen für Cyberangriffe und Malware. Darüber hinaus sind bestehende industrielle Anlagen und Maschinen auf Lebenszyklen von 20 Jahren und länger konzipiert, aber verfügen oft kaum über Cybersecurity-Funktionalitäten [1]. Bisher wurden deshalb Security-Maßnahmen üblicherweise nachträglich umgesetzt. Diese nachträgliche Umsetzung führt dazu, dass Security bis heute ein "Fremdkörper" in der Domäne des Automation-Engineering und auch in den zugehörigen Automatisierungs-Gremien ist. Erschwerend kommt hinzu, dass die Schutzmaßnahmen der Cybersecurity häufig die Verfügbarkeit und Performance der Produktionseinheiten verschlechtern. All diese Punkte sprechen dafür, Security-Engineering zu systematisieren. Noch immer erfolgt Security-Engineering zu unsystematisch, oft getrieben von Best-Practice-Lösungen. Wie gut eine Lösung ist, hängt vor allem von dem Wissensumfang der beteiligten Security-Experten ab. Damit ein systematischeres Security-Engineering für Automatisierungsingenieure bewältigbar bleibt, sind digitale Modelle notwendig. Modelle, die es ermöglichen, eine Vogelperspektive über das gesamte technische System zu schaffen. Nicht nur lesbar für den Menschen, um technische Systeme mit der Security-Perspektive zu verbessern, sondern auch maschinenverarbeitbar, um unter anderem security-relevante Merkmale zu visualisieren. Somit muss in Zukunft Security-Engineering ein selbstverständlicher und organischer Bestandteil des Automatisierungs-Engineering-Prozesses werden. Dafür werden neue Methoden benötigt, welche parallel mit dieser Arbeit („Security-Entscheidungen „by Design“ in das Engineering prozesstechnischer Anlagen integrieren“) bei der Automation 2022 vorgestellt wird.

Im Kontext von Industrie 4.0 und sich schneller ändernden und zunehmend modular aufgebauten Automatisierungslösungen gewinnt die Maschine-zu-Maschine-Kommunikation an Bedeutung. Diese soll sicherstellen, dass Maschinen einander ihre Zustände und Eigenschaften eindeutig und universell verständlich mitteilen können. Wie solch ein Modell für ein Automation-Security-Engineering-Prozess aussehen kann und wie ein Security-Domänenmodell im Engineering sowie auch in der späteren Betriebsphase helfen kann, untersuchen die Autoren im Rahmen des öffentlich geförderten Projektes IDEAS "Integrated Data Models for the engineering of Automation Security". Dazu wird unter anderem ein elektronisches Security-Informationsmodell, das in elektronischer und standardisierter Form den Austausch zwischen beteiligten Planungswerkzeugen maschinenlesbar und

maschinenverarbeitbar ermöglicht, entwickelt. Es fungiert dabei als modelltechnische Grundlage des elektronischen Workflows und ermöglicht weitreichende Funktionen wie Änderungsmanagement, Tracking von Entscheidungen über Phasen und Zeiten hinweg, sowie Möglichkeiten zur Archivierung und Speicherung von Musterlösungen und als Basis für neue Geschäftsmodelle der Zukunft.

Als Ausgangspunkt ist ein Ansatz zur Strukturierung des Security-Engineering-Prozesses vom Projektpartner admeritia GmbH erarbeitet worden [2]. Die Bild 1 zeigt ein methodenneutrales Prozessmodell für das Security-Engineering. Dieser beinhaltet vier aufeinander aufbauende Ebenen, in die sich bestehende Vorgehensweisen einordnen lassen. Der Leuchtturm beschreibt ein allgemeines Security-Engineering-Prozessmodell für den gesamten Security-

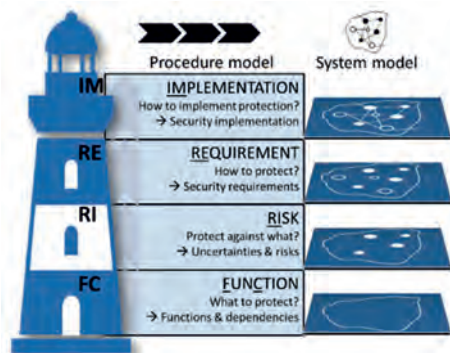


Bild 1: Allgemeines Prozessmodell für Security-Engineering [2]

Engineering-Prozess: Ausgehend von der Identifikation schutzbedürftiger Funktionen (FC) werden Risiken (RI) abgeleitet, Security-Anforderungen (RE) gegen diese Risiken definiert und in einer Security-Lösung (IM) implementiert. [2]

Damit die Methodenneutralität des Modells bestehend bleibt, können die verschiedenen Security-Risikoanalyse- oder Systems-Security-Engineering-Methoden in den vier Ebenen eingesetzt werden, sodass keine Entscheidungen für oder gegen einen bestehenden Standard getroffen werden müssen.

Im Rahmen dieses Beitrags werden die Autoren für die erste Ebene des Turmes (FC) die Anforderungen an ein Security-Informationsmodell sowie vier Ansätze für ein konkretes Modell mit einem Beispielen beschreiben. Dazu werden die Modellierungsansätze mit AutomationML mit ihren Vor- und Nachteilen beschrieben und verglichen. Die anderen Ebenen sind Gegenstand zukünftiger Arbeiten.

Darüber hinaus wurde seitens der Autoren bereits in [3] anhand von Anwendungsfällen die Notwendigkeit eines Security-Engineering-Informationsmodells beschrieben und eine Anforderungsanalyse an das Informationsmodell abgeleitet. Es beinhaltet einen Modellierungsansatz in AutomationML zur Unterstützung der Anwendungsfälle und der Erfüllung der Anforderungen. Dennoch existieren mehrere Modellierungsansätze für Security-Merkmale, welche in der bisherigen Arbeit nicht erwähnt worden sind.

2. Stand der Technik

Als Technologie für das Informationsmodell wählen die Autoren die neutrale Modellierungssprache AutomationML (AML) [4] [5]. Innerhalb des Automation-Engineerings mit seiner Vielfalt an beteiligten Ingenieurdisziplinen von Verfahrenstechnik über Regelungstechnik und Robotik bis hin zu funktionaler Sicherheit wird die Herausforderung der verschiedenen Ingenieursprachen, diese in Einklang zu bringen, mit Hilfe von AutomationML verfolgt. Insbesondere für die Zielsetzung des vorliegenden Security-Engineering-Informationsmodells kann AutomationML angewendet werden.

Darüber sind in der Literatur erfolgreiche Ansätze in AutomationML beschrieben, auf denen das neu zu entwickelnde Security-Informationsmodell aufbauen kann. AutomationML ermöglicht die Modellierung von Anlagenkomponenten und Kommunikationssystemen in Form hierarchischer Objektmodelle. Auch die Beschreibung von Geometrie, Kinematik, Dokumentation, Schnittstellen oder Logik können in Form von Objekten und separaten Dateiformaten abgelegt werden [6]. Es ist anzumerken, dass für das Security-Engineering-Informationsmodell ein gewisser Abstraktionsgrad notwendig ist und nicht alle Teilaspekte von AML notwendig sind.

Die Modellierung von Kommunikationsnetzwerken ist für das Security-Engineering ein wesentlicher Bestandteil des Informationsmodells. Durch die AML-Reihe IEC 62714 (insbesondere der 2022 standardisierte Teil 5) werden Modellierungsempfehlungen für Kommunikationsnetzwerke geliefert, die eine Klassifikation der Verbindungen zwischen logischer sowie physischer Sichten liefert [7]. Für die Security-Analyse ist die logische Sicht des Netzwerkes essenziell. Solch eine Modellierung wurde bereits in [8] publiziert, dass den ersten Ansatz für ein Security-Modell in AML vorstellt. Eine Analyse, der für das Security-Engineering zu modellierenden Informationen und Vorschläge für ihre Abbildung in der Verwaltungsschale ist in [9] erfolgt.

Zur Erleichterung der Modellierung von Prozessabläufen in komplexen Anlagen wird die Dreiteilung der Daten in AML empfohlen. Hierzu wird das Produkt-Prozess-Ressource-Konzept (kurz: PPR-Konzept) angewendet. Das PPR-Konzept erlaubt die interaktive Modellierung von Arbeitsabläufen, Prozessabläufen und Systemkonfigurationen. Für die Security-Analyse können hiermit Funktionen modelliert werden, d. h. eine systematische Beschreibung der zu schützenden Systemkomponenten sowie deren Kommunikation miteinander, der für die Erfüllung eines bestimmten Zwecks im System relevant ist. [4]

Darüber hinaus beschreibt [10] eine umfassende Analyse zu existierenden funktionsbasierten Engineeringansätzen in verschiedenen Domänen. Es beinhaltet die Umsetzung eines funktionsbasierten Systemmodells für die Fertigungsindustrie, welche in Form einer AML-Modellierung und Bibliotheken vorliegt.

Auch wird mit AMLsec nicht nur eine ausführliche Risikobetrachtung nach ISA/IEC 62443-3-2 [11] vorgegeben, sondern auch eine AML-Bibliothek, die security-relevante Informationen wie Security-Geräte und Security-Konfigurationen beinhaltet, welche für das vorliegende Security-Informationsmodell genutzt werden. [12]

In Standardisierung und Forschung werden Konzepte wie die Verwaltungsschale [13], IEC Common Data Dictionary [14], eCI@ss [15] oder das Common Information Model [16] vorangetrieben. Für Security ist hierzu noch kein umfassendes Merkmalmodell vorhanden und wird von den diversen Konzepten zusammenfassend betrachtet und ggf. abgeleitet.

3. Die erste Ebene im Security-Engineering-Prozessmodell: Funktion (FC)

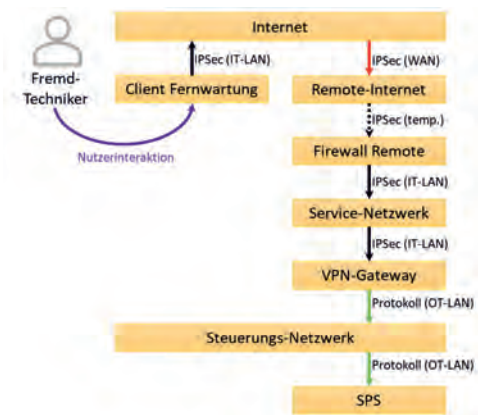


Bild 2: Einfache Funktion - Remote Control

Ein Modell soll eine klare Sicht auf die Anlage als System schaffen. Solch eine Systemmodellierung ist Voraussetzung für die Entwicklung einer Security-Lösung. Das Informationsmodell soll die Anlage explizit und technisch fundiert abbilden und die Komplexität auf security-relevante Aspekte reduzieren können. In dieser sollen alle betroffenen Systemkomponenten und menschlichen Rollen sowie deren Verbindungen mit Informationen zueinander abgebildet werden.

Um Security-Lösungen entwickeln zu können, müssen Risiken im System

erfasst werden. Und um diese zu ermitteln, werden aus dem Systemmodell Funktionen abgeleitet, in der die security-relevanten Informationen im Detail betrachtet werden. Eine Funktion ist eine Zusammenfassung von Systemkomponenten sowie deren Verbindungen, dass einen Zweck im System erfüllt. Sie definiert konkret die zu schützende Systemfunktion im Anwendungsbereich sowie deren Abhängigkeiten von Menschen, Hardware und Software. Als Grundlage für die Modellierung einer Funktion und der ersten Ebene des Security-Engineering-Prozessmodells wird eine vereinfachte Netzwerkstruktur angewendet. **Fehler! Verweisquelle konnte nicht gefunden werden.** zeigt exemplarisch eine Funktion für das Bedienen, Beobachten, Engineering und die Fernwartung durch einen externen Dienstleister. Dabei greift ein Fremd-Techniker über die VPN-Schnittstelle in das Prozessnetz ein und programmiert eine SPS. Im Folgenden werden vier Ansätze zur Modellierung von Funktionen

vorgestellt und bewertet.

4. Anforderungen an die Modellierung der ersten Ebene (FC)

A1 – Transparente Teilmodelle:

Eine der wichtigsten Eigenschaften des Security-Informationsmodells ist die Reduzierung der Komplexität aufgrund der großen Menge an Informationen aus verschiedenen Domänen. Für die Modellierung der ersten Ebene ist die Informationsgewinnung von security-relevanten Eigenschaften wesentlich und setzt einen höheren Abstraktionsgrad an die jeweiligen Objekte vor. Somit wird eine Transparenz des Informationsmodells vorausgesetzt, das alle Information einer Funktion leicht zugänglich und verständlich in klarer und einfacher Form widerspiegelt.

A2.1 – Modellierung von Objekten für Funktionen:

Das Security-Engineering-Modell benötigt für ein rein technisches System zusätzliche Informationen. In einer Funktion werden Objekte mit den Eigenschaften von menschlichen Rollen oder IT/OT-Systemen modelliert, welche Interaktionen und Kommunikation zwischen diesen Objekten beinhalten. Jede Funktion zeigt den Teil des Systems in Form von Objekten, der für die Erfüllung eines bestimmten Zwecks relevant ist. Darüber hinaus kann in der Funktion eine Verbindung zum Risiko (zweite Ebene des Leuchtturms) hergestellt werden.

A2.2 – Modellierung des Informationsflusses:

Eine Funktion bildet den Informationsfluss zwischen den beteiligten technischen Komponenten, Menschen sowie Interaktionen mit Kommunikationsverbindungen ab. Diese müssen übersichtlich und abrufbar sein, da solche technischen Komponenten eine Vielzahl von Verbindungen zu anderen Systemkomponenten besitzen.

A3 – Vermeidung der doppelten Datenhaltung:

In dem Security-Informationsmodell soll Inkonsistenz vermieden werden, indem keine Daten mehrfach mit der gleichen Bedeutung abgelegt werden. Dies ist unter anderem der Fall, wenn für eine Funktion aus dem gesamten Systemmodell eine Kopie des Objektes erstellt wird. Somit besteht nicht die Möglichkeit, eine Änderung im Systemmodell durchzuführen, wenn Eigenschaften eines Objektes in einer Funktion manipuliert werden. Eine doppelte Datenhaltung führt zu mehr Arbeitsaufwand, da Daten mehrfach angepasst werden müssen. Dies kann auch zu einer Mutationsanomalie führen, wenn diese Änderungen nicht fehlerfrei durchgeführt werden, oder zu Verarbeitungsfehlern, wenn sich die Daten widersprechen.

A4 – Nutzbarkeit von vordefinierten Bibliotheken:

Für die wichtigsten Systemfunktionen kann eine zusätzliche Bibliothek genutzt werden, in der bereits etablierte Funktionsvorschläge als Musterlösungen abgelegt sind. Diese beschreiben alle notwendigen Informationen wie betroffene Anlagenkomponenten aus dem Netzwerk, Rollen und Zugehörigkeiten von Risikoszenarien. Darüber hinaus kann mit einer

Rollenvergabe eine bessere Durchsuchbarkeit und maschinelle Verarbeitbarkeit der Objekte durch das Filtern und Selektieren gewährleistet werden. Auch muss die Möglichkeit gegeben sein, dass die jeweiligen Ansätze zur Modellierung der ersten Ebene in Bibliotheken abgeleitet werden kann.

5. Modellierungsansätze für die erste Ebene (FC) mit AutomationML

5.1. Übersicht

Dieses Kapitel widmet sich der Modellierung mit der Objektmodellierungssprache AutomationML. Analog zu einer gesprochenen Sprache, mit der eine Aussage auf verschiedene Weisen formuliert werden kann, kann AutomationML einen gewünschten Sachverhalt mit Hilfe verschiedener Modellierungsansätze unterschiedlicher Ausdruckskraft abbilden. Im Folgenden stellen die Autoren verschiedene Ansätze vor, um ein Security-Informationsmodell mit AML zu modellieren. In Kapitel 1 erfolgt ein Vergleich und Empfehlung.

5.2. Ansatz 1: das AML-Gruppen-Konzept

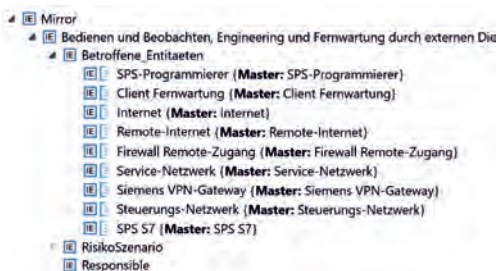


Bild 3: Mirror-Konzept

In einem AML-Informationsmodell kann eine Separierung der Objekte von Strukturen oder Netzwerken mithilfe des AML-Gruppen-Konzeptes realisiert werden. Die Objekte in der Funktion werden durch ein gespiegeltes Objekt (Mirror-Objekt) modelliert. Durch die Referenzierung im Mirror-Objekt wird mit der eindeutigen globalen ID (Globally

Unique Identifier; kurz: GUID) des Masters ein Verweis auf das Informationsmodell generiert. Dadurch besteht stets ein direkter Bezug und Änderungen können ohne doppelte Datenhaltung durchgeführt werden. Die Funktion wird anhand einer Hierarchie modelliert wie in Bild 3 dargestellt. Anzumerken ist, dass die Eigenschaften von und zwischen den Objekten der Funktion ausschließlich in den Master-Objekten modifiziert werden können. Sollten mehrere Verbindungen zwischen den jeweiligen Objekten bestehen oder mehrere Protokolle werden ausgetauscht, kann in diesem Konzept nicht entschieden werden, welche für die Funktion genutzt werden. Darüber hinaus werden menschliche Interaktionen im Informationsmodell nicht abgebildet (siehe [3]) und die Art der Nutzerinteraktion bezogen auf das Folgeobjekt kann nicht modelliert werden. Durch das Gruppen-Konzept wird lediglich eine Model-

lierung einer Hierarchie sowie alle Verbindungen der betroffenen Entitäten abgebildet, die für

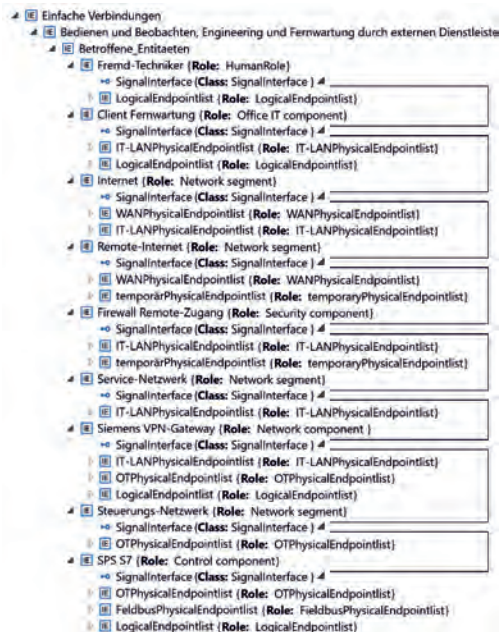


Bild4: Funktion mit einfachen Verbindungen

besteht die Möglichkeit, die Nutzerinteraktion einer menschlichen Rolle im Modell abbilden zu können. Demzufolge können nur Informationen zu den Verbindungen in Form von Attributen definiert werden. Eine Modellierung von weiteren Objekten (CAEX InternalElements) an dem InternalLink ist nicht möglich.

die Funktion nicht von Bedeutung ist.

5.3. Ansatz 2: Verbindungen zwischen den Entitäten

Ein zweiter Ansatz (siehe Bild4) beschreibt die Funktion, indem die Objekte aus dem gesamten technischen Systemmodell kopiert werden. Dadurch besteht kein direkter Bezug mehr zum Modell und Änderungen müssen nachträglich durchgeführt werden. Die „ExternalInterface“-Schnittstelle, welche die Verbindungen zwischen den Objekten herstellt, muss bei der Modellierung der Funktion nachträglich hinzugefügt werden. Zusätzliche security-relevante Informationen, wie die Richtung, Reihenfolge oder Art der Verbindung werden als Attribut im InternalLink (Verbindung zwischen Objekten) gespeichert. Damit

5.4. Ansatz 3: Verbindungen zwischen den Entitäten nach IEC 62714-5

Für die Modellierung von Kommunikationsstrukturen wird das empfohlene Konzept für Netzwerkkommunikationen von AML angewendet. Das Netzwerk besteht aus zwei Ebenen von Informationen: die logische und die physische Sicht, welche mithilfe der Bibliotheken aus [7] dargestellt werden. Für die Zuordnung der Sichten wird das OSI-Modell [17] herangezogen. Dabei entspricht die 1. und 2. Schicht für die physische und die 3. bis 7. für die logische Ebene. In der Regel ist die logische Sicht für die Security-Analyse von Bedeutung. Auch wird in der 7. Schicht (Anwendungsschicht) die menschliche Interaktion beschrieben. Der dritte Modellierungsansatz ähnelt dem Ansatz aus 5.2. Als zusätzliche Änderungen werden die

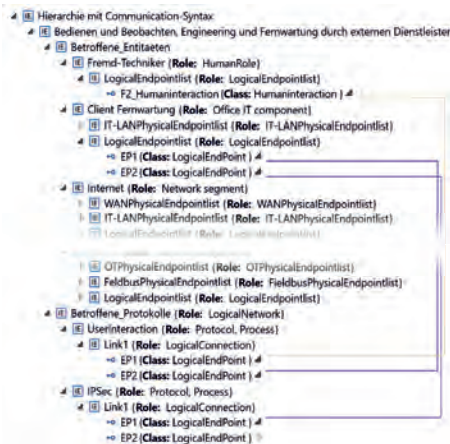


Bild 5: Ausschnitt der Funktion mit der Kommunikationssyntax aus IEC 62714-5

Verbindung zwischen den Objekten als InternalElements zusätzlich modelliert. Dementsprechend können zusätzliche Information in Form von weiteren Objekten (Internal-Elements) bei den Verbindungen zugeordnet werden. Die betroffenen Komponenten aus dem Informationsmodell werden in die Funktion kopiert, dass wiederum zu einer doppelten Datenhaltung führt. Dementsprechend hat die Funktion keinen direkten Bezug zum gesamten technischen Systemmodell.

5.5. Ansatz 4: „Sender-Process-Receiver“-Konzept (SPR-Konzept)

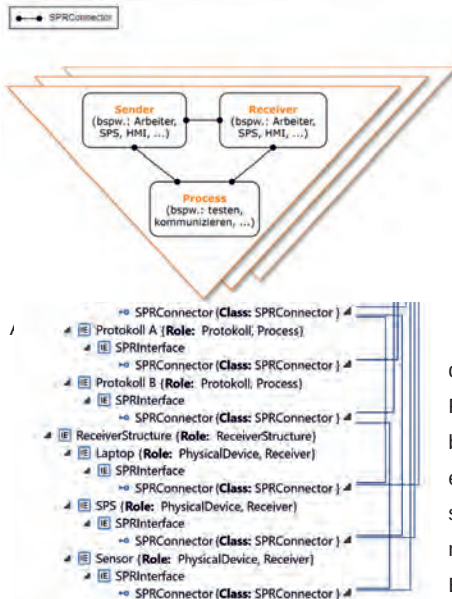


Bild 6: Funktionsmodellierung mit dem SPR-Konzept (aufgrund der Größe der Funktion wurden die Inhalte angepasst)

Mit dem folgenden Absatz wird der vierte Ansatz (Bild 6) beschrieben: Bei der Strukturierung komplexer Anlagen hat sich in der Praxis der AutomationML die Dreiteilung der Daten in Ressourcen, Prozesse und Produkte bewährt, die auch als PPR-Konzept bezeichnet wird [4]. Dieses Konzept bietet die Grundlage zur Darstellung der Informationsmodellierung der Funktionen. Aufgrund der Tatsache, dass das klassische PPR-Konzept auf Produkten, Ressourcen und Prozessen basiert, sind menschliche Rollen nicht enthalten. Dementsprechend ist eine Unterscheidung von Produkt und Ressource nicht möglich. Deshalb empfehlen die Autoren eine Erweiterung des PPR-Konzeptes für das Security-Informationsmodell. Folgende Rollen werden hierfür eingesetzt: Sender, Prozess sowie Receiver (Empfänger) und die Schnittstelle „SPRConnector“. Ein Sender führt einen Prozess am Receiver durch. Mit diesem Konzept können die Funktionen wie einfache Sätze gelesen werden: "Techniker programmiert SPS". Die Funktion besteht aus mehreren Interaktionsdreiecken (siehe Abbildung 7). Dadurch werden für den Betrachter der Funktion drei Blick-richtungen mit ihren Objekten und Rollen generiert. Die Reihenfolge der Interaktions-dreiecke wird in der AML-Datei mit der InterfaceClass „Order“ realisiert, welche sich beim Sender befindet. Die Objekte werden, wie bei den beiden vorherigen Ansätzen aus dem Informationsmodell kopiert. Eine Spiegelung der Objekte (siehe 5.2 – Gruppen-Konzept) ist nicht möglich, da die Rollen *Sender* und *Receiver* jeweils getrennt an die gleichen Objekte in der Funktion vergeben werden. Dadurch gibt es bspw. die SPS zweimal in der Funktion mit jeweils der Rolle Sender und Receiver in unterschiedlichen Interaktionsdreiecken.

6. Vergleich der Ansätze und Ableitung einer Modellierungs-Lösung

In diesem Abschnitt werden die vorgestellten Modellierungsansätze verglichen und bewertet. Anhand dieser Entscheidung können nächste Schritte für die Modellierung von Risiken, Anforderungen und Implementierungsmöglichkeiten erstellt werden. Die folgende Tabelle zeigt, ob die jeweiligen Ansätze die Anforderungen erfüllen.

Tabelle 1: Vergleich – Gegenüberstellung der Anforderungen zu den Ansätzen (grün: erfüllt; orange: nicht empfohlen; rot: nicht erfüllt)

	A1 - Transparente Teilmodelle	A2.1 - Modellierung von Objekten für Funktionen	A2.2 - Modellierung des Informationsflusses	A3 - Vermeidung der doppelten Datenhaltung	A4 - Nutzbarkeit von vordefinierten Bibliotheken
#1 Gruppen-Konzept	orange	orange	rot	grün	grün
#2 Einfache Verbindungen	grün	orange	grün	rot	grün
#3 Verbindungen nach IEC 62714-5	grün	grün	grün	rot	grün
#4 SPR-Konzept	grün	grün	grün	rot	grün

Für die Security-Analyse ist es erheblich, dass alle relevanten Informationen aus dem gesamten Informationsmodell in einer Funktion abgebildet sind. Da das (#1) Gruppen-Konzept nicht aufweisen kann, welche Protokolle für die Funktion genutzt werden, ist dieser Ansatz nicht zu empfehlen. Darüber hinaus wird die Sicht der Funktion – das Denkmodell – verlassen, indem das vollständige Informationsmodell betrachtet wird. Der Ansatz mit den (#2) einfachen InternalLinks zwischen den Objekten zeigt, dass die Vollständigkeit der Informationen nicht gewährleistet werden kann, da alles in Form von Attributen gespeichert werden muss. Wird der Ansatz mit den (#3) Verbindungen nach IEC 62714-5 mit dem (#4) SPR-Konzept verglichen, so ist ersichtlich, dass der Anwender diverse Informationen aus dem SPR-Konzept mit den Sichten durch die Rollenvergabe gewinnen kann. Da die doppelte Datenhaltung bei großen Netzwerkstrukturen auch zu einem Risiko führen kann, wird die Mirror-Eigenschaft aus dem (#1) Gruppen-Konzept mit dem (#4) SPR-Konzept vereint.

Die Zusammenführung der Mirror-Eigenschaft mit den AML-Objekten wird ebenfalls bei „Module Type Package“-AutomationML-Dateien [18] verfolgt. Die Modellierung zur Vermeidung der doppelten Datenhaltung wird in Bild 8 dargestellt. Dabei wird dem jeweiligen

Objekt in der Funktion ein Attribut mit dem Type „xs:IDREF“ vergeben, dass auf das Master-Objekt mit dieser GUID verweist. Dadurch wird gewährleistet, dass durch eine verarbeitende Software die jeweiligen Objekte im Modell selektiert und dementsprechend angepasst werden.

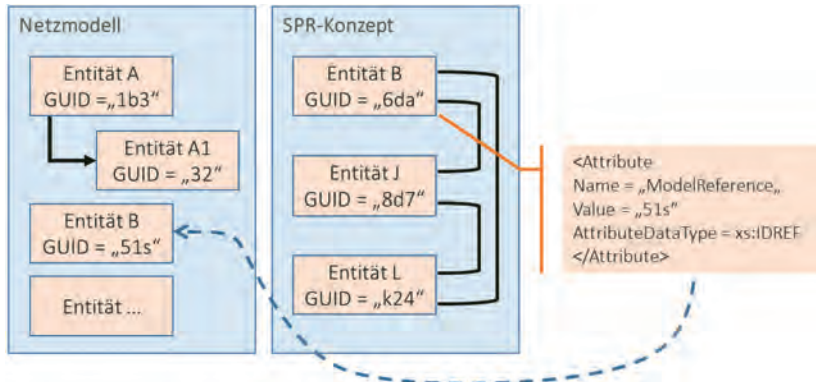


Bild 8: Vermeidung der doppelten Datenhaltung mit dem SPR-Konzept

7. Zusammenfassung und Ausblick

Mit den vorgestellten Ansätzen werden Modellierungsmöglichkeiten für die erste Ebene des Security-Engineering-Prozessmodells evaluiert mit dem Ziel, die Workflows von Security- und Automation Engineering zu vereinen. Dadurch wird eine Grundlage für die Durchführung einer Security-Analyse, insbesondere für die Ableitung von Risiken, Security-Anforderungen und Security-Lösungen geschaffen. Mithilfe des SPR-Konzeptes und dem Zusatz zur Vermeidung der doppelten Datenerhaltung kann das zu schützende System mit allen notwendigen security-relevanten Informationen modelliert werden.

Die Integration von security-relevanten Informationen in maschinenlesbare, maschinenverarbeitbare und maschinenverständliche Datenmodelle wie AutomationML, die bereits im Bereich der Automatisierungstechnik etabliert sind, ist ein wichtiger Schritt zur Integration von Security in den gesamten Prozess der Automatisierungstechnik. AutomationML war bereits Grundlage für eine Vielzahl von Informationsmodellen in der Industrie und ist als Technologie zur Modellierung von Teilmodellen für Engineering-Informationen in die Industrie 4.0 Verwaltungsschale bereits eingebettet.

Zusätzlich ist die Kompatibilität mit der Verwaltungsschale (VWS) ein bedeutsamer Aspekt für die Zukunftstauglichkeit des Security-Informationsmodells. Das im Rahmen des Projektes IDEAS zu entwickelnde methodenneutrale Informationsmodell dient als Datenlieferant für das

Teilmodell „Security-Engineering“ der VWS. Teilmodelle sind zentraler Bestandteil der „Industrie 4.0 Verwaltungsschale“ und können innerhalb einer VWS voneinander getrennt betrachtet werden. Unterschiedliche Teilmodell repräsentieren die verschiedenen Aspekte eines Assets über den gesamten Lebenszyklus hinweg [19] [20].

In diesem Beitrag wird das Fundament für die Informationsmodellierung des Security-Engineerings gelegt. Die Modellierung der nächsten Schritte im Security-Engineering-Prozess (restliche Ebenen im Security-Engineering-Prozessmodell: Risiken, Anforderungen und Implementierung) ist Teil zukünftiger Arbeiten im Forschungsprojekt.

In Zusammenarbeit mit dem NAMUR-Arbeitskreis 1.3 (Informationsmanagement und Werkzeuge) wird das Informationsmodell für das Security-Engineering als UML-Modell modelliert. Darüber hinaus wird die Verwendung des Security-Informationsmodells als Teilmodell der Verwaltungsschale angestrebt.

8. Danksagung

Diese Arbeit wird unterstützt vom Bundesministerium für Bildung und Forschung (BMBF) im Rahmen des Projektes „IDEAS“ (Integrated Data Models for the Engineering of Automation Security, Vorhaben 16KIS1269K). [21]

9. Literatur

- [1] BMWi, „Der IT-Sicherheitsmarkt in Deutschland: Grundstein für eine makroökonomische Erfassung der Branche.“, 2013.
- [2] S. Fluchs und H. Rudolph, „Wie OT-Security-Engineering eine Ingenieurwissenschaft wird - Ein Denkmodell und ein Datenmodell,“ atp magazin, 2019.
- [3] E. Tastan, S. Fluchs und R. Drath, „Warum wir ein Security-Engineering-Informationsmodell brauchen - Motivation, Anwendungsfälle und Konzept für ein neues Domänenmodell für Security-Engineering,“ Hochschule für Technik, Wirtschaft und Kultur Leipzig, 18. AALE-Konferenz. Pforzheim, 2022.
- [4] „IEC 62714 Ed. 2: Engineering data exchange format for use in industrial automation systems engineering (AutomationML). International Electrotechnical Commission,“ IEC, 2018.
- [5] R. Drath (Ed.), AutomationML – The Industrial Cookbook, Berlin: De Gruyter Oldenbourg, 2021.
- [6] M. Rentschler, R. Drath, J. Hinze und A. Gatterburg, „AutomationML als generische Beschreibungssprache für den Digitalen Zwilling,“ VDI-Verlag, Baden-Baden, 2019.
- [7] „DIN EN 62714-5: Datenaustauschformat für Planungsdaten industrieller Automatisierungssysteme - Automation markup language - Teil 5: Kommunikation,“ 2020.
- [8] S. Fluchs, N. Schmidt und A. Mendl-Heinisch, „Ein Systemmodell für Security-Engineering,“ TeSi, vol. 9, no. 11–12, 2019.
- [9] S. Fluchs, „On Modelling for Security Engineering. fluchsfriktion 37,“ 2021.
- [10] A. Scholz, „Unterstützung des Engineerings von fertigungstechnischen Produktionssystemen mit Hilfe von Maschinenfunktionen - Methode, Modell und Beschreibungsmittel für ein funktionsorientiertes Planen,“ Helmut-Schmidt-Universität / Universität der Bundeswehr Hamburg, Hamburg, 2018.
- [11] „ISA/IEC 62443-3-2: Security for industrial automation and control systems - Part 3-2: Security risk assessment for system design,“ 2020.
- [12] M. Eckhart, A. Ekelhart und E. R. Weippl, „Automated Security Risk Identification Using AutomationML-based Engineering Data,“ IEEE Trans. Dependable and Secure Comput., 2020.
- [13] BMWi, Plattform I4.0: Usage View of Asset Administration Shell, Berlin:, 2019.

- [14] „IEC 61987-4 - SC 65E/WG 2: Common Data Dictionary,“ [Online]. Available: <https://cdd.iec.ch/cdd/iec61987/iec61987.nsf/>.
- [15] A. Bondza, C. Eck, R. Heidel, M. Reigl und S. Wenzel, „eCI@ss: Toward Smart Manufacturing with Data and Semantics,“ eCI@ss e. V., Köln, 2018.
- [16] DMTF, „Common Information Model (CIM) Metamodel,“ 2014.
- [17] H. Zimmermann, „OSI Reference Model--The ISO Model of Architecture for Open Systems Interconnection,“ IEEE Transactions on Communications.
- [18] M. Hoernicke, K. Stark und L. Urbas, „MTP – Automation Engineering of Modular Process Plants,“ in *AutomationML - The Industrial Cookbook*, De Gruyter Oldenbourg, 2021, pp. 125-164.
- [19] R. Drath, S. Malakuti, S. Grüner, J. Grothoff, C. Wagner, U. Epple und M. Hoffmeister, „Die Rolle der Industrie 4.0 „Verwaltungsschale“ und des „digitalen Zwillings“ im Lebenszyklus einer Anlage - Navigationshilfe, Begriffsbestimmung und Abgrenzung,“ VDI-Verlag - Tagungsband zur Automation 2017, Baden-Baden, 2017.
- [20] BMWi - Plattform Industrie 4.0, „Verwaltungsschale in der Praxis,“ 2020.
- [21] BMBF, „IDEAS - Integrierte Datenmodelle zur Berücksichtigung von IT-Sicherheit im Entwicklungsprozess industrieller Anlagen,“ 2021. [Online]. Available: <https://www.forschung-it-sicherheit-kommunikationssysteme.de/projekte/ideas>.

IT-Security für Automatisierungssysteme mit Ethernet-APL-Feldgeräten

Anforderungen und Schutzmaßnahmen

Karl-Heinz Niemann, Hochschule Hannover;
Simon Merklin, Endress + Hauser Digital Solutions, Reinach, Schweiz

Kurzfassung

Die Konvergenz der Netze ist ein zunehmender Trend im Bereich der Automatisierung. Immer mehr Anlagenbetreiber streben eine Vereinheitlichung der Netzwerke in ihren Anlagen an. Dies führt zu einer vereinfachten Überwachung und einem reduzierten Schulungsaufwand für das Personal, da nur noch eine einheitliche Netzwerktechnologie beherrscht werden muss. Ethernet-APL ist ein Puzzlestück für ein solches konvergentes Netzwerk. APL ermöglicht Ethernet bis zum Sensor/Aktuator und unterstützt verschiedene Protokolle wie z. B. PROFINET, EtherNet/IP sowie OPC UA. Trotz der erheblichen Vorteile für die Prozessindustrie, die Ethernet-APL mit sich bringt, sollte die IT-Sicherheit solcher Systeme mit betrachtet werden. Dieser Beitrag gibt einen Überblick über die Auswirkungen von Ethernet-APL auf die IT-Security und macht einen Vorschlag, wie die IT-Security für diese Geräte gewährleistet werden kann.

Abstract

The convergence of networks is an increasing trend in the field of automation. More and more plant operators are striving to standardize the networks in their plants. This leads to simplified monitoring and reduced training requirements for personnel, as only one uniform network technology needs to be mastered. Ethernet APL is a piece of the puzzle for such a converged network. APL allows Ethernet up to the sensor/actuator and supports various protocols such as PROFINET, EtherNet/IP as well as OPC UA. Despite the significant benefits for the process industry that APL brings, the OT security of such systems should also be considered. This paper provides an overview of the IT security implications of Ethernet APL and suggests how to ensure It security for these devices.

1. Ethernet-APL in der Prozessindustrie

Der Standard IEEE 802.3cg [1] spezifiziert ein Zweidraht-Ethernet welches für den Anschluss von Sensoren, Aktoren und anderen Geräten vorgesehen ist. Wesentliches Merkmal ist eine Datenrate von 10 Mbit/s (voll-duplex) und die Energieversorgung der Endgeräte über die Ethernet-Leitung. Diese Spezifikation wurde um eine Port-Profile-Specification [2] erweitert, die es erlaubt, das Zweidraht-Ethernet auch in rauen Umgebungen und in explosionsgefährdeten Bereichen zu betreiben. Das Ergebnis nennt sich Advanced Physical Layer for Ethernet – kurz Ethernet-APL [3]. Mögliche Systemstrukturen und die Zuordnung zu Ex-Zonen sind in [4] dargestellt. Im Rahmen dieses Beitrages wird von der in Bild 1 dargestellten Systemstruktur ausgegangen.

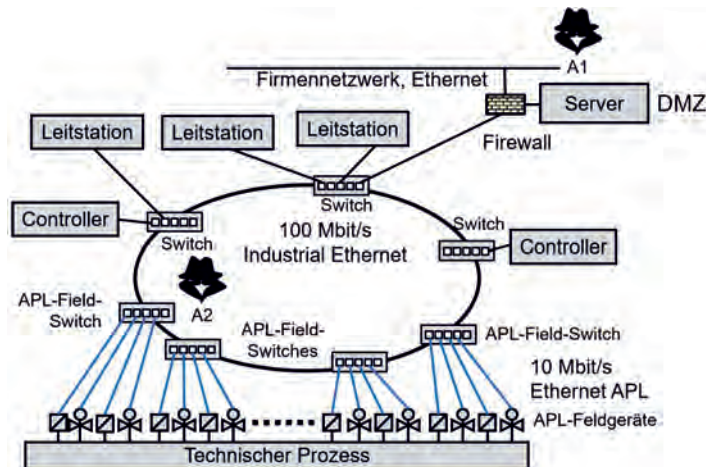


Bild 1: Systemstruktur mit APL-Feldgeräten

Es zeigt sich, dass die Ethernet-APL-Topologie eine Reihe von Vorteilen mit sich bringt: Einheitliches Netzwerk, hohe Datenraten, einfache Übertragung von Diagnosedaten an ein Asset-Management-System werden ermöglicht. Im Hinblick auf die Cybersicherheit hat die flache Netzwerkarchitektur jedoch einen Nachteil: Die Ethernet-APL-Feldgeräte können leichter von einem Eindringling angegriffen werden. Das folgende Kapitel 2 wird sich mit diesem Thema im Detail befassen.

2. Die Security-Herausforderungen für Ethernet-APL Feldgeräte

Der typische Sicherheitskontext eines APL-Feldgeräts ist wie folgt definiert: Die Produktionsanlage ist durch eine Umzäunung geschützt. Der physische Zugriff auf das Gerät ist nur für das in der Anlage arbeitende Personal und für externe Dienstleister, wie Inbetriebnahme- oder Wartungspersonal, möglich. Das Feldgerät ist über Ethernet-APL mit dem Automatisierungsnetzwerk verbunden. Das Ethernet-Netzwerk ist vom Büronetzwerk durch eine Sicherheits-Appliance, wie z.B. eine Firewall, ggf. mit Nutzung einer demilitarisierten Zone (DMZ), getrennt. Das Feldgerät hat keine direkte Verbindung zum Internet.

Im Allgemeinen ist ein Feldgerät über verschiedene Schnittstellen für Cyberangriffe anfällig. Dies kann ein lokales Display in Kombination mit lokalen Tasten, Bluetooth, Wireless HART, Handheld mit kabelgebundenem HART, die Stromschleife, der Feldbus oder die Ethernet-APL-Verbindung sein. Eine detaillierte Risikoanalyse für ein Feldgerät, die alle beschriebenen Kanäle abdeckt, ist in [5] verfügbar. In diesem Beitrag wird nur auf die Ethernet-APL-Verbindung eingegangen.

An dieser Stelle ist festzuhalten, dass APL den Physical-Layer des verwendeten Ethernet-Netzwerkes beschreibt (Layer 1). Oberhalb dieses Layers laufen dann Ethernet (Layer 2) und dann Industrial Ethernet-Protokolle wie PROFINET, EtherNet/IP oder OPC UA. (Layer 7). Die nachfolgend beschriebenen Szenarien betreffen im Wesentlichen die verwendeten Ethernet-basierenden Protokolle. Der Advanced-Physical-Layer steht für den Sachverhalt, dass Ethernet bis an den Sensor bzw. den Aktor geführt wird.

Die Exposition eines Feldgeräts muss in den Kontext möglicher Angreifer, entsprechender Angriffsorten und Angriffsorte gestellt werden. In diesem Beitrag werden gemäß Bild 1 zwei verschiedene Standorte von Angreifern betrachtet:

- A1: Remote Angreifer, der im Büronetz operiert, die Security Appliance überwindet und über das Ethernet-Netzwerk in die OT-Domäne kommunizieren kann. Dieser Angreifer kann aus dem Büronetz dann genauso agieren, wie der im Folgenden beschriebene Angreifer A2.
- A2: Lokaler Angreifer (Innentäter), der Zugriff auf das Ethernet-Netzwerk in der OT-Domäne hat. Das Thema Innentäter ist im Bereich Automatisierungstechnik existent und wird z. B. in [6] und [7] beschrieben.

Zusätzlich zum Standort des Angreifers sind zwei Arten von Angriffen zu berücksichtigen. Die eine Angriffsart ist ein Denial-of-Service-Angriff (DoS). Ein DoS verhindert den autorisierten Zugriff auf Ressourcen oder die Verzögerung von zeitkritischen Operationen [8]. Diese Angriffsart macht die Anlage funktionsunfähig (Shutdown), führt aber nicht dazu, dass manipulierte IO-Werte übertragen werden können. Ein DoS-Angriff beeinträchtigt das Schutzziel der

"Verfügbarkeit". Ethernet-APL-Feldgeräte sind in der Regel auf Grund des Explosionsschutzes auf einen geringen Energieverbrauch optimiert. Die damit einhergehende limitierte Rechenleistung macht diese Geräte somit zu einem einfachen Ziel für einen DoS-Angriff.

Die zweite Angriffsart betrachtet die Manipulation von IO-Werten. Dies würde dazu führen, dass die Steuerung manipulierte, verfälschte Eingangswerte ausliest, aber der Angriff führt nicht zu einem sofortigen Stopp der Anlage, wie im ersten Fall. Dieser Angriff beeinträchtigt das Schutzziel "Integrität". Die folgenden Abschnitte werden sich auf diesen Aspekt konzentrieren.

Nachdem der Angreifer (Innentäter) sich Zugang zum Netzwerk verschafft hat, kann das APL-Feldgerät direkt identifiziert und angegriffen werden, da es z. B. Teil der PROFINET- oder EtherNet/IP-Kommunikation ist. Beispielhafte Angriffe auf ein PROFINET-Gerät sind z. B. in [11] und [12] beschrieben. Hierbei wird z. B. die Kommunikation zwischen der Steuerung und dem Feldgerät durch einen „Man-in-the-Middle-Angriff“ abgefangen und die Original-Messwerte werden durch manipulierte Werte ersetzt. Diese Art des Angriffs betrifft generell alle Geräte, die dieses Protokoll verwenden und nicht nur Ethernet-APL-Feldgeräte.

3. OT-Security-Bedrohungen und Schutzmaßnahmen für Ethernet-APL-Feldgeräte

Die im vorangehenden Abschnitt beschriebenen Angriffe lassen sich durch den Einsatz kryptografisch abgesicherter Protokolle verhindern. Die Standards Developing Organizations (SDOs) bieten heute bereits eine sichere Kommunikation für EtherNet/IP [10, 11], PROFINET [9] oder OPC UA [12, 13] an, bzw. arbeiten daran.

Man kann festhalten, dass Ethernet-APL-Feldgeräte die Sicherheitsfunktionen des von ihnen unterstützten Kommunikationsprotokolls unterstützen müssen. Bei der Entwicklung sind Vorkehrungen hinsichtlich der Rechenleistung, der Speichergröße und gegebenenfalls der Bereitstellung eines Secure Elements (z. B. eines Trusted Platform Module) zu berücksichtigen. DoS-Angriffe müssen bis zu einem gewissen Grad berücksichtigt werden, auch wenn davon ausgegangen wird, dass die Geräte nicht direkt mit dem Internet verbunden sind und in einer geschützten Umgebung arbeiten. Der nächste Abschnitt zeigt im Detail, welche Sicherheitsanforderungen ein Ethernet-APL-Feldgerät erfüllen muss.

Für die Definition der Sicherheitsanforderungen von Ethernet-APL-Feldgeräten sollten die Design- und Schutzmechanismen einer verfahrenstechnischen Anlage Grundkonzepten nach

IEC 62443-1-1 [14] berücksichtigt werden. Dieser Schritt ist wichtig, da viele Komponentenanforderungen (CRs) und Anforderungserweiterungen (REs) der IEC 62443-4-2 [15] zur Erfüllung der Anforderungen beitragen.

Ein Ethernet-APL-Feldgerät sollte nach dem Secure Development Lifecycle gemäß 62443-4-1 [16] entwickelt werden. Dadurch wird sichergestellt, dass das Gerät die Sicherheitsanforderungen erfüllt, die für den beabsichtigten Einsatz des Geräts angemessen sind. Beispiele für die erfolgreiche Durchführung des Prozesses sind eine Bedrohungsanalyse und ein Anforderungsmanagement, eine Defense-in-Depth-Strategie, sichere Programmierrichtlinien sowie eine Dokumentation der Sicherheitsmerkmale.

Die in Tabelle 1 aufgeführten Beispiele sind eine Auswahl potenzieller Bedrohungen und Schutzmaßnahmen, die für ein sicheres Ethernet-APL-Feldgerät zu berücksichtigen sind. Aufgrund des begrenzten Umfangs dieses Beitrages gibt Tabelle 1 nur die wichtigsten Anforderungen wieder. Um weiteren Bedrohungen in einer Ethernet-APL-Prozessanlage begegnen zu können, sind ggf. weitere Schutzmaßnahmen notwendig.

Tabelle 1: Beispiele für potenzielle Bedrohungen und entsprechende Schutzmaßnahmen

Nr.	Potenzielle Bedrohungen	Mögliche Schutzmaßnahmen	Bezug IEC 62443-4-2 [15]
1	Der Angreifer verändert die Kommunikationsdaten im Netzwerk	Integrität und Authentizität der Kommunikation - Die Integrität und Authentizität der Netzkomunikation zwischen Ethernet-APL-Feldgeräten sollte durch kryptografische Methoden gesichert werden. Dies kann durch ein sicheres Protokoll in Kombination mit einer Public Key Infrastructure (PKI) erreicht werden.	CR 3.1 CR 3.1 RE (1)
2	Angreifer nimmt ohne Authentifizierung an der Netzwerkkommunikation teil	Ethernet-APL-Feldgeräte und andere Komponenten müssen in der Lage sein, sich gegenseitig zu authentifizieren. Dies kann durch ein sicheres Protokoll in Kombination mit einer Public Key Infrastructure in der Anlage erreicht werden.	CR 1.8

3	Angreifer gibt sich als APL Field Device aus	Das Ethernet-APL-Feldgerät muss in der Lage sein, die Echtheit der Firmware zu prüfen. Dies kann mit Hilfe von kryptographischen Hashes erfolgen.	EDR 3.12
4	Angreifer manipuliert die Firmware des Geräts	Die Integrität und Authentizität des Bootvorgangs ist ein entscheidendes Merkmal, um sicherzustellen, dass die Gerätefirmware nicht von einem Angreifer manipuliert wurde. Aus diesem Grund muss das Ethernet-APL-Feldgerät während des Bootvorgangs Authentizitätsprüfungen durchführen, um sicherzustellen, dass das Gerät nicht in einen unsicheren oder manipulierten Zustand bootet.	EDR 3.14 EDR 3.14 RE (1)
5	Angreifer manipuliert Firmware durch Update	Das Ethernet-APL-Feldgerät ist updatefähig und akzeptiert nur signierte Firmware vom Hersteller.	EDR 3.10 EDR 3.10 RE (1)
6	Angreifer verschafft sich Zugang zu sensiblen Daten (wie Anmeldeinformationen und Authentifikatoren) von Ethernet-APL-Geräten	Vertraulichkeit der kritischen Daten: Das Ethernet-APL-Feldgerät schützt Daten, die für den sicheren Betrieb der Anlage entscheidend sind.	CR 4.1
7	Der Angreifer führt einen DoS-Angriff auf das Netz durch	Das APL-Feldgerät schützt sich gegen DoS-Angriffe, um die wesentliche Funktionalität des Gerätes zu erhalten. Dies kann z.B. dadurch geschehen, dass Datenpakete, die das Gerät erreichen und für die wesentliche Funktion des Gerätes nicht relevant sind, verworfen werden.	CR 7.1
8	Besondere Anforderungen an Ethernet-APL-Feldgeräte, wie lange	Beim Gerätedesign sollte berücksichtigt werden, dass Ethernet-APL-Feldgeräte mitunter eine sehr lange Produktlebensdauer haben. Ein angemessenes Schutzniveau der verwendeten kryptografischen Algorithmen sollte	---

	Produktlebensdauer und Ressourcenbeschränkungen	ebenfalls berücksichtigt werden, ebenso wie die besonderen Anforderungen für Geräte mit eingeschränkten Ressourcen [17].	

Es lässt sich aus den vorangehenden Beschreibungen ableiten, dass APL-Feldgeräte die gleichen OT-Security-Anforderungen erfüllen müssen, wie andere Automatisierungskomponenten auch. Die Ressourcenbeschränkung in Bezug auf Rechenleistung und Speicherausbau stellt dabei eine besondere Herausforderung dar. Durch die direkte Einbindung von APL-Feldgeräten in das Automatisierungsnetzwerk ist davon auszugehen, dass diese Feldgeräte möglicherweise auch Schwachstellen aufweisen können, die durch ein Softwareupdate zu beheben sind. Daher ist dem Aspekt des Schwachstellenmanagements und der Updatefähigkeit besondere Aufmerksamkeit zu widmen.

4. Schwachstellenmanagement

Der sichere Entwicklungslebenszyklus der IEC 62443-4-1 sieht vor, dass Hersteller, die ihre Kunden über Schwachstellen informieren. Der Prozess zur Offenlegung von Schwachstellen ist in der IEC ISO/IEC 29147 [18] beschrieben. Den Prozess zur Behandlung von Schwachstellen beschreibt die ISO/IEC FDIS 30111 [19]. Ein beispielhafter Melde-Prozess ist in Bild 2 dargestellt:



Bild 2: Melde- und Bearbeitungsprozess für Schwachstellen

Meldung: Bei der Schwachstellenmeldung kontaktiert der Melder (z. B. Anwender oder Sicherheitsforscher) die Herstellerfirma über einen vorgegebenen (vorzugsweise verschlüsselten) Kanal. Meldungen sollten den Produkttyp, die Seriennummer, Firmware und die Beschreibung der Schwachstelle umfassen. **Analyse:** Im nächsten Schritt analysiert die Herstellerfirma, inwiefern das Produkt von der Schwachstelle betroffen ist und ordnet die Schwachstelle nach einem Scoring-System (CVSS-Score) [20] ein. **Lösung:** Falls die Herstellerfirma von der Schwachstelle betroffen ist, erarbeitet das Unternehmen das sogenannte Advisory. Dies ist ein Dokument für die Bekanntmachung von Schwachstellen. Diese enthalten eine Beschreibung der Schwachstelle, die betroffenen Produkte, sowie eine Mitigation (extern oder Update) für die Produkte. **Veröffentlichung:** Für eine strukturierte Veröffentlichung von Advisories bietet sich einerseits die Website der Herstellerfirma, sowie sogenannte übergeordnete CERTs an.

Oft müssen neben einem Advisory auch sogenannte Security Updates von Herstellern zur Verfügung gestellt werden. Ein Update ist laut IEC 62443-4-2 [15]: „[Der] schrittweiser Übergang zu neuerer Hardware oder Software, um IT-Sicherheitslücken, Programmfehler, Zuverlässigkeits- oder Funktionsprobleme zu beseitigen“. Die prozessualen Anforderungen bezüglich Security Updates sind in der IEC 62443-4-1 [16] in den Anforderungen SUM-1 bis SUM-5 definiert. Darin werden, neben der Notwendigkeit für Updates selbst, auch die Anforderungen an die Dokumentation, das Testen, sowie die sichere Auslieferung der Updates beschrieben. Weitere Anforderungen an die Geräte finden sich in der 62443-4-2 [15] in der Anforderung EDR 3.10 und RE (1).

5. Software Updates für APL-Feldgeräte

Wird ein Software Update vom Hersteller bereitgestellt, stellt sich dem Anlagenbetreiber die Frage, ob das Update tatsächlich vom Hersteller kommt und somit authentisch ist. Für PC-basierte Systeme (Windows) existiert schon seit längerem ein Verfahren zur Signierung von SW-Updates unter Verwendung digitaler Zertifikate [21]. Dieses Verfahren ist für Embedded Devices, wie z. B. APL-Messumformer auf Grund der die Einsatzbedingungen in einer OT-Umgebung nur bedingt einsetzbar, da man davon ausgehen muss, dass möglicherweise keine Verbindung zum Internet und keine Uhrzeitführung vorhanden ist. Aus diesem Grund kommt für die Sicherstellung der Authentizität und Integrität von Software Updates ein an die OT-Umgebung angepasstes Verfahren zum Einsatz.

Das hier beschriebene Verfahren nutzt ein asymmetrisches Kryptographie-Verfahren für die Sicherstellung von Integrität und Authentizität von Updates [22]. Ein asymmetrisches Schlüsselpaar besteht immer aus einem Private Key und einem Public Key. Der Private Key muss geheim gehalten werden und dient der Signaturerstellung. Der Public Key ist öffentlich und wird verwendet, um die Signatur zu prüfen.

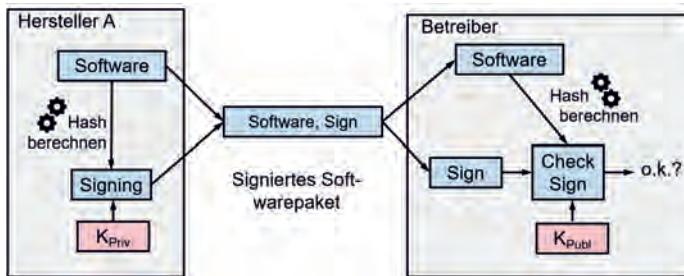


Bild 3: Vereinfachter Signierungs- und Prüfprozess für Software

Bild 3 zeigt ein vereinfachtes Verfahren zur Signierung von Software für eingebettete Geräte, wie APL-Messumformer. Hersteller A erstellt ein Software-Update-Paket für einen Messumformer. Aus diesem Paket wird eine Prüfsumme (Hash) berechnet. Unter Verwendung des Hashwertes und des privaten Schlüssels (K_{Priv}) wird eine Signaturinformation ($Sign$) berechnet und an das Softwarepaket angehängt und an den Betreiber geliefert. Der Betreiber möchte nun das signierte Software Update installieren. Im zu aktualisierenden Gerät ist der öffentliche Schlüssel (K_{Publ}) des Herstellers A sicher gespeichert. Vor der Installation wird die mit dem Paket verknüpfte Signaturinformation verifiziert. Hierbei wertet der Prüfalgorithmus den Inhalt der Software-Lieferung aus und prüft unter Nutzung von K_{Publ} , ob Inhalt der Lieferung und Signaturinformation zusammenpassen und ob die Signierung mit dem privaten Schlüssel K_{priv} des Herstellers erfolgt ist. Ist die Prüfung erfolgreich sind Integrität (das Paket wurde nicht verändert) und Authentizität (das Paket stammt von Hersteller A) gewährleistet. Ist ein Update Paket nicht mit dem korrekten Private Key signiert, wird das Update-Paket vom APL-Feldgerät nicht akzeptiert. Das hier exemplarisch beschriebene Verfahren ist eines von mehreren möglichen. Einen Überblick über weitere Verfahren liefert [23].

Die Sicherheit des beschriebenen Verfahrens hängt an der sicheren Verwahrung des privaten Schlüssels beim Hersteller A. Dieser ist mit besonderer Sorgfalt zu verwahren. So fordert die ISO 27002 [24] „Alle kryptographischen Schlüssel sollten gegen Veränderung und Verlust geschützt sein. Zusätzlich sollten geheime und private Schlüssel gegen unbefugte Benutzung und Offenlegung geschützt werden. Einrichtungen zur Erzeugung, Speicherung und Archivierung von Schlüsseln sollten physisch geschützt sein.“ Verlust des privaten Schlüssels, würde dazu führen, dass Dritte in der Lage wären, signierte Firmware-Updates zu erzeugen. Hier ist also besondere Sorgfalt angebracht [25]. Da, anders als bei digitalen Zertifikaten, bei dem beschriebenen Verfahren kein Mechanismus zum Rückruf existiert, kommt in diesem Fall nur

eine arbeitsintensive Aktualisierung des öffentlichen Schlüssels in den Geräten in Frage. Dieses und weitere Verfahren sind in [26] beschrieben.

6. Zusammenfassung und Ausblick

Das Papier zeigte, dass Ethernet-APL Field-Geräte potenziellen Angriffen ausgesetzt sind. Die flache Netzwerkstruktur bietet Angreifern einen einfachen Zugang zu den Geräten, da sie direkt mit dem Anlagennetzwerk verbunden sind. Diese potenzielle Angreifbarkeit gilt jedoch nicht nur für APL-Feldgeräte, sondern für alle mit dem Netzwerk verbundenen Automatisierungskomponenten. Es wurde erläutert, dass Ethernet-APL letztlich nur eine Übertragungsphysik definiert und dass die Absicherung der Kommunikation auf den höheren Ebenen des Protokollstacks erfolgt bzw. erfolgen muss. Es wurde gezeigt, dass für eine kryptografische Absicherung der typischen Protokolle (PROFINET, EtherNet/IP, OPC UA) bereits Konzepte existieren, bzw. in Arbeit sind. Ethernet-APL-Geräte sollten, genauso wie andere Automatisierungskomponenten nach dem sicheren Entwicklungslebenszyklus der IEC 62443-4-1 [16] entwickelt werden. Dies sichert die umfassende Berücksichtigung von IT-Security Anforderungen. Hersteller von Ethernet-APL-Feldgeräten sollten die zukünftige Integration einer Sicherheitsschicht planen und ausreichend Ressourcen (Speicher, Rechenleistung, evtl. ein Secure-Element wie ein Trusted Platform Module) für die Verwahrung von Schlüsselinformationen in ihren Geräten vorsehen.

6. Literatur

- [1] IEEE Standard for Ethernet - Amendment 5: Physical Layer Specifications and Management Parameters for 10 Mb/s Operation and Associated Power Delivery over a Single Balanced Pair of Conductors, IEEE 802.3cg-2019, IEEE Computer Society, Nov. 2019. URL: https://standards.ieee.org/standard/802_3cg-2019.html
- [2] APL Project, Ethernet APL Port Profile Specification: Ethernet-APL Network and Port Requirements. URL: <https://www.profibus.com/download/port-profile-specification-ethernet-apl>
- [3] PROFIBUS und PROFINET International, ODVA Inc., OPC-Foundation, and FieldComm Group, Ethernet to the field: White Paper. URL: <https://www.profibus.com/index.php?eID=dumpFile&t=f&f=107608&token=97abc376e3e83e010df1902d93deba94d1a30d22>
- [4] K.-H. Niemann, Engineering-Richtlinie Ethernet-APL: Planung, Installation und Inbetriebnahme von Ethernet-APL-Netzwerken. URL: <https://www.profibus.com/download/engineering-guideline-ethernet-apl>

- [5] K.-H. Niemann and M. Hoh, "Anforderungen an die IT--Sicherheit von Feldgeräten: Schutzlösungen für hoch vernetzte Produktionsanlagen," atp-edition, vol. 59, no. 12, pp. 42–53, 2017. URL: https://atpinfo.de/wp-content/uploads/2022/02/atp-edition_2017_12_Beitrag_Niemann_Hoh_de.pdf
- [6] Verizon Enterprise, 2019 Data Breach Investigations Report. [URL: <https://enterprise.verizon.com/resources/reports/2019-data-breach-investigations-report.pdf>]
- [7] Bundesamt für Sicherheit in der Informationstechnik, Industrial Control System Security: Innentäter. URL: https://www.allianz-fuer-cybersicherheit.de/ACS/DE/_downloads/technik-risikomanagement/BSI-CS_061.html.
- [8] An Introduction to Information Security, NIST 800-12, NIST National Institute of Standards and Technology. URL: <https://doi.org/10.6028/NIST.SP.800-12r1>.
- [9] PROFIBUS Nutzerorganisation e.V., Security Extensions for PROFINET - PI White Paper for PROFINET. URL: <https://www.profibus.com/download/pi-white-paper-security-extensions-for-profinet/>.
- [10] ODVA Inc., Securing Ethernet/IP Networks. URL: https://www.odva.org/wp-content/uploads/2020/05/PUB00269R1.1_ODVA-Securing-EtherNetIP-Networks.pdf.
- [11] J. Visoky and J. Wiberg, CIP Security and IEC 62443-4-2. URL: https://www.odva.org/wp-content/uploads/2020/05/2020-ODVA-Conference_CIP_Security_and_IEC_62443_Visoky_Wiberg_Final.pdf.
- [12] OPC unified architecture - Part 2: Security Model, IEC TR 62541-2:2020, IEC- International Electrotechnical Commission, Nov. 2020. URL: <https://webstore.iec.ch/publication/61110>.
- [13] Bundesamt für Sicherheit in der Informationstechnik, Sicherheitsanalyse OPC UA. URL: https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Publikationen/Studien/OPCUA/OPCUA.pdf.jses:sionid=DC17E3A6357837DF12ED131165F080A2.2_cid368?blob=publication-File&v=2.
- [14] Industrial communication networks - Network and system security - Part 1-1: Terminology, concepts and models, IEC TS 62443-1-1:2009, IEC- International Electrotechnical Commission. URL: https://webstore.iec.ch/preview/info_iec62443-1-1%7Bed1.0%7Den.pdf.
- [15] IT-Sicherheit für industrielle Automatisierungssysteme - Teil 4-2: Technische Sicherheitsanforderungen an Komponenten industrieller Automatisierungssysteme (IACS) (IEC 62443-4-2:2019); Deutsche Fassung EN IEC 62443-4-2:2019, DIN EN IEC 62443-4-2, DKE-Deutsche Kommission Elektrotechnik, Elektronik Informationstechnik DIN und VDE, Dec. 2019.

- [16] IT -Sicherheit für industrielle Automatisierungssysteme - Teil 4-1: Anforderungen an den Lebenszyklus für eine sichere Produktentwicklung (IEC 62443-4-1 :2018); Deutsche Fassung EN IEC 62443-4-1 :2018, DIN EN IEC 62443-4-1 (VDE 0802-4-1), DKE-Deutsche Kommission Elektrotechnik, Elektronik Informationstechnik DIN und VDE, DIN Deutsches Institut für Normung e. V, Berlin, Oct. 2018.
- [17] K.-H. Niemann, "IT-Security-Konzepte für die Prozessindustrie: Anforderungen im Kontext von Industrie 4.0," atp-edition, vol. 56, 7-8/2014, pp. 62–69, 2014. URL: http://media.di-verlag.de/atp/atp_07-08_2014_Niemann.pdf,
- [18] Information technology — Security techniques — Vulnerability disclosure, ISO/IEC 29147:2018(E), ISO - International Standardization Organization; IEC- International Electrotechnical Commission, Oct. 2018.
- [19] Information technology — Security techniques — Vulnerability handling processes, ISO/IEC FDIS 30111:2019(E), ISO - International Standardization Organization; IEC- International Electrotechnical Commission, Jul. 2019.
- [20] First.org, Common Vulnerability Scoring System SIG. URL: <https://www.first.org/cvss/>
- [21] Microsoft, Introduction to Code Signing. URL: [https://docs.microsoft.com/en-us/previous-versions/windows/internet-explorer/ie-developer/platform-apis/ms537361\(v=vs.85\)?redirectedfrom=MSDN](https://docs.microsoft.com/en-us/previous-versions/windows/internet-explorer/ie-developer/platform-apis/ms537361(v=vs.85)?redirectedfrom=MSDN)
- [22] D. Wätjen, Kryptographie: Grundlagen, Algorithmen, Protokolle, 3rd ed. Wiesbaden, Heidelberg: Springer Vieweg, 2018.
- [23] N. S. Mtetwa, P. Tarwireyi, A. M. Abu-Mahfouz, and M. O. Adigun, "Secure Firmware Updates in the Internet of Things: A survey," in 2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC): Proceedings : 21-22 November 2019, Emerald Resort & Casino, Vanderbijlpark, Gauteng, South Africa, Vanderbijlpark, South Africa, 2019, pp. 1–7.
- [24] Informationstechnik – Sicherheitsverfahren – Leitfaden für Informationssicherheitsmaßnahmen (ISO/IEC 27002:2013 einschließlich Cor 1:2014 und Cor 2:2015); Deutsche Fassung EN ISO/IEC 27002:2017, DIN ISO/IEC 27002:2017, DIN Deutsches Institut für Normung e. V, Jun. 2017. URL: <https://www.iso.org/standard/54533.html?browse=tc>
- [25] L. Seltzer, Securing Your Private Keys As Best Practice for Code Signing Certificates. URL: http://publishing.insightnews.us/download/Securing_Your_Private_Keys_as_Best_Practice_for_Code_Signing_Certificates.pdf
- [26] A. Regenscheid, "Platform firmware resiliency guidelines," Gaithersburg, MD, 2018. URL: <https://www.nist.gov/publications/platform-firmware-resiliency-guidelines>

Materialverfolgungsbasierter Softsensor für die Überwachung von Treibhausgasemissionen in der Prozessindustrie

Material Tracking based Soft Sensor for Greenhouse Gas Emission Monitoring in Process Industries

Dr. C. Song, Dr. J. Schlake,

ABB Corporate Research Center, Ladenburg;

Sophie Emmrich, ABB Division Process Industries, Cottbus

Kurzfassung

Die Treibhausgasemissionen (GHG, engl. Greenhouse Gas) sind heutzutage der Hauptfaktor für den Klimawandel, und sie haben in den letzten drei Jahrzehnten massiv zugenommen [1]. Vor allem in den letzten Jahren hat die Zahl der extremen Wetterereignisse dramatisch zugenommen, so dass wir die Treibhausgasemissionen weltweit dringend reduzieren müssen. Aus diesem Grund hat jedes Land eigene Programme für eine kohlenstoffneutrale Regierung angekündigt. Die Messung der Treibhausgasemissionen ist dazu der grundlegende, erste Schritt, bevor ein kohlenstoffneutraler Staat zertifiziert werden kann [2]. Der Industriesektor trägt weltweit mehr als 25 % zu den Treibhausgasemissionen bei, was bedeutet, dass alle produzierenden Unternehmen in naher Zukunft verpflichtet sein werden, eine GHG-Emissionslizenz zu erwerben. Somit wird die Überwachung der Treibhausgasemissionen in allen Produktionsschritten zur nächsten Herausforderung für industrielle Monitoring- und Regelsysteme. Derzeit werden Treibhausgasemissionen lediglich punktuell, z.B. mittels optischer Sensoren auf Satelliten gemessen, so dass die spektroskopische Beobachtung von Treibhausgasen vom Weltraum aus erfolgen kann [3]. Diese Methoden sind in der Regel aufgrund der erforderlichen Hardware sehr teuer, und die Messqualität wird durch verschiedene Faktoren, wie Wetter, Streuung des Lichts in der Atmosphäre, beeinflusst. Andererseits bieten ISO-Normen einen Leitfaden für die (lokale) Berechnung des CO₂-Fußabdrucks. Doch dafür sind feingranulare Prozessinformationen erforderlich. In dieser Studie schlagen wir die Entwicklung eines datengesteuerten Soft-Sensor-Modells vor, das auf einer diskreten Ereignissimulation [4] basiert. Der Materialfluss wird auf der Grundlage von Prozessdaten generiert, so dass der Simulator die Fertigungsereignisse in situ nachbildet und feinkörnige

Prozessinformationen über den Materialfluss abgeleitet werden können. Darüber hinaus können die Treibhausgasemissionen entweder pro Produkteinheit oder pro Prozessschritt geschätzt werden. Um eine hohe Qualität der Simulationsergebnisse zu erreichen, entwickeln wir einen Rahmen für einen materialzentrierten Simulationsworkflow, der auf einer Vorwärtsquantifizierung der Unsicherheit und einem Mechanismus zur Zustandsaktualisierung beruht [4]. Die Methode wird am Beispiel der Wertschöpfungskette im Bergbau erläutert. Auf der Grundlage des auf der diskreten Ereignissimulation basierenden Soft-Sensors werden Treibhausgasemissions Key-Performance-Indicators (KPIs) berechnet und in Form einer Wertstromkarte dargestellt.

Abstract

Nowadays, greenhouse gas (GHG) emissions are the main factor for climate change, and they have been massively increased over the past three decades [1]. Especially in recent years, the number of extreme weather events increases dramatically; we need to urgently slash greenhouse gas emissions globally. Therefore, every country announced their carbon neutral government program, and the GHG emission measurement is the fundamental step before any carbon neutral state can be certified [2]. Moreover, the industry sector contributes more than 25% GHG emissions worldwide, which implies that all manufacturing companies will be obliged to obtain GHS emission license soon. Thus, being able to monitor greenhouse gas emissions among all manufacturing steps is becoming the next challenge for industrial monitoring and control systems. Currently, GHG emissions are measured for example by employing the optic sensors on satellites, such that the spectroscopic observation of GHG can be done from space [3]. Those methods are generally very expensive due to the hardware requirements, and the measurement quality is also influenced by weather, atmospheric light scattering, or other factors. On the other hand, ISO standards provide the guidance for carbon footprint calculation, but it requires fine grained process information. In this study, we propose to build a data-driven soft sensor model, which is based on discrete event simulation [4]. The material flow is generated based on process data, such that the simulator replicates manufacturing events *in situ*, and fine-grained process information about material flow can be derived. Furthermore, the greenhouse gas emissions can be estimated either per unit product or on process steps. In order to obtain high quality of simulation results, we develop a framework of manufacturing material centric simulation workflow, which is based on forward uncertainty quantification (UQ) and State Update mechanism [5]. The method is illustrated using the mining value chain as an example. Based on the discrete event simulation-based

soft sensor, greenhouse gas emission key performance indicators (KPIs) are calculated and presented as a value stream map.

1. Introduction

The global climate change is mainly driven by carbon dioxide (CO₂) emissions [1], the usage of fossil fuel is the primary source. The global carbon emissions from fossil fuels have significantly grown since last century, today's carbon emissions are twice as in 1970s, the trend is still increasing [4]. Despite the renewable energy usage is increasing, the global energy is yet dominated by fossil fuels [6]. Industry sectors' energy mix remained basically unchanged from last two decades, more than 60% energy consumed by industry comes from fossil fuel [7], their activities are still the main contributor to climate change.

The Paris Agreement on climate change is a legally binding agreement, the member countries should pursue efforts towards avoiding global warming. Many countries launched consequently their own targets by establishing policies in various areas, e.g.: carbon pricing, clean energy standard, etc. Estimating and monitoring carbon footprints in different sectors attract significant attention, the focus is placed on the value chain emissions [10], the upstream and downstream activities connected to production operations needs to be taken into consideration for measuring and quantifying the baseline emission.

In process industries, materials are prepared following by manufacturing recipes including batch and continuous operations. In process control system the KPIs are often computed on production units. In contrast, the GHG emission is often calculated per unit of material, which requires rather information on material tracking. ISO standard [9] suggests the GHG emission certificates should be evaluated per balance area, e.g., production plants. The CO₂ emissions about upstream of the supply chain needs to be taken into consideration. For example, GHG of the mineral raw materials must be calculated per ton. Moreover, the optimization of reducing CO₂ emissions for production plants also necessitates the information on fine-grained level, the average emission per plant is not sufficient.

A soft sensor [10] is a numerical model that combines process data and a physical model to predict target quantities, e.g. CO₂ emissions within a system; it can be considered as a digital twin during the operation phase performing indirect measurements. Soft sensors can recapitulate manufacturing operations digitally, especially those operations which cannot be measured by physical sensors. The application areas of soft sensors can be divided according to target quantities, e.g., online prediction or monitoring, and process control.

The discrete event simulation (DES) [4] is widely accepted within industries to model event scheduling of manufacturing system; it is commonly used for virtual inspection of considered system before the implementation. Our suggestion is to deploy discrete event simulation as a soft sensor to integrate physical sensor data and event model with the intention of replicating manufacturing processes digitally. The discrete event modeling can fill the gap between the process centric data, which are collected by a control system, and material tracking information, which is required for carbon footprint estimation. Besides, DES is a framework focusing on scheduling, material tracking demands the chain of information on materials. Thus, we introduce the material blob concept, which is an information container of digital materials travelling through process units. During the simulation, physical models can alter and record material blobs and associated information in order to construct the material flows.

Moreover, DES is a data-driven prediction model, the prediction quality highly depends on input data and the underlying physical model. Uncertainty Quantification [11] is a framework to investigate the correlation of quantities of interest with variations in the input data. The uncertainty propagation can estimate the degree of deviations of model output influenced by uncertain data. Hence, in order to reach the deployment level of such a digital twin, we also consider the uncertainty quantification method to determine and increase the model reliability or at least provide an indicator where to improve the model.

2. Soft Sensor and Discrete Event Modeling

A sensor is device that converts the state of an observed physical quantity into information. For example, a sensor converts temperature into a display on a thermometer device. In general, sensors can be divided into two types: physical sensors and soft sensors. Physical sensors are devices which measure physical quantities, e.g. pressure and speed. Soft sensors, also known as virtual sensors, are a software that can handle multiple measurements at the same time. The soft sensor concept is rooted in control theory and sometimes referred to as observers (or estimator, filter), which are used indirectly for measurement [12].

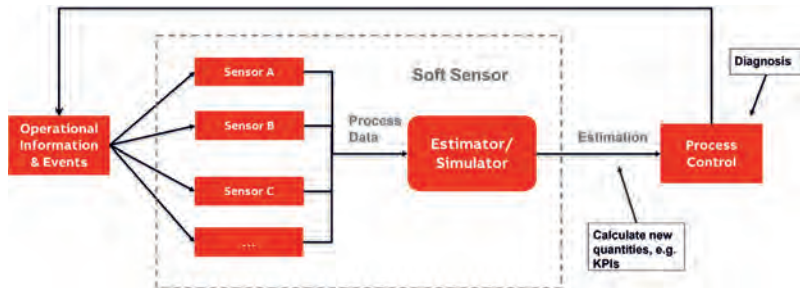


Fig. 1: Basic workflow of soft sensor

Fig. 1 illustrates the basic principle of a soft sensor. The DES-based soft sensor is generally divided into two essential components: the first step is to collect and extract event information from physical sensor data, the second step is to create a simulator, which simulates the process of the considered system. The outcomes of the simulator can be used to evaluate the system states and necessary KPIs, such that a process control system or the operator can use it to diagnose the operation's status and to take decisions accordingly.

Discrete event simulation is commonly used when the system can be described naturally as a sequence of operations to simulate the operation of a system. In a discrete system one can always find a point in time to label changes in system, namely at a timestamp some materials enter and leave the system. Between successive events, it is assumed that no major changes occur in the system, the simulation time can directly jump to the time of occurrence of the next event, between two events considered KPIs can be calculated based on underlying physical models. This concept is also applicable to continuous processes, since one can synthetically discretize the system process [4].

For tracking the flow of materials within a production plant, we decide to utilize the discrete event simulation as estimator in soft sensor framework (Fig. 1). Discrete event modeling enables the data fusion of different instruments and provides the means to investigate complex value chain systems, important aspects can be mapped into sub-model of the simulator, whose behavior can be analyzed in place of the real system of which it represents.

3. Material Blob

Although discrete event simulation can model operations in complex systems, it is still a process model and focuses on the view of event scheduling. We introduce the material blob as auxiliary concept into DES to calculate the material flow. Material blobs can be regarded as

material batches and are information container of material-related information. Due to the nature of the discrete event simulation, material blobs are not describing continuous material but discretized units of materials. Assuming the discrete event model depicts an industrial process properly the material blobs are the equivalent to a product unit in discrete manufacturing allowing to use the same features as in discrete manufacturing.

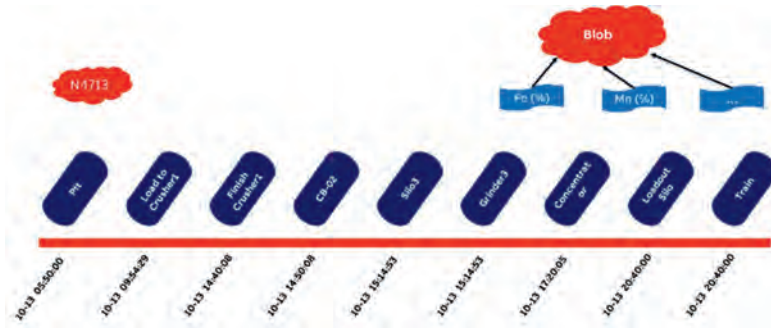


Fig. 2: Material blob in a mining system

4. State Update and Uncertainty Quantification

Fig. 3 illustrates the workflow of deploying DES for a soft sensor system. The process data consist of operational information and event data, which are responsible for defining process units' actions and material re-allocation. DES requires also hyper-parameters for model setting. The outputs are the system states and target KPIs, which can be verified with process data. However, the data quality is one of the key issues for any data-driven simulation, data uncertainty can accumulate in DES and lead to undesired results.

Moreover, DES is built upon a certain level of approximation, missing knowledge about process details may also raise the challenge of inaccurate prediction. We introduce an additional step, State Update, into the discrete event simulation. The State Update is implemented as an extra monitoring system, it inspects the overall system state with a pre-defined frequency. Whenever the State Update is active, the material flow on each process unit is compared with the process data; when the difference exceeds certain threshold the corresponding material blobs are updated accordingly.



Fig. 3: Discrete event simulation workflow and state update process

The State Update mechanism is responsible for easing the uncertainties in process data, especially in event data. However, DES necessitates hyper-parameters, which are generally obtained from field operators or process engineers, hardware specifications or purely simulation settings. Due to the lacking knowledge about the production plant and data uncertainty those parameters can also have a huge influence on the prediction results.

Many applied models inevitably contain uncertainties arising from systematic, observational, and cognitive errors. Calculating and analyzing uncertainties would help to understand cognitive problems and also assist decision making process. Uncertainty Quantification is one of the active research areas today, whose main purpose is to quantify the uncertainty in mathematical modeling and improve the predictive power of numerical results [11].

In order to investigate the influence of the uncertainty parameters of the model, it is common to define a mapping of these parameters on the model outputs. Such a mapping is called a forward uncertainty quantification model [11]. Uncertain parameters or input quantities are represented by probability distributions. The aim is to determine the probability distribution of the solution to the model or the derived variable (the output variable).



Fig. 4: Workflow of forward uncertainty quantification

The forward uncertainty quantification [11] procedure is summarized in Fig. 4. The most common technique is the sampling method, for example Monte Carlo. One generates random samples of input uncertain parameters based on their probability distribution, then run the numerical model with different settings. However, a fully random method like Monte Carlo method often leads to slower convergence. If the behavior of an uncertain parameter is rather homogeneous, then quasi Monte Carlo methods [12] can provide a better convergence rate than the classical Monte Carlo methods. Moreover, if the dimension of the stochastic space is small, e.g., less than 20, Sparse Grid [13] and Stochastic Collocation [14] methods can be applied; the number of samples needed can be reduced exponentially.

5. Value Chain and Greenhouse Gas Emissions

Currently, companies focus only on CO₂ emissions from their own operations, but it is more important to understand the need to also account for GHG emissions along their value chains and product portfolios to comprehensively manage GHG-related risks and opportunities. The value chain (VC) starting with raw materials and ending with the delivery of the products to customers is a set of activities of company operations, all related information, e.g., energy consumption, auxiliary materials, CO₂ emissions, etc., to processed materials needs to be tracked through the value chain [15].

The World Resources Institute (WRI) releases the protocol [16] and guidance for calculating greenhouse gas emissions. The guidance emphasizes on GHG management of companies' value chain. However, the way of calculating GHG per unit material in a production is not clearly defined. For the most manufacturing processes, simulating corresponding material flows within a production plant using soft sensors, e.g., discrete event simulation, can support the implementation of value chain emission monitoring.

The IPCC 2006 Guidelines on how to calculate emissions cover a wide range of greenhouse gases, amongst them Carbon dioxide, Methane, Nitrous oxide, and Sulphur hexafluoride. Since then, researchers have gathered data to come up with models on how to calculate the emissions from various activities. Focus of the efforts is to cover the most relevant sources of emissions first. The results are publicly available in an Emission Factor Database. For practitioners this database [17] can be used to calculate the emissions within a defined balancing area. The IPCC also proposes (depending on the availability of data and the balancing area) activity equations on how to estimate emissions. For a fleet of trucks, the general process is depicted in Fig. 5; similar equations exist for other emission sources as well.

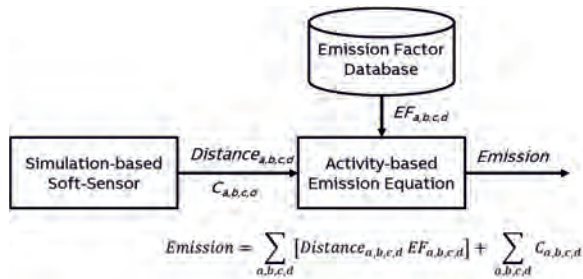


Fig. 5: Activity-based Emission Calculation [18]

The equation itself is using information on the traveled distance, where

- a: fuel type
- b: vehicle type
- c: emission control technology
- d: operating conditions
- $C_{a,b,c,d}$: emissions during warm up phase

If the Distance information is not measured it can be derived via a soft sensor. The framework can be adopted so online estimations are feasible.

6. A Case Study

We use mining process as a case study in this paper, Fig. 6 illustrates the general process units in a mine and corresponding data acquisition. Due to the decentralization of different units and control systems the data management is also different. For example, hauling data is typical event data which track the amount of raw material of each truck dump. Whereas, process data of continuous processing units, e.g. crushers and concentrators, are continuous time series data, which document the operational status. The discrete event simulation plays a role for data fusion in modeling mining process. It integrates different types of field measurement data into numerical models, such that digital states of the mining plant can be generated.

As mentioned earlier the discrete event simulation provide the means of tracking the material flow within a production plant, whereas it requires to cope with process data with a large degree of uncertainties. Hence, building an accurate discrete event simulation model is important for tracking the material flow as well as for calculating KPIs.

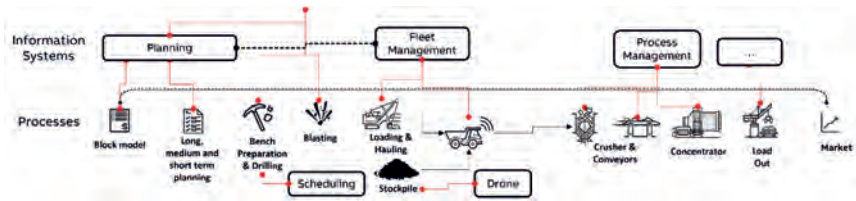


Fig. 6: Process units in general mine and data management

Fig. 7 shows a comparison of two days numerical results from the discrete event simulator and the measurement of the storage level on two different silos in a mine. One observes that the discrepancy between real data and simulated data is significant due to the data quality and data uncertainty issue. Fig. 8 demonstrates the same comparison, whereas State Update mechanism in the discrete event simulator is enabled and is taken place every 4 hours; DES results are more accurate because the accumulated errors are corrected by State Update process. However, some observable difference still exists. And one cannot easily estimate the source of those errors because the data uncertainty and model inaccuracy are mixed on this level. Fig. 9 exhibits the numerical results with forward uncertainty quantification, the gray area pictures the confidence interval defining with two standard deviations envelop, i.e. creating a region, whose lower window is defined by the mean value minus two standard deviations and upper window is defined by the mean value plus two standard deviations, respectively. The forward UQ is chosen as the stochastic collocation method [11], which is a sampling-based method similar to Monte Carlo but with less sampling points for low dimensional uncertain system. By considering the uncertainty in data and model hyperparameters, the measurements can be well contained within the confidence interval, UQ results improve the reliability of our discrete event simulation prediction.

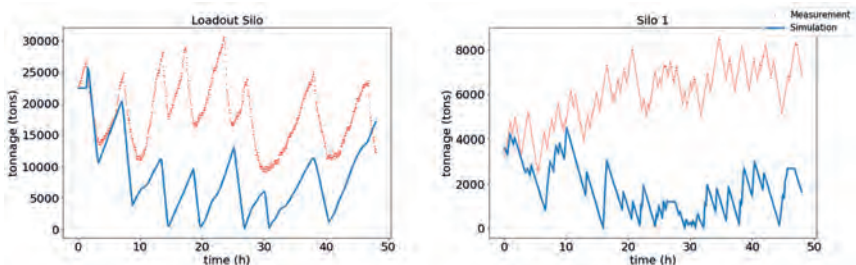


Fig. 7: Comparison of discrete event simulation and measure for storage level on two silos in a mine

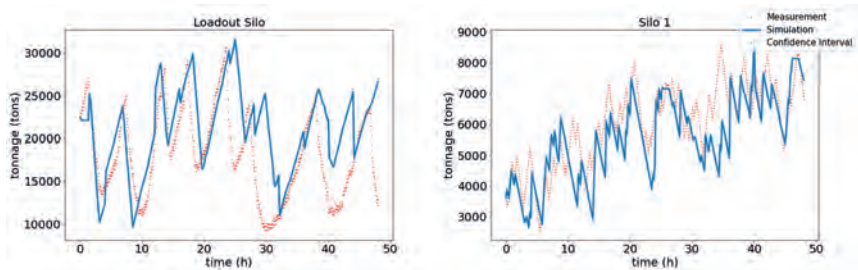


Fig. 8: Same comparison as in Fig. 7 with State Update mechanism for DES

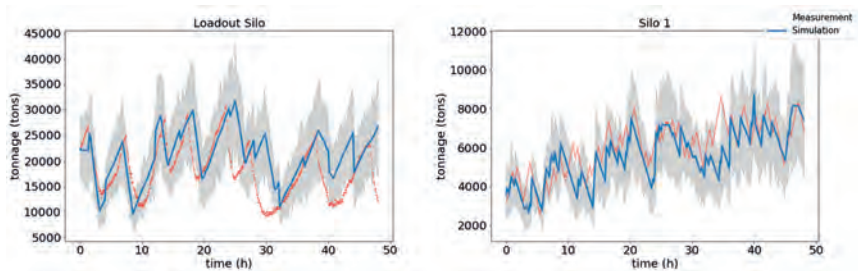


Fig. 9: Same comparison as in Fig. 8 with uncertainty quantification

Fig. 10 shows the value stream mapping [19] for a single material blob (ID:5110) in the considered use case [20]. The numerical results are obtained from a discrete event simulator in combination with process data. The studied mine layout for process units is presented in the middle. On the left, the time consumption on each process unit traversed by selected material blob is displayed. On the bottom the energy consumption and CO₂ emission are added to this material blob on different unit are also tracked.



Fig. 10: Value stream mapping

7. Conclusion

In this work, we propose to apply the discrete event simulation as modeling framework for simulating the industrial processes in production plants. By combining the material blob concept and DES, the modeling framework can not only calculate the manufacturing events scheduling, but also be able to track the material flows with assigned KPIs throughout a system in a flexible way. In addition, the data quality in industry applications is an inevitable issue, but it has very great influence on data-driven models' outputs. Hence, we also discussed the importance of state update and uncertainty quantification, the prediction improvements illustrated with a use case in mining. The primary objective is to obtain a reliable modeling framework, such considered KPIs related to processed materials can be calculated correctly. Regarding the trend of establishing greenhouse gas emissions footprint, our proposed framework can be utilized for calculating value chain emissions. We demonstrated our implementation in mining, the material blobs assigned CO₂ emission information as well as other data can be visualized across different process units.

Reference

- [1] "Causes of climate change." https://ec.europa.eu/clima/climate-change/causes-climate-change_en (accessed Apr. 26, 2022).
- [2] M. E. Sotos, "GHG Protocol Scope 2 Guidance," 2015.
- [3] "ABB-built sensor improves methane emission monitoring from space." <https://new.abb.com/news/detail/73664/abb-built-sensor-improves-methane-emission-monitoring-from-space> (accessed Nov. 10, 2021).
- [4] A. Soofastaei, S. M. Aminossadati, M. S. Kizil, and P. Knights, "A discrete-event model to simulate the effect of truck bunching due to payload variance on cycle time, hauled mine materials and fuel consumption," *International Journal of Mining Science and Technology*, vol. 26, no. 5, pp. 745–752, Sep. 2016, doi: 10.1016/J.IJMST.2016.05.047.
- [5] G. Lugaresi and A. Matta, "Generation and Tuning of Discrete Event Simulation Models for Manufacturing Applications," *Proceedings - Winter Simulation Conference*, vol. 2020-December, pp. 2707–2718, Dec. 2020, doi: 10.1109/WSC48552.2020.9383870.
- [6] "Global Energy Review: CO2 Emissions in 2021 – Analysis - IEA." <https://www.iea.org/reports/global-energy-review-co2-emissions-in-2021-2> (accessed Apr. 20, 2022).
- [7] P. Levi, T. Vass, H. Mandová, and A. Gouy, "Tracking Industry 2020," 2020.
- [8] G. Shanahan, "Managing government regulation for carbon reporting and action," *Aarti Krishnan and Simon Maxwell are alone responsible for errors and omissions.*, p. 63, 2020.
- [9] "DIN EN ISO 14064-1 - 2019-06 - Beuth.de." <https://www.beuth.de/en/standard/din-en-iso-14064-1/291289049> (accessed Apr. 25, 2022).
- [10] "Soft Sensors for Monitoring and Control of Industrial Processes," *Soft Sensors for Monitoring and Control of Industrial Processes*, 2007, doi: 10.1007/978-1-84628-480-9.
- [11] R. Ghanem, H. Owhadi, and D. Higdon, "Handbook of uncertainty quantification," *Handbook of Uncertainty Quantification*, pp. 1–2053, Jun. 2017, doi: 10.1007/978-3-319-12385-1.

- [12] W. J. Morokoff and R. E. Caflisch, "Quasi-Monte Carlo Integration," *Journal of Computational Physics*, vol. 122, no. 2, pp. 218–230, Dec. 1995, doi: 10.1006/JCPH.1995.1209.
- [13] J. Garcke and M. Griebel, Eds., "Sparse Grids and Applications," vol. 88, 2013, doi: 10.1007/978-3-642-31703-3.
- [14] D. Xiu, "Stochastic Collocation Methods: A Survey," *Handbook of Uncertainty Quantification*, pp. 1–18, 2015, doi: 10.1007/978-3-319-11259-6_26-1.
- [15] V. C. M. PORTER'S, "What Is Value Chain," 1985.
- [16] W. R. I. Wbcsd, "The greenhouse gas protocol," *A corporate accounting and reporting standard, Rev. ed. Washington, DC, Conches-Geneva*, 2004.
- [17] "CGE Workshops | UNFCCC." <https://unfccc.int/process/bodies/constituted-bodies/consultative-group-of-experts/calendar-of-meetings/archive-of-meetings> (accessed Apr. 28, 2022).
- [18] H. S. Eggleston, L. Buendia, K. Miwa, T. Ngara, and K. Tanabe, "2006 IPCC guidelines for national greenhouse gas inventories," 2006.
- [19] T. Manos, "Value stream mapping-an introduction," *Quality Progress*, vol. 39, no. 6, pp. 64–69, 2006.
- [20] S. Bhadauria, "LEAN DIGITAL TWIN FRAMEWORK: An integration of Lean Tools with Digital Twin Architecture," Hochschule Ansbach, 2021.

Anforderungen aus der dynamischen Nutzung von grünem Wasserstoff in Power-to-X-Wertschöpfungsketten

Dipl.-Ing. **I. Viedt**, **M. Mock**, Dipl.-Ing. **J. Lorenz**, Dipl.-Ing. **J. Mädler**,
Prof. Dr.-Ing. **L. Urbas**,
Professur für Prozessleittechnik und Arbeitsgruppe
Systemverfahrenstechnik, Technische Universität Dresden;
Dipl.-Ing. **A. Klose**, M.Eng. **H. Lange**,
Process-to-Order-Lab, Technische Universität Dresden;
M.Sc. **D. Erdmann**, Verband Deutscher Maschinen- und Anlagenbau
(VDMA), Frankfurt am Main;
M.Sc. **L. Scholz**, M.Sc. **A. Markaj**, Prof. Dr.-Ing. **A. Fay**,
Helmut-Schmidt-Universität / Universität der Bundeswehr Hamburg,
Hamburg

Kurzfassung

Im Kontext des globalen Klimawandels wird die Nutzung von nachhaltigen Alternativen fossiler Energieträger, wie z.B. grüner Wasserstoff, immer relevanter. In Power-to-X-Wertschöpfungsketten können hierbei die benötigten Chemikalien anhand von Prozessketten auf Basis von grünem Wasserstoff bereitgestellt werden. Die Produktion von Wasserstoff kann unter Berücksichtigung der benötigten Energie und der verschiedenen Wertschöpfungsketten optimiert werden. Dazu muss der lastabhängige Energieverbrauch und der pfadspezifische Wasserstoffbedarf der nachfolgenden Prozesse berücksichtigt werden, was wiederum in Prozessführungskonzepten für Elektrolyseure resultiert.

In diesem Beitrag werden zwei relevante Power-to-X-Wertschöpfungsketten - Herstellung von Ammoniak aus grünem Wasserstoff und die Nutzung von grünem Wasserstoff bei der Stahlproduktion - dargestellt und hinsichtlich ihrer charakteristischen Dynamik und ihres Lastfahrverhaltens analysiert. Aus der Analyse der relevanten Abnahmepfade werden spezifische Anforderungen an die Prozessführung der eingesetzten Elektrolysetechnologien abgeleitet. Diese Anforderungen bilden die Basis für die Simulation und den Entwurf von Prozessführungskonzepten.

1. Einsatz von Wasserstoff in der Prozessindustrie

Die Industrie ist mit einem Anteil von 38 % an der gesamten Endenergienachfrage der größte Endverbrauchssektor und ist für 26 % der weltweiten CO₂-Emissionen des Energiesystems

verantwortlich [1]. Aktuell basieren die meisten relevanten Wertschöpfungsketten in der chemischen Prozessindustrie auf nicht-erneuerbaren Rohstoffen wie Erdgas, Kohle und Erdöl. Aufgrund der Endlichkeit dieser Rohstoffbasis und den entstehenden CO₂-Emissionen werden alternative, kurzfristig möglicherweise teurere Produktionspfade betrachtet. Hierbei sind Power-to-X-Wertschöpfungsketten ein relevanter Lösungsansatz für die Defossilisierung der Prozessindustrie [2]. Im Zuge dieses Wandels in der Prozessindustrie (incl. Eisen und Stahl) prognostiziert die OECD bis 2050 eine Steigerung des industriellen Wasserstoffbedarfs von 51 Mt auf 130 Mt Wasserstoff (siehe Abbildung 1) für das Netto-Null-Emissionsszenario [1]. Neben den in Projekten bereits untersuchten Anwendungen von Wasserstoff z.B. als Zusatz für Reduktionsmittel werden neue Anwendungsbereiche in ähnlicher Größenordnung hinzukommen (Abbildung 1). Das Projekt eModule als Teil des Wasserstoffleitprojekts H₂Giga untersucht den Beitrag von Modularisierung und Steuerung zum effizienten Betrieb von Elektrolyseanlagen. Betrachtet werden dabei zwei Szenarien: zum einen die mittelfristige Anpassung der Anlagen an die Nachfrage, zum anderen die kurzfristige Eignung als Bindeglied zwischen einer volatilen nachhaltigen Stromversorgung [3] und den auf grünem Wasserstoff basierenden PtX-Wertschöpfungsketten [3].

Um die Anforderungen an die Prozessführung von Elektrolyseanlagen aus Sicht der Wasserstoffabnahmepfade zu erheben, wird in Abschnitt 2 zunächst die Relevanz ausgewählter Wertschöpfungsketten betrachtet und anschließend diese Verfahren hinsichtlich ihrer kommerziellen Anwendung charakterisiert. In Abschnitt 3 werden dann die zwei ausgewählten Verfahren, die Ammoniaksynthese und das Direktreduktionsverfahren von Eisenerz, hinsichtlich ihrer charakteristischen Dynamik und der Prozessführung analysiert und die sich daraus ergebenden Anforderungen an die vorgeschaltete Elektrolyse abgeleitet. Diese Anforderungen dienen später im Projekt eModule als Ausgangspunkt für die Simulation und Validierung der entwickelten Konzepte.

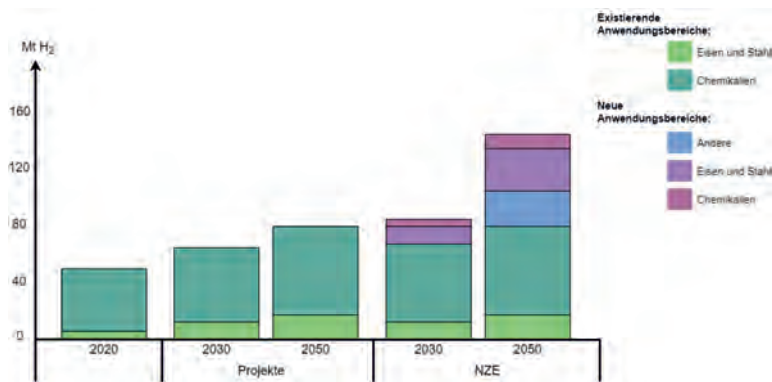


Bild 1: Prognostizierte Veränderung des totalen Wasserstoffbedarfs von 2020 zu 2030 und 2050 für aktuell geplante Wasserstoffprojekte und Netto-Null-Emissionsszenario (NZE - Net Zero Emission) (vgl. [1])

2. Power-to-X-Wertschöpfungsketten

Der Begriff Power-to-X (PtX) beinhaltet zahlreiche Verfahren und wird in der Literatur auf verschiedenste Weise definiert [2, 4, 5, 6]. Die von Berger et al. [7] aufgestellte umfassende Definition von PtX-Technologien lautet:

„Unter Power-to-X-Technologien werden alle Technologien beziehungsweise Verfahren mit neuen oder neu eingeführten Optionen der Stromnutzung verstanden, in denen Strom aus einem variablen Stromangebot auf Basis erneuerbarer Energien in Verfahren zur Herstellung von Wertschöpfungsprodukten oder zur Bereitstellung von Dienstleistungen direkt oder indirekt genutzt wird.“ [7]

Im Allgemeinen umfassen Power-to-X-Wertschöpfungsketten eine Vielzahl von verschiedenen Verfahren, die ein breites Spektrum an Anwendungsfällen abdecken. Je nach Produkt oder Leistung werden in der Fachliteratur Kategorien wie Power-to-Power, Power-to-Gas, Power-to-Fuel, Power-to-Chemicals und Power-to-Heat [8] verwendet.

Im Rahmen dieses Artikels werden zwei ausgewählte PtX-Wertschöpfungsketten betrachtet werden und dienen als Basis für die Anforderungsanalyse. Zur Auswahl der zu betrachtenden Wertschöpfungsketten wird der aktuelle (Abbildung 2) und zukünftige Wasserstoffbedarf zur chemischen Weiterproduktion herangezogen. Die Verteilung des Wasserstoffbedarfs in der Prozessindustrie für das Jahr 2020 ist in Abbildung 2 dargestellt [1].

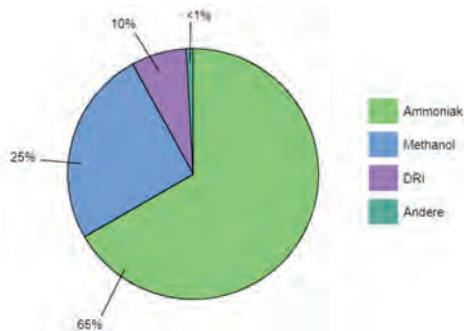


Bild 2: Verteilung des Wasserstoffbedarfs in der Prozessindustrie in 2020 nach [1]

Mit einem Anteil von 65% hat die Ammoniaksynthese den höchsten relativen Wasserstoffbedarf an produziertem Wasserstoff [1] und qualifiziert sich damit als erster Kandidat für eine detaillierte Betrachtung. Um die Übertragbarkeit der Methodik zu explorieren, ist ein weiterer, grundsätzlich anderer Abnahmepfad in einem anderen Industriezweig sinnvoll. Ausgewählt wurde die direkte Nutzung von Wasserstoff (Power-to-Wasserstoff) zur Direktreduktion von Eisenerz (DRI).

2.1 Power-to-Ammoniak (PtA)

Der Power-to-Ammoniak-Pfad (PtA) hat hohe Priorität, da Ammoniak in den verschiedensten Anwendungen benötigt wird und Grundlage für die Produktion von Düngemittel, Sprengstoff und verschiedensten Spezialmaterialien ist. Ammoniak kann weiterhin sowohl für die verlustarme Zwischenspeicherung von Überschussenergie sowie für den Betrieb von Brennstoffzellen genutzt werden [5, 9].

2.1.1 Ammoniaksynthese

Für die PtA-Wertschöpfungskette steht der Haber-Bosch-Prozess als konventioneller und kommerziell eingesetzter Verfahrensweg im Fokus [10]. Der grundsätzliche Ablauf sowie die typischen Betriebsparameter in Großanlagen haben sich dabei seit Beginn der großtechnischen Produktion im frühen 20. Jahrhundert kaum verändert [10]. Konventionelle Anlagen nutzen Wasserstoff auf Basis von Erdgas, Kohle und Öl [1, 11]. Im Kontext von PtX soll jedoch

die Produktion aus grünem Wasserstoff durch Elektrolyse betrachtet werden. Die grundlegenden Prozessschritte der PtA-Wertschöpfungskette sind in Abbildung 3 dargestellt.

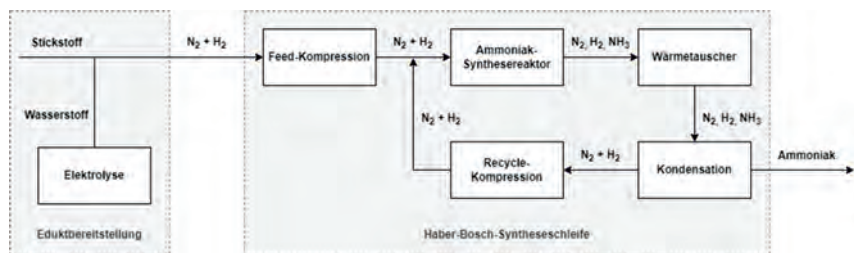


Bild 3: Verfahrensschema Haber-Bosch-Prozess zur Ammoniaksynthese nach [12]

Beim Haber-Bosch-Verfahren reagieren Wasserstoff und Stickstoff bei hohem Druck und hohen Temperaturen über eine katalytische Reaktion an einem Eisenkatalysator zu Ammoniak (siehe Gleichung 1).



Typische Betriebsparameter konventioneller Anlagen liegen bei 250 – 300 bar und 450 – 550°C [12, 11]. Durch die Nutzung neuer Ruthenium-basierter Katalysatormaterialien, besonders in kleinskaligen Anlagen, lässt sich die Reaktion auch unter milderen Reaktions- und Trennungsbedingungen durchführen [13]. Der Energiebedarf einer großtechnischen PtA-Anlage beträgt etwa 42 GJ pro Tonne produziertem Ammoniak [12].

2.1.2 Prozessführung des Haber-Bosch-Prozesses

Die Ammoniaksynthese im Haber-Bosch-Verfahren ist eine Gleichgewichtsreaktion, die Verlagerung des Gleichgewichts zur Ammoniakseite wird durch hohen Druck und niedrige Temperaturen begünstigt wird [12, 14]. Dies steht im Gegensatz zur Reaktionskinetik, für die hohe Temperaturen notwendig sind [12]. Der Gleichgewichtsumsatz von Ammoniak über konventionellen Eisenkatalysatoren liegt bei 500 bar und 350°C bei maximal 60% [10], dennoch wird üblicherweise lediglich ein Umsatz von 20 – 40% erreicht [12]. Daher ist das Abtrennen von Ammoniak und eine Rückführung der unverbrauchten Edukte, also Wasserstoff und Stickstoff, notwendig, um die Prozessausbeute zu erhöhen.

Die Reaktion ist stark exotherm, weshalb eine Kühlung zur Abfuhr der entstehenden Wärme notwendig ist [15]. Dabei sind mehrere Wärmetauscher im System sinnvoll. Diese befinden sich üblicherweise vor dem Reaktor, direkt im Reaktor und vor der Ammoniakseparation [9].

Um ausreichenden Prozessdruck einzustellen, wird eine Kompressorschaltung für die Eingangsströme ins System und eine Kompressorschaltung für die rückgeführten Edukte eingesetzt [9]. Die Prozessführung beinhaltet daher eine Temperaturkontrolle und -regelung zur Anpassung der abgeführten Wärme durch das Kühlmittel in den Wärmetauschern. Weiterhin findet eine Druck- sowie Durchflussüberwachung und -regelung durch die Anpassung der Kompressoren statt [13]. Die konventionelle Ammoniaksynthese muss kontinuierlich betrieben werden. Grund dafür ist, dass sowohl Kaltstarts als auch Warmstarts bei den Hochdruck- und Hochtemperaturverfahren sehr teuer und zeitaufwändig sind. Zudem führen häufige Druck- und Temperaturschwankungen zu einer deutlichen Verkürzung der Lebensdauer wesentlicher Anlagenbestandteile [10]. Die Syntheseschleife mit mehreren Reaktoren, Wärmetauschern, Flash-Drums und Kompressoren macht den dynamischen Betrieb überaus komplex [12]. Die Ramp-Up-Rate des Haber-Bosch-Prozesses liegt bei 20% des zu erzielenden Nennwertes pro Stunde, wobei das korrekte Einstellen des H_2/N_2 -Verhältnis den zweitgrößten Einfluss auf die Performance während des Anfahrens hat [15].

2.1.3 Verhalten bei schwankenden Eingangsbedingungen der Ammoniaksynthese

Um den kontinuierlichen Betrieb der Ammoniaksynthese auch bei schwankenden Eingangsströmen sicherzustellen, ist ein variierender Lastbetrieb notwendig. Im Fall auftretender Lastschwankungen existieren unterschiedliche Prozessführungsstrategien wie die Steuerung durch Druckvariation, Parallelisierung und Variation des Inertgasanteils [13]. Die Lastregelung durch Druckvariation reguliert die Leistung des Recyclingkompressors. Dies wird in der Praxis jedoch nicht angewandt, da die Druckvariation eine höhere Belastung des Equipments nach sich zieht und die Lebensdauer der Anlage verkürzt [13]. Die Steuerung durch Parallelisierung setzt einen ansatzweise modularen Aufbau voraus, bei dem ausgewählte Anlagenbestandteile mehrfach parallel gefahren werden. Dabei können, ähnlich der Lastkontrolle von Stacks innerhalb eines Elektrolysesystems, einzelne Blocks zu- oder abgeschaltet werden. Dies steigert die Anforderungen an die Prozessführung durch die Leitsystemebene und zusätzlich müssen das jeweilige Ramp-Up der Teilanlagen berücksichtigt werden, wodurch höhere Investitionskosten entstehen. Zudem kommt es durch den daraus entstehenden diskontinuierlichen Betrieb zur Schädigung der Anlage [10, 13]. Daher hat sich in der Praxis das Lastmanagement durch die Anpassung des Inertgasanteils durchgesetzt. Die Inertgasfraktion des Eduktstroms wird durch veränderte Spülströme mit gezielter Zugabe von Argon oder erhöhten Stickstoffzugabe angepasst was den Umsatz im Reaktor senkt und die Leistungsaufnahme der energieintensiven Anlagenbestandteile wie Kühlung und Separation des Ammoniaks reduziert. Der durch die Modifikation des Inertgasanteils realisierbare Teillastbetrieb kann technisch bis 0%

reduziert werden, unterhalb eines Lastbetriebs von 20% ist die Produktion allerdings ökonomisch nicht mehr haltbar [10]. Ohne einen stabilen Betriebspunkt kommt es jedoch zum Abbruch der Reaktion [13]. Der Lastbereich liegt daher bei 20 – 100%. Eine dynamische Lastvariation ist innerhalb von Stunden möglich (Zeitkonstante von 2h), wobei nicht beliebig an- und abgefahren werden kann, da eine Änderung der Ammoniakproduktion nur einmal alle vier Stunden erfolgen kann [17]. Elektrolyse (H_2) und Luftzerlegung (N_2) reagieren schnell im Bereich weniger als 1h auf Lastwechsel. Die Ammoniaksynthese ist daher im PtA-Verfahren (siehe Abbildung 3) der entscheidende Schritt für die Laständerungsgeschwindigkeit ist [10, 13].

2.2 Power-to-Wasserstoff für die Stahlherstellung

Auf den Eisen- und Stahlsektor entfallen heute 10% des gesamten Wasserstoffbedarfs in der Industrie [1]. Für eine Senkung der CO_2 -Emissionen kann der Anteil von Wasserstoff im konventionellen Hochofen kurzfristig gesteigert werden, für eine langfristige Defossilisierung des Sektors ist jedoch eine grundsätzliche Umstellung der Stahlindustrie notwendig [16].

2.2.1 Direktreduktion von Eisenerz mittels Wasserstoff

Ein Lösungsansatz für die höhere Nachhaltigkeit ist der Übergang zur Direktreduktion [1]. In der konventionellen Stahlproduktion wird Eisenerz mittels Kohlenstoffmonoxids unter Bildung von Kohlendioxid zu Roheisen reduziert (Gleichung 2).



Alternativ kann für eine nachhaltige Produktion von Roheisen anstatt Kohlenstoffmonoxid im Direktreduktionsverfahren auch reiner Wasserstoff eingesetzt werden (Gleichung 3).



Wasserstoff wird bereits in der konventionellen Stahlproduktion als Reduktionsmittel, meist als CO/H_2 -Synthesegas mit Wasserstoffanteilen von 0 - 70% eingesetzt [1]. Für die kommerzielle Umsetzung des Direktreduktionsverfahrens (DR-Anlagen) sind erhebliche Anpassung der bestehenden Anlagen notwendig [16]. DR-Anlagen produzieren im Gegensatz zu konventionellen Hochofen kein flüssiges Roheisen, sondern festen Eisenschwamm (DRI - Direct Reduced Iron). Damit der Eisenschwamm zu hochwertigem Stahl weiterverarbeitet werden kann, muss

er in einem weiteren Prozess, beispielsweise in einem Elektrolichtbogenofen zu einem roheisenähnlichen Produkt eingeschmolzen werden (Abbildung 4) [6].

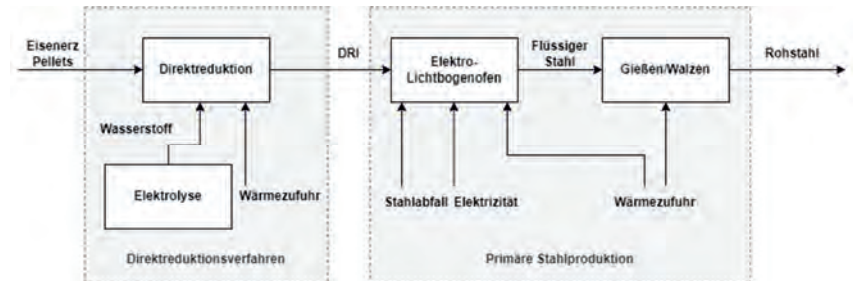


Bild 4: Schematische Darstellung der Herstellung von Rohstahl mittels Direktreduktionsverfahrens und 100% Wasserstoff nach [18]

2.2.2 Prozessführung der Eisen- und Stahlherstellung

Anders als die Ammoniaksynthese via Haber-Bosch-Verfahren kann die Direktreduktion von Eisenerz bisher nicht flexibel betrieben werden und unterliegt einer rigiden Prozessführung [9]. Die Direktreduktion, kombiniert mit einem Elektrolichtbogenofen, erzeugt kontinuierlich ein flüssiges Produkt. Der Schachtofen in dem die Umwandlung stattfindet, wird mit den Rohstoffen kontinuierlich von oben befüllt und flüssiges Eisen wird unten abgezogen. Die Anlagen können im niederen Druckbereich bei 6-8 bar betrieben werden. Da es sich bei der Direktreduktion um einen endothermen Prozess handelt, wird der Wasserstoffstrom mit etwa 900°C mittig eingelassen, während meist vorgeheizte und vorgereinigte Eisenerz-Pellets mit 800°C von oben in den Schachtofen eingeführt werden [19, 16]. Um eine Verbesserung der Reaktionskinetik zu erreichen, wird der Wasserstoff überstöchiometrisch dem Prozess zugeführt [19]. Weiterhin muss aufgrund der endothermen Reaktionseigenschaften der Direktreduktion für eine ausreichende Wärmezufuhr gesorgt werden [19]. Eine Abschätzung der Reaktionsgeschwindigkeit während des Prozesses ist nur bedingt möglich [5]. Die Verweilzeit der Erze während der Direktreduktion beträgt etwa acht Stunden, wobei 2/3 (ca. 5,25h) für die Reaktion zu festem Eisenschwamm im oberen Bereich benötigt werden und 1/3 der Zeit (ca. 2,75h) für die Abkühlung unterhalb der Wasserstoffzufuhr aufgebracht werden müssen [5]. Aus der Verweilzeit lassen sich auch grob die Zeiten für Anfahren und Abfahren abschätzen. Die benötigte Wasserstoffmenge für die Direktreduktion des Eisenerzes hängt von Prozesskinetik, Thermodynamik sowie Eigenschaften des Eisenerzes und des Wärmetransports ab [16]. Durchschnittlich werden etwa 40kg Wasserstoff pro Tonne Eisenschwamm benötigt [1, 5].

3. Anforderungen anhand der Prozessführung der Abnahmepfade

Anhand der dargestellten Charakteristika der beiden Abnahmepfade in Bezug auf Prozessführung und Dynamik werden im Folgenden Anforderungen an die vorgeschaltete Elektrolyse abgeleitet werden. Hierbei wird insbesondere auf das Anfahrverhalten und die kontrollierte Produktion im Abnahmepfad eingegangen werden.

Für das PtA-Verfahren gilt, dass die Reaktion wirtschaftlich sinnvoll bei einer Teillast von 20 - 100% betrieben werden kann [10]. Der über den Tagesverlauf fluktuierend anfallende Wasserstoff aus erneuerbaren Energiequellen wie beispielsweise Photovoltaik kann somit direkt in der PtA-Wertschöpfungskette verwendet werden. Als Anforderung an die vorgeschaltete Elektrolyse ergibt sich je nach PtA-Anlage, dass eine Produktionsmenge in Form von Grundlast von bis zu 600t Wasserstoff pro Tag bereitgestellt werden muss [20], um heute typische Abnahmemengen bei einem Bedarf von 180kg H₂ pro produzierter Tonne Ammoniak abzudecken [1, 20]. Damit eine direkte Kopplung von Elektrolyse und PtA-Prozess möglich wird, muss die Anfahrcharakteristik identisch sein, und somit auch die vorgeschaltete Elektrolysetechnik einen Ramp-Up von ca. 20% Wasserstoff pro Stunde abbilden können. Aus der Charakteristik des DR-Verfahrens für die Stahlproduktion ergibt sich für die Elektrolyse, dass eine feste Wasserstoffmenge während der Produktionsphase über einen Zeitraum von mindestens acht Stunden bereitgestellt werden muss. Um das durchschnittliche Produktionsvolumen einer DR-Anlage abzudecken, muss die vorgeschaltete Elektrolyse mindestens einem Bedarf von 40 kg H₂ pro produzierte Tonne Eisenschwamm bereitstellen können [1, 20]. Da von den beiden betrachteten Power-to-X-Wertschöpfungsketten nur der PtA-Pfad eine flexible Betriebsfähigkeit aufweist, ist die Gesamtprozessflexibilität der Abnahme- bzw. Nachfrageseite stark eingeschränkt in Bezug auf die Stahlproduktion. Um diese Diskrepanz zu überbrücken sind Speichersysteme für die Pufferung und Bereitstellung notwendig, die eine kontinuierliche Wasserstoffbereitstellung im geforderten Rahmen ermöglichen. Für einen wirtschaftlich sinnvollen Betrieb und den möglichen industriellen Einsatz ist es daher notwendig, Auslastungsfaktoren, Speicher- und Bedarfscharakteristika in Einklang zu bringen [11].

Insgesamt ist die stärkste Beschränkung für einen an die Fluktuation nachhaltiger Energien angepassten Betrieb von Wasserelektrolyseanlagen die begrenzte Flexibilität der nachgeschalteten Syntheseschritte in den PtX-Ketten. Die wetterbedingte wechselnde Verfügbarkeit von erneuerbaren Energien (EE) aus Wind und Sonne spielen hierbei für die Verfügbarkeit des grünen Wasserstoffs eine entscheidende Rolle für den Betrieb der Elektrolyseanlagen. Es entsteht also ein Optimierungsproblem, welches die Nachfrage von H₂ und Verfügbarkeit von EE mit den jeweiligen Lastbereichen und der Dynamik der Prozesse in Einklang bringen muss.

Eine konstante Nachfrage kann zum einen mit ausreichend dimensionierten Wasserstoffspeichern, zum anderen mit einem Grundlastbetrieb der Elektrolyse beantwortet werden. Letzterer erlaubt immer noch ein gewisses Maß an Flexibilität, d.h. ein Hoch- oder Herunterfahren je nach den auftretenden Netzanforderungen und entsprechend den Anforderungen aus den nachgeschalteten Syntheseschritten. Für die notwendige zuverlässige Bereitstellung von H_2 für den DR-Prozess sollte die Elektrolyse und der DR-Prozess aber für eine minimale Verfügbarkeit von EE ausgelegt sein. Spitzen in der EE-Verfügbarkeit gehen damit verloren.

4. Zusammenfassung und Ausblick

Die Analyse der Randbedingungen für die bedarfsgesteuerte Elektrolyse zeigt die Relevanz von Power-to-X-Wertschöpfungsketten für die nachhaltige Transformation der Prozessindustrie auf. Um diese Transformation zu ermöglichen, ist die Produktion von grünem Wasserstoff auf Basis von Elektrolyse ein wichtiger Schritt. Für die zwei verschiedenen Nutzungsszenarien für grünen Wasserstoff wurden die aktuellen Prozessführungsstrategien untersucht und Anforderungen an die Elektrolyse aufgestellt. Für die Ammoniaksynthese im Rahmen des Power-to-Ammoniak-Verfahrens lässt sich ableiten, dass aufgrund der flexibel möglichen Prozessführung eine entsprechend des Stromangebots fluktuierende Wasserstoffproduktion ohne Pufferspeicher umsetzbar ist. Die zweite am Beispiel der Direktreduktion untersuchte Prozessklasse von PtX-Anlagen benötigt hingegen eine konstante Bereitstellung von Wasserstoff [4]. Für die effiziente Produktion sind daher entweder eine ausreichende Speicherkapazität oder technische Innovationen der Prozesse notwendig, die eine dynamischere Prozessführung erlauben [14]. Für eine Abschätzung bietet sich an die Auswirkungen eines volatilen Wasserstoffangebots auf die Produktivität, als Kombination von Effizienz und Verfügbarkeit für verschiedene Technologievarianten simulationsbasiert intensiver zu untersuchen. Weiterhin müssen für die stoffliche Nutzung auch andere relevante Wertschöpfungsketten wie Power-to-Methanol untersucht werden, die 25% des derzeitigen H_2 -Bedarfs ausmacht, oder Verfahren wie Power-to-Methan, die für das Netto-Null-Emissionsszenario bis 2050 relevant sind (Abbildung 1). Für das ambitionierte Ziel von Netto-Null-Emissionen werden jedoch technologie- und prozessorientierte Analysen wie die hier vorgestellte nicht ausreichen. Vielmehr werden multiperspektivische Ansätze notwendig sein, die zudem auch das Verhalten der Märkte und der Marktteilnehmer und eine im Sinne des o.g. Ziels optimale Verteilung der nach wie vor beschränkten Ressource grüner Wasserstoff zwischen thermischen und chemischen Wertschöpfungsnetzen adressieren.

Danksagung

Dem Bundesministerium für Bildung und Forschung sowie dem Projektträger Jülich sei für die finanzielle Unterstützung im Rahmen des Forschungsprojektes eModule (FKZ 03HY116A-E) der Leitplattform H₂Giga gedankt.

Quellen

- [1] International Energy Agency. Global Hydrogen Review 2021. 2021, S. 223. DOI: <https://doi.org/https://doi.org/10.1787/39351842-en>. URL: <https://www.oecd-ilibrary.org/content/publication/39351842-en>
- [2] Hermesmann, M., Grübel, K., Scherotzki, L., & Müller, T. E. (2021). Promising pathways: The geographic and energetic potential of power-to-x technologies based on regeneratively obtained hydrogen. *Renewable and Sustainable Energy Reviews*, 138, 110644.
- [3] Scholz, L., Markaj, A., Fay, A., Erdmann, D., Beisswenger, L., Lorenzen, L., Urbas, L., Bittorf, L. (2022). Anforderungen an modulare Elektrolyseanlagen. Erkenntnisse aus dem H₂Giga Projekt eModule. *atp magazin*. (in print)
- [4] Bünger, U., Michalski, J., Schmidt, P., & Weindorf, W. (2017). Wasserstoff–Schlüsselement von Power-to-X. In Wasserstoff und Brennstoffzelle (pp. 327-368). Springer Vieweg, Berlin, Heidelberg.
- [5] Domnick, C. (2019). Vergleichende Untersuchung und Bewertung verschiedener wasserstoffbasierter Direktreduktionsverfahren zur Senkung der Treibhausgasemissionen in der Stahlproduktion. Hamburg: Hochschule für angewandte Wissenschaften Hamburg.
- [6] Tang, J., Chu, M.-s., Li, F., Feng, C., Liu, Z.-g., & Zhou, Y.-s. (2020). Development and progress on hydrogen metallurgy. *International Journal of Minerals, Metallurgy and Materials*, S. 713-723.
- [7] Berger, A., Bluhm, H., Ehrenstein, U., Heinbach, K., Dunkelberg, E., & Hirschl, B. (2020). Systematik im Power-to-X-Ansatz–Identifikation, Charakterisierung und Clusterung der Power-to-X-Technologien. *Zeitschrift für Energiewirtschaft*, 44(3), 177-193.
- [8] Chehade, Z., Mansilla, C., Lucchese, P., Hilliard, S., & Proost, J. (2019). Review and analysis of demonstration projects on power-to-X pathways in the world. *International Journal of Hydrogen Energy*, 44(51), 27637-27655.
- [8] Cholewa, T., Semmel, M., Mantei, F., Güttel, R., & Salem, O. (2022). Process Intensification Strategies for Power-to-X Technologies. *ChemEngineering*, 6(1), 13.
- [9] Bañares-Alcántara, R., Dericks III, G., Fiaschetti, M., Grünewald, P., Lopez, J. M., Tsang, E., ... & Zhao, S. (2015). Analysis of islanded ammonia-based energy storage systems. University of Oxford.

- [10] Schulte Beerbühl, S., Fröhling, M., & Schultmann, F. (2015). Combined scheduling and capacity planning of electricity-based ammonia production to integrate renewable energies. *European Journal of Operational Research*, S. 851-862.
- [11] Nazir, H., Muthuswamy, N., Louis, C., Jose, S., Prakash, J., Buan, M. E., ... & Kannan, A. M. (2020). Is the H₂ economy realizable in the foreseeable future? Part III: H₂ usage technologies, applications, and challenges and opportunities. *International journal of hydrogen energy*, 45(53), 28217-28239.
- [12] Kelley, M. T., Do, T. T., & Baldea, M. (2022). Evaluating the demand response potential of ammonia plants. *AIChE Journal*, S. e17552.
- [13] Rouwenhorst, K., Van der Ham, A., Mul, G., & Kersten, S. (2019). Islanded ammonia power systems: Technology review & conceptual process design. *Renewable and Sustainable Energy Reviews*, S. 198338.
- [14] Cholewa, T., Semmel, M., Mantei, F., Güttel, R., & Salem, O. (2022). Process Intensification Strategies for Power-to-X Technologies. *ChemEngineering*, 6(1), 13.
- [15] Verleysen, K., Parente, A., & Contino, F. (2021). How sensitive is a dynamic ammonia synthesis process? Global sensitivity analysis of a dynamic Haber-Bosch process (for flexible seasonal energy storage). *Energy*, S. 121016.
- [16] Bhaskar, A., Assadi, M., & Homam, N. (2020). Decarbonization of the Iron and Steel Industry with Direct Reduction of Iron Ore with Green Hydrogen. *Energies*, S. 758.
- [17] Allman, A., & Zhang, Q. (2020). Distributed cooperative industrial demand response. *Journal of Process Control*, S. 81--93.
- [18] Bäck, E., Schenk, J., Badr, K., Sormann, A., & Plaul, J. F. (2015). Wasserstoff als Reduktionsmittel für die Eisen-und Rohstahlerzeugung—Ist-Situation, Potentiale und Herausforderungen. *BHM Berg-und Hüttenmännische Monatshefte*, 160(3), 96-102.
- [19] Bhaskar, A., Abhishek, R., Assadi, M., & Somehesaraei, H. (2022). Decarbonizing primary steel production: Techno-economic assessment of a hydrogen based green steel production plant in Norway. *Journal of Cleaner Production*, S. 131339.
- [20] Schlomann, B., & Eichhammer, W. (2013). Energieverbrauch und CO₂-Emissionen industrieller Prozesstechnologien: Einsparpotenziale, Hemmnisse und Instrumente. T. Fleiter (Ed.). Stuttgart: Fraunhofer-Verlag.

Steuerung und Regelung von Abwasserkanalnetzen – ein praxisgerechter Ansatz zur modell-prädiktiven Regelung

Dr. M. Schütze PhD DIC, Dr.-Ing. J. Alex,
ifak e. V., Magdeburg

Kurzfassung

Die Steuerung und Regelung von Abwasserkanalnetzen stellt eine wichtige, aber in vielen Fällen herausfordernde, Aufgabe dar: Zum einen sind Messwerte über den Systemzustand des Abwasserkanalnetzes nur schwierig bzw. oft fehlerhaft zu gewinnen (Messungen im Rohabwasser), zum anderen sind die dem Wellenablauf in Kanalnetzen zugrundeliegenden Prozesse weiträumig örtlich verteilt und nichtlinear (System partieller Differenzialgleichungen) und schließlich sind die Eingangsgrößen des Systems (Niederschlag) hochgradig stochastisch. Auf der anderen Seite erlaubt es eine dynamische globale Abflussregelung, das kostspielige System effizienter als bisher zu nutzen. Insbesondere der Reduktion von Mischwasserüberläufen (verdünntes, aber ungereinigtes Abwasser) in Gewässer kommt eine besondere Bedeutung zu.

Dieser Beitrag stellt die Integration eines MIMO (multiple input-multiple output)-Bausteins zur modell-prädiktiven Regelung in ein weltweit in der Praxis etabliertes Simulationssystem für siedlungswasserwirtschaftliche Systeme vor. Dieser Baustein ermöglicht die Integration eines beliebigen internen Prozessmodells („Digitaler Zwilling“) in die modell-prädiktive Regelung; auch die Zielfunktion für die Anwendung mathematischer Optimierungsverfahren im Rahmen der modell-prädiktiven Regelung kann vom Benutzer flexibel definiert werden. Für die Optimierung stehen dem Nutzer eine Vielzahl von lokalen und globalen Optimierungsverfahren zur Verfügung. Über ein Beobachtermodell und ein Update der Komponenten des Zustandsvektors wird sichergestellt, dass die Optimierung der zu treffenden Regelungsentscheidung für den aktuell vorliegenden Systemzustand getroffen wird. Auch die Definition der Vorhersage- und Steuerungshorizonte sind für den Anwender flexibel möglich.

Der Beitrag erläutert die Umsetzung dieser Verfahren im Simulationssystem Simba# und illustriert die Anwendung einer modell-prädiktiven Steuerung am (in der Siedlungswasserwirtschaft etablierten) Benchmark-Kanalnetz „Astlingen“ (DWA, 2018 [1]). Eine Steuerung mit Hilfe der modell-prädiktiven Regelung (im Englischen: „model-based predictive control“, MBPC) ermöglichte über einen kontinuierlichen Zeitraum von einem Jahr die Reduktion des Mischwasserüberlaufvolumens von ca. 10 % bei ausschließlicher Verwendung von aktuellen Wasserstands- und Niederschlagsinformationen (ohne Niederschlagsvorhersage) als Sensor-Input.

Die Anbindung einer derartigen modell-prädiktiven Kanalnetzsteuerung an ein reales Automatisierungssystem mittels der Laufzeitumgebung ifakFAST (Alex *et al.*, 2020 [2]) wird ebenfalls im Beitrag skizziert.

Abstract

Real time control (RTC) of urban drainage systems represents an important, yet challenging task: Information about the current state of the system (obtained by sensors) is difficult to obtain (measures in raw wastewater), processes are highly distributed and nonlinear (systems of partial differential equations) and system input (precipitation) is of stochastic nature. On the other hand, dynamic control of the flow processes in the sewer system can assist in making better use of this system of critical infrastructure. In particular, the reduction of (essentially untreated) discharges of wastewater into receiving water bodies is of high importance.

This paper describes the integration of a MIMO (multiple input-multiple output) module for model-based predictive control into a worldwide established simulation system. This module allows the integration of an arbitrary internal process model („Digital Twin“) into a model-based predictive control algorithm; also the objective function to be optimised within the model-based predictive control can be defined by the user in a highly flexible way. A number of different local and global optimisers is available to the user. The use of an observer model and an update of the components of the state vector ensures that the determination of the control decision is based on the current state of the system. Also prediction and control horizons can be configured by the user.

In the present paper, the implementation of these concepts within the simulation system Simba# is illustrated for the Astlingen benchmark example of German Water Association (DWA, 2018 [1]). Application of model-based predictive control for this example results in a reduction of overflow volume of around 10 %, using only water level and rainfall sensor information; the evaluation was done by continuous long-term simulation over one year. Furthermore, the connection of such a model-based predictive control to a real-life automation system, using the ifakFAST runtime environment (Alex *et al.*, 2020 [2]), is discussed in the paper.

1. Motivation

Abwasserkanalnetze stellen eines der wertvollsten Infrastruktursysteme menschlicher Siedlungen dar – wie wertvoll sie sind, ist spätestens bei der Kostenermittlung für Bau- oder Sanierungsmaßnahmen von Abwasserkälen festzustellen. Die Kanalnetze nehmen häusliche sowie ggf. industrielle Abwässer (Schmutzwasser) auf; dazu kommen Fremdwasserzuflüsse (beispielsweise durch Infiltration) sowie (im Mischwasserkanalnetzen) Niederschlagsabflüsse.

Die sichere Ableitung der Abwässer zur anschließenden Reinigung auf der Kläranlage ist Hauptziel von Kanalnetzen. Ein weiteres Ziel ist der Gesundheitsschutz, insbesondere durch die Vermeidung von Überflutungen, wie sie insbesondere bei Starkregenereignissen auftreten können.

Bestmögliche Nutzung dieses Infrastruktursystems ist somit von großer Bedeutung. Doch wird von diesen – meist im Wesentlichen statischen – Systemen erwartet, dass sie auch auf hochdynamische Eingangsgrößen (Anfall von Niederschlags- und Schmutzwasser) zufriedenstellend reagieren. Wie bei anderen technischen Systemen lässt sich eine bessere Nutzung jedoch in vielen Fällen durch ein bewusstes Eingreifen – beispielsweise über bewegliche Schieber, Wehre und Pumpen – erreichen. Obwohl es sich hier um eine Regelungsaufgabe handelt, hat sich hierfür in der siedlungswasserwirtschaftlichen Fachwelt der (etwas unscharfe) Begriff „Abflusssteuerung“ (im Englischen „real time control“, RTC) etabliert. Der vorliegende Beitrag stellt die modell-prädiktive Regelung als einen der möglichen Ansätze und ihre Implementierung in einer praxisgerechten und nutzerfreundlichen Simulationsumgebung vor.

2. Methodische Ansätze zur Abflusssteuerung

Die Grundaufgabe der Abflusssteuerung besteht darin, zu jedem Steuerungszeitschritt (beispielsweise 1 Minute oder 5 Minuten) aus den verfügbaren Sensor-Informationen zum Systemzustand des Abwassersystems (beispielsweise Wasserstands- und Durchflussmessungen; Niederschlagsinformationen; ggf. Messungen der Abwasserqualität; ggf. Prognosen von Niederschlägen und Abflüssen), getrieben durch die Ziele des Abwassersystems (u. a. Verringerung von Abschlägen ungereinigten Abwassers in die Gewässer; Überflutungsvermeidung; Kostenreduktion), in diesem Sinne bestmögliche (oder zumindest gute) Eingriffe in die Fließvorgänge im System (durch Aktoren, wie z. B. Pumpen, bewegliche Schieber, Wehre) zu bewerkstelligen.

Hierbei handelt es sich also um eine komplexe Regelungsaufgabe mit besonders herausfordernden Eigenschaften: Hohe Nichtlinearität des Systems, herausfordernde Messbedingungen (Rohabwasser!), stochastische Haupteingangsgröße (Niederschlag), u. U. hohe Komplexität (Vielzahl der Stelleinrichtungen) und Mehrzieligkeit. Daher ist die Kanalnetzsteuerungsaufgabe vielen Verfahren (einschleifige Regelkreise mit linearen PI(D)-Reglern) nicht oder nur unter starken Einschränkungen zugänglich.

Für die Lösung der Kanalnetzsteuerungsaufgabe haben sich folgende drei Herangehensweisen entwickelt; alle drei von ihnen befinden sich auch im praktischen Einsatz in realen Kanalsystemen:

Regelbasierte Steuerung: Hierzu gehören Wenn-Dann-Regelsysteme und Entscheidungsmatrizen und ihre Erweiterungen (z. B. Fuzzy-Regeln), die für sämtliche denkbare Systemzustände die dynamisch vorzunehmenden Einstellungen der Aktoren definieren. Erweiterungen bestehen in der Optimierung von festen Regel-Parametern (Schütze *et al.*, 1999 [3]) oder im Einsatz von maschinellen Lernverfahren zur selbständigen Verbesserung und Erweiterung des Regelsatzes (Khelil, 1993 [4]). Vorteil eines einmal definierten Regelsystems ist die sehr schnelle Abarbeitung während des laufenden Prozesses; jedoch ist seine Erstellung langwierig und verlangt viel Heuristik und Intuition. Zur Entwicklung und Parametrierung der Regelbasis wird immer mit einem Simulationswerkzeug gearbeitet, das ein Modell des Kanalnetzes mit einem Modell der Steuerung (Regel-Netzwerk) kombiniert. Die Funktion der Kanalnetzsteuerung wird zunächst anhand des Verhaltens bei Einzel-Regenereignissen und abschließend in einer Langzeitsimulation (1 - 30 Jahre) mit gemessenen historischen Regenintensitäten bewertet. Hauptkriterium der Bewertung sind häufig die Abschlagsmengen (Mischwasser, das ungereinigt in das Gewässer geleitet wird). Die Simulationsergebnisse insbesondere im Vergleich zum ungesteuerten System sind auch Grundlage der behördlichen Genehmigung des Systems.

Wurde ein einfaches Wenn-Dann-Regelsystem entwickelt, dann kann dieses problemlos in ein klassisches Automatisierungssystem als z.B. SPS-Code implementiert werden. Als Herausforderung bleibt die fehlerfreie Übersetzung der entwickelten Regeln aus dem verwendeten Simulationssystem in SPS-Code. Eine elegante Variante ist hier die Nutzung eines Simulationswerkzeuges, in dem eine SPS-Programmiersprache zur Formulierung der Regeln verwendet werden kann (Alex 2015 [5]), die dann direkt in das SPS-Programmierwerkzeug kopiert werden können. Werden komplexere Beschreibungsansätze verwendet, muss in der Regel Spezialsoftware zur Ausführung dieser Regeln an das Automatisierungssystem gekoppelt werden (z. B. Fuzzy Control, verwendet in Fuchs *et al.* 2005 [6], 1997 [7]).

Verallgemeinerte vorkonfektionierte Regler: Basierend auf der Aufteilung der Regelungsaufgabe auf mehrere – miteinander verbundene – Regelungsaufgaben für Teilsysteme (z. B. Stränge des Kanalnetzes) haben Alex *et al.* (2008 [8]) ein Konzept hierarchischer und miteinander kommunizierender Regler erarbeitet, das aufwandsarm auf (nahezu) beliebige Kanalnetze angewendet werden kann. Hierfür wurde, im Gegensatz zu den meisten anderen Veröffentlichungen im Themenfeld „Abflusssteuerung“, auch eine eingehende regelungstheoretische Untersuchung (basierend auf Wurzelortskurven) durchgeführt. Eine Praxis-Implementierung beschreibt Pabst *et al.* (2011 [9]); auf diesem Konzept basierende Weiterentwicklungen finden sich bei Dirckx (2011 [10]), Gresch (2018 [11]), Kroll (2019 [12]). Spezielle Konzepte für lokale

(und nicht miteinander verbundene) Regler finden sich bei Eulogi *et al.* (2020 [13]). Auch diese Regler müssen in einem Simulationswerkzeug implementiert werden, um das Verhalten des geregelten Kanalnetzes analysieren zu können (gleiches Vorgehen wie für regelbasierte Steuerung). Im Gegensatz zur Entwicklung des Regelwerkes für ein regelbasiertes Steuerungssystem, müssen hier nicht sehr viele Simulationsläufe zur empirischen Verbesserung der Struktur und Parametrierung der Regelbasis durchgeführt werden. Hier werden nur wenige Simulationsläufe zum Tuning weniger Freiheitsgrade im Algorithmus benötigt und es muss der finale Vergleich des geregelten Systems im Vergleich zum ungeregelten System in einer Langzeitsimulation durchgeführt werden.

Diese verallgemeinerten Regler sind relativ einfach strukturiert und lassen sich mit Standard-SPS-Programmiersprachen formulieren. Ist ein entsprechender Funktionsbaustein einmal erstellt und getestet, ist der Implementierungsaufwand sehr gering.

Modell-basierte Regelung: Klassische modellbasierte und modell-prädiktive Ansätze finden sich auch in der Kanalnetzsteuerung (u. a. Marinaki und Papageorgiou (2005 [14]), Ocampo Martinez (2010 [15]), Pleau *et al.* (2005 [16]), Rohlfing (1993 [17]), Lund *et al.* (2018 [18])). Vorteile liegen in der optimalen Anpassung an das betrachtete Kanalnetz. Das vorhandene Steuerungspotential lässt sich weitgehender ausschöpfen. Herausforderungen liegen hier in der Definition des internen Modells (Modellbildung bzw. Modellvereinfachung für ein hochgradig nichtlineares, komplexes System), das für die Optimierung der Steuerungseingriffe verwendet wird, sowie in der ausgewogenen Definition der für diese Optimierung angewendeten Zielfunktion. Für die Entwicklung von prädiktiven Kanalnetzsteuerungen entstehen weitergehende Anforderungen an das eingesetzte Simulationssystem. Der Simulator muss nicht nur das Kanalnetz und einfache Steuerungsfunktionen simulieren können, sondern auch der prädiktive Regler (bestehend aus mehreren internen Modellen und entsprechenden Optimierungsalgorithmen) muss simuliert werden. Diese Anforderung ist in den bisher verfügbaren und ingenieurtechnischen Simulationswerkzeugen nicht gegeben. Bisherige akademische Lösungen kombinieren unterschiedliche Spezialsoftwaresysteme (z. B. Matlab, PySWMM) und organisieren das Zusammenspiel mit einer geeigneten Skriptsprache z.B. Python (siehe u.a. Stenotto *et al.*, 2020 [19]).

Wenn diese akademische Herausforderung gemeistert wurde, muss das resultierende komplexe modellbasierte Regelungssystem noch möglichst praxistauglich implementiert und an ein Automatisierungssystem gekoppelt werden. Eine direkte Integration in existierende Automatisierungssysteme ist nicht verfügbar. Hier muss das Konstrukt aus akademischen Softwaresystemen an das Automatisierungssystem gekoppelt und in Echtzeit ausgeführt werden.

3. Modell-prädiktive Abflusssteuerung

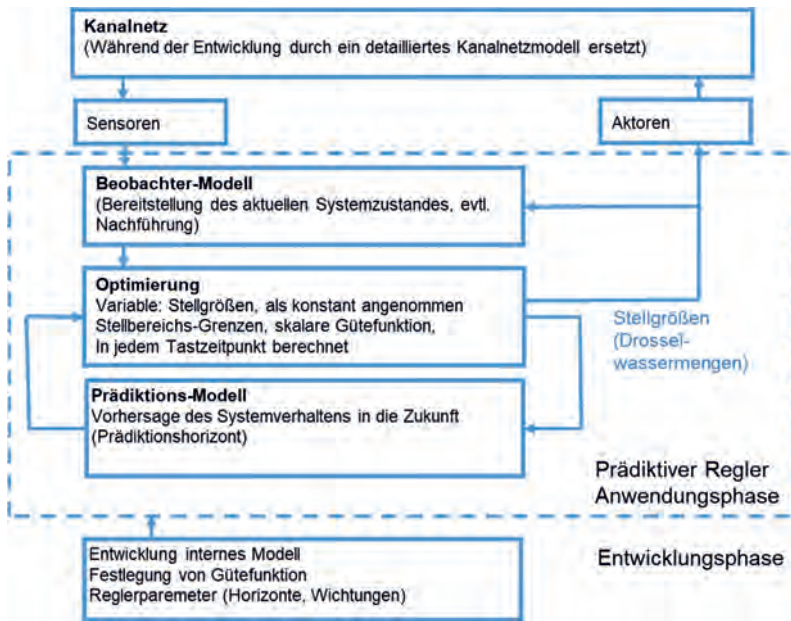


Bild 1: Modell-basierte Regelung eines Kanalnetzes

Abbildung 1 skizziert das generelle Konzept einer modell-prädiktiven Regelung (Alex 2015 [5]) im Anwendungsfeld „Kanalnetz“. Die erfolgreiche Funktion einer modell-prädiktiven Regelung ist von folgenden vier Faktoren abhängig:

Internes Modell („Digitaler Zwilling“): Von besonderer Bedeutung ist hier das interne Prozessmodell („Digitaler Zwilling“), welches in der Anwendung eines mathematischen Optimierungsverfahrens zu jedem Steuerungszeitschritt viele Male simuliert wird, um die jeweilige Zielfunktion auszuwerten und ihren Wert (o. B. d. A.) zu minimieren. Das interne Modell muss einfach genug sein, um eine Echtzeitoptimierung zu ermöglichen, aber detailliert genug, um die für die Kanalnetzsteuerung wichtigen Phänomene zu erfassen (insbesondere Einstau und Überläufe). Für diese Aufgabe bieten sich u. U. hydrologische Modelle des Netzes an.

Zur Entwicklung des Systems muss auch ein Ersatzmodell des realen Systems bereitgestellt werden. Dies kann ein Kopie des internen Modells des Reglers sein, aber für einen realistischen Test sollte ein detaillierteres, in der Regel hydrodynamisches, Modell genutzt werden. Das externe Simulationsmodell dient zum simulativen Test der Steuerung und wird in der endgültigen praktischen Anwendung durch das reale System ersetzt (virtuelle Inbetriebnahme). Das externe Modell muss es erlauben, zu jedem Steuerungszeitschritt die Routinen zur Ermittlung der Steuerungsentscheidung aufzurufen und die Simulation dann mit diesen Einstellungen fortzusetzen. Diese Notwendigkeit erfordert einen Eingriff in den Quellcode des externen Programms oder das Vorliegen (oft kostenpflichtiger) spezieller Schnittstellen.

Zustandsschätzung: Für die Prädiktion des künftigen Systemverhaltens wird eine möglichst gute Abschätzung des aktuellen Systemzustandes benötigt (nichtlineares System). Dies kann entweder, wenn möglich, aus der Messung des vollständigen Systemzustandes des realen Systems oder durch ein Beobachtersystem erfolgen. In dem typischen Fall der Nutzung eines Beobachters wird eine zweite Instanz des internen Modells parallel zum realen System simuliert und optional ein Abgleich des Zustandes des Beobachtermodells mit dem realen System implementiert. Der jeweilige Zustand des Beobachtermodells kann dann als Startzustand für das strukturgleiche Prädiktionsmodell verwendet werden.

Gütefunktion: Für die Optimierung muss eine skalare Gütefunktion bereitgestellt werden, die dann minimiert werden soll. Für klassische Regler gibt es generische Vorschläge (Norm der Abweichung des Regelgrößenverlaufs vom Sollwert). Für den gewünschten Kanalnetzbetrieb müssen spezielle Gütefunktionen definiert werden. Diese Aufgabe ist nicht trivial und erfordert einige Versuche in der Entwicklungsphase des Reglers. Um alle Anforderungen an das System in einer skalaren Gütefunktion zu integrieren, müssen Wichtungsfaktoren verwendet werden. Hier entstehen sensitive Reglerparameter, die in der Entwurfsphase sorgfältig bestimmt werden müssen. Einen wichtigen Einfluss hat auch der verwendete Prädiktionshorizont, typischerweise ein definierter Zeitbereich in der Zukunft.

Optimierer: Die Optimieraufgabe stellt sich als lokale, beschränkte Optimierung einer skalaren Gütefunktion mit einem Vektor von unabhängigen Optimierungsvariablen dar, für die keine Gradienteninformationen vorliegen. In der in den allermeisten Fällen verfolgten „Step-Input-Policy“ werden Stellgrößen optimiert, die über den Prädiktionshorizont als konstant angenommen werden. Damit hat der Vektor der unabhängigen Variablen die Dimension des Stellgrößenvektors. Als Stellgrößen werden typischerweise Drosselwassermengen für Regenüberlaufbauwerke

angesetzt. Lokale Regler setzen diese Vorgabe um. Für die Stellgrößen wird ein erlaubter Stellbereich festgesetzt. Minimale Drosselwassermengen (>0) resultieren aus dem Wunsch, den Einstau von reinem Schmutzwasser zu verhindern, maximale Drosselwassermengen resultieren aus hydraulischen Kapazitäten des jeweils untenliegenden Systems. Theoretisch wären globale Optimierverfahren zu bevorzugen, aber der signifikante erhöhte Bedarf an Gütefunktionsauswertungen ist in Echtzeit nicht zu leisten. In den von den Autoren untersuchten Anwendungen konnten mit den Algorithmen COBYLA (Constrained optimisation by linear approximation, a constrained version of Powell's derivative-free search algorithm, Powell 1998 [20]) und BOBYQA (Bound optimisation by quadratic approximation, Powell 2009 [21]) aus der NLOpt-Bibliothek (Johnson [22]) gute Erfahrungen gesammelt werden.

4. Implementierung von Kanalnetz-MBPC in ein ingenieurtaugliches Entwicklungs- und Laufzeitsystem

Der in vielen anderen industriellen Anwendungen weit verbreitete Lösungsansatz einer modellprädiktiven Regelung ist also auch für die Kanalnetzsteuerungsaufgabe (mit ihren besonderen Anforderungen, s.o.) von Interesse. In bisher umgesetzten praktischen Anwendungen (z. B. Ocampo Martinez (2010 [15]), Pleau *et al.* (2005 [16]), Stentoft *et al.*, 2020 [19]) war Spezialsoftware für die Definition der internen Modelle sowie ihrer Ankopplung an das externe Simulationsmodell bzw. den realen Prozess erforderlich.

Um nun die Anwendung von modellprädiktiven Regelungen im Kanalnetzbereiche auch in größerem Rahmen zu unterstützen, sind entsprechende Module in ein aktuelles, nutzerfreundliches (und weltweit zum Einsatz bei Ingenieurbüros, Betreibern, Abwasserverbänden, Behörden und Hochschulen kommendes) Simulationssystem für Wassersysteme (SIMBA ifak, 2021 [23]) integriert worden.

Dieses Simulationswerkzeug erlaubt die Simulation von Kanalnetzen mit vereinfachten (hydrologischen) oder auch mit detaillierten (vollständig hydrodynamischen) Modellen, ggf. auch mit detaillierter Betrachtung der Abwasserinhaltsstoffe. Für die untersuchten Kanalnetze lassen sich alle oben beschriebenen Varianten zur Kanalnetzsteuerung (regelbasierte Steuerung, verallgemeinerte vorkonfektionierte Regler, MBPC) simulieren. Der Simulator ist aber auch als Laufzeitumgebung für komplexere Regler (MBPC) zur Regelung realer Prozesse geeignet (Förster und Alex, 2021 [24]). Damit sind Entwurf, virtuelle Inbetriebnahme und Einsatz als reale Steuerung in einem Werkzeug integriert. Für ein einfaches akademisches Beispiel ist ein entsprechendes Simulationsmodell mit einfachem Prozessmodell, Beobachtermodell mit empirischer Nachführung und MBPC-Reglerblock in Abbildung 2 dargestellt. Die internen Modelle

(normales Simulationsmodell ohne Einschränkungen) werden im selben System entwickelt und die Projektnamen werden als Parameter an den Beobachterblock (*Observer*) und den modell-prädiktiven Regler übergeben. In dem Beispiel wird die empirische Nachführung des Beobachtermodells (Blöcke *Add n*, *Gain*, *Saturation*) genutzt, um die unbekannte Störung am Prozesseingang z zu schätzen und gleichzeitig den Zustand des Beobachtermodells an das Prozessmodell anzugleichen. Bei dem in diesem Beispiel definierten linearen Prozess könnte das Beobachtermodell zur Vorgabe des Startzustandes für die Prädiktion entfallen, für den allgemeinen nichtlinearen Fall ist dieses Modell aber erforderlich.

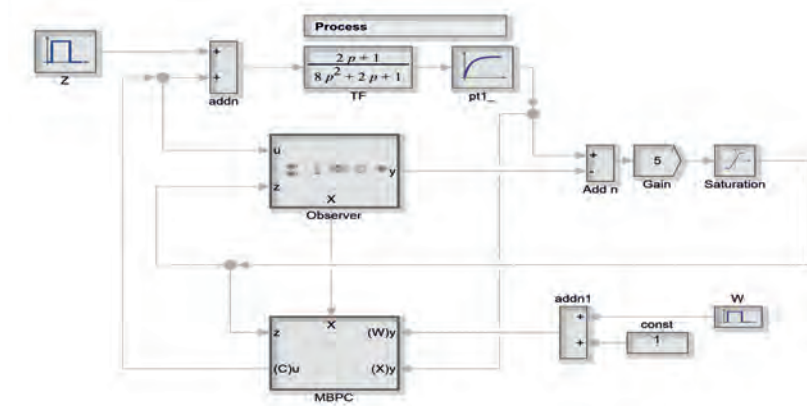


Bild 2: Simulationsmodell mit einfachem Prozessmodell, Beobachtermodell mit empirischer Nachführung und MBPC-Reglerblock

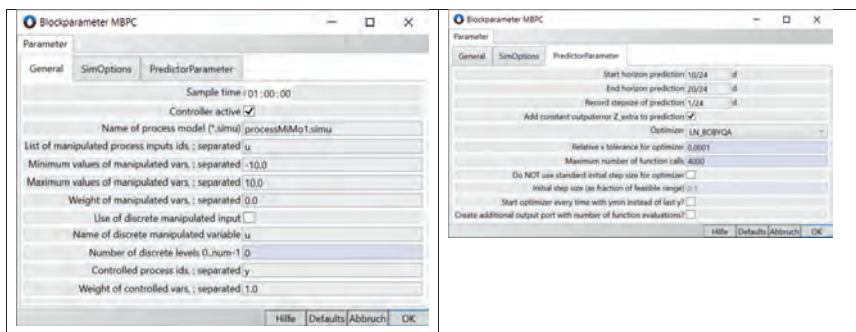


Bild 3: Parameter MBPC-Reglerblock

In den in Abbildung 3 dargestellten Parameter-Dialogen des MBPC-Blocks wird erkennbar, wie das interne Modell eingestellt, die Stellgrößen ausgewählt und die Stellbereiche konfiguriert werden. In dem Reiter *PredictorParameter* werden dann u. a. der Prognose- und Auswer-tehorizonte definiert und der Optimierungsalgorithmus gewählt. Da der MBPC-Block generell nur „Step-Input-Policy“ unterstützt, muss der Stellhorizont nicht definiert werden. Der Block kann als Mehrgrößenregler konfiguriert werden und unterstützt neben mehreren analogen Stellgrößen auch einen diskreten Stelleingang für gemischt diskret analoge Regler.

5. Anwendungsbeispiel Astlingen

Das Beispielkanalnetz Astlingen ist von der DWA-Arbeitsgruppe „Integrale Abflusssteuerung“ erarbeitet und in (DWA, 2018 [1]) ausführlich dokumentiert; eine Zusammenfassung findet sich in Schütze *et al.* (2017 [25]). Abbildung 4 zeigt eine schematische Darstellung.

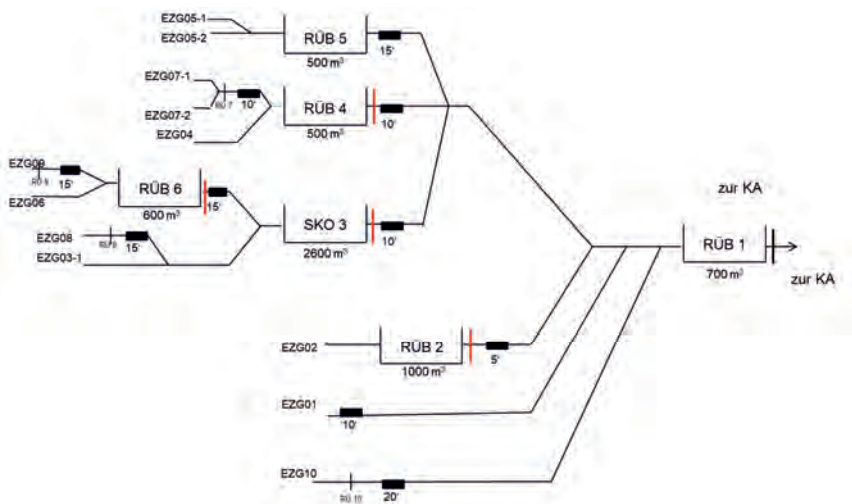


Bild 4: Schema Astlingen (adaptiert nach [1])

Das angenommene System lässt sich mit folgenden Stichpunkten charakterisieren:

- Größe der Kommune: 33.000 Einwohner,
- Abwasseranfall: 49 l/s Schmutzwasser + 39 l/s Fremdwasser,
- Gesamtfläche: 3.100 ha,

- Speichervolumen: 5.900 m³ (32.7 m³/ha A_u),
- 6 Mischwasserbauwerke, von denen vier steuerbar sind, und 4 Regenüberläufe.

Die Steuerungsaufgabe besteht nun darin, zu jedem Steuerungszeitschritt die Abflüsse der Becken so einzustellen, dass die Ziele der Steuerung erreicht werden. Die Ziele der Steuerung sind durch die nachstehend formulierte Zielfunktion beschrieben. Über Wichtungsfaktoren werden die Zielkriterien „Abweichung der Beckenfüllgrade vom durch das untenliegende Becken gegebenen Referenzfüllgrad“, „Überlaufvolumen“, „Abweichung vom Nominalabfluss (aus statischem Systementwurf)“, „Auslastungsdefizit der Kläranlage“, „Variation der Einstellungen in aufeinanderfolgenden Zeitschritten“ gewichtet und zu einem Funktional zusammengefasst.

Die Zielfunktion ist in diesem Anwendungsbeispiel definiert als

$$J = \sum_{i=n_1}^{n_2} \left[a_1 \frac{1}{4} \sum_{b=1}^4 (F_{b,i} - F_{ref,i})^2 + a_2 \frac{q_{ov,i}}{q_{ov0}} + a_3 \frac{1}{2} \sqrt{\sum_{b=1}^4 \left(\frac{q_{d,b,i} - q_{nom,b}}{q_{nom,b}} \right)^2} + a_4 \left(1 - \frac{q_{KA,i}}{q_{KA,max}} \right) \right] + \sum_{b=1}^4 f_{u,j} \frac{1}{T_0} \sqrt{(q_{d,b,i+1} - q_{d,b,i})^2}$$

mit

$n_1 \dots n_2$	Prädiktionshorizont
$a_1, \dots, a_4$	Gewichte zur Wichtung der Teilziele (z. B.: 2; 10; 0.1; 0.1)
$F_{b,i}$	Füllgrad Becken b zum Zeitpunkt i
$F_{ref,i}$	Füllgrad Referenzbecken (letztes Becken vor Kläranlage) zum Zeitpunkt i
$q_{ov,i}$	Gesamtüberlaufvolumen zum Zeitpunkt i
q_{ov0}	Wichtungsparameter für Gesamtüberlaufvolumen
$q_{d,b,i}$	Eingestellter Drosselabfluss für Becken b zum Zeitschritt i
$q_{nom,b}$	Nominal-Drosselabfluss für Becken b (z. B. aus Systementwurf)
$q_{KA,i}$	Zufluss zur Kläranlage zum Zeitschritt i
$q_{KA,max}$	Maximal zulässiger Zufluss zur Kläranlage
T_0	Tastzeit (hier: 5 Min.)

Im hier beschriebenen Anwendungsbeispiel ist, bei einem Steuerungszeitschritt von 5 Minuten, der Prädiktionshorizont auf das Zeitintervall $[+10'; +30']$ in der Zukunft gesetzt, der Steuerungshorizont auf 30' (wobei nur die Steuerungsentscheidung für die ersten 5 Minuten an die Aktoren übermittelt wird) und der Vorhersagehorizont wahlweise auf 0' (danach Nullannahme oder letzte-Wert-Annahme für die Niederschlagsdaten) oder auf die volle Länge des Prädiktionshorizontes gesetzt.

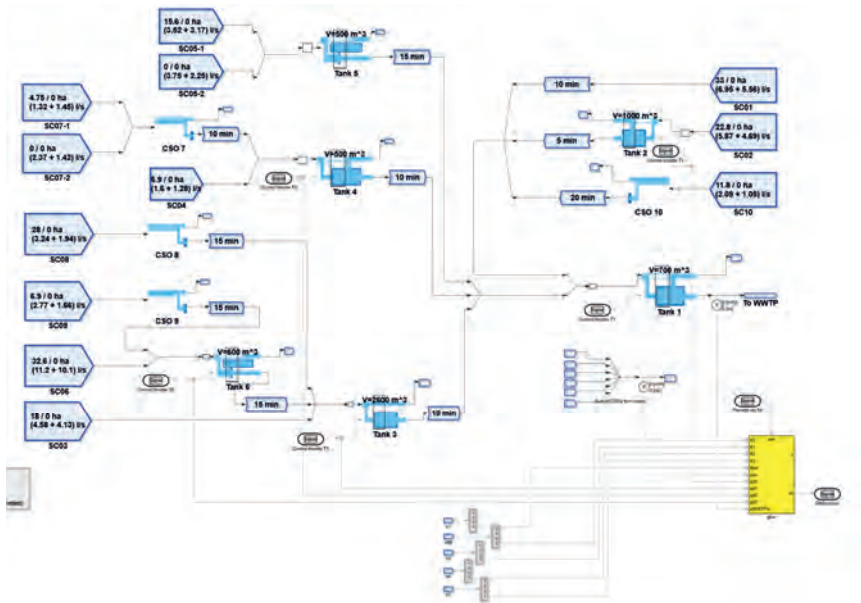


Bild 5: Auszug internes Modell Kanalnetz Astlingen

In Abbildung 5 ist das interne Modell für die MBPC (und den Beobachter) dargestellt. Die hier beschriebene Gütefunktion wurde grafisch programmiert und im Block *gfun* berechnet. Dieser Block ist ein sogenannter Konverterblock, der zur Flexibilität des Simulationswerkzeuges beiträgt. Dieser Block kann durch den Anwender mit textuellen Formeln parametrisiert werden. In Abbildung 6 ist der Formeleditor dieses Blocks dargestellt.

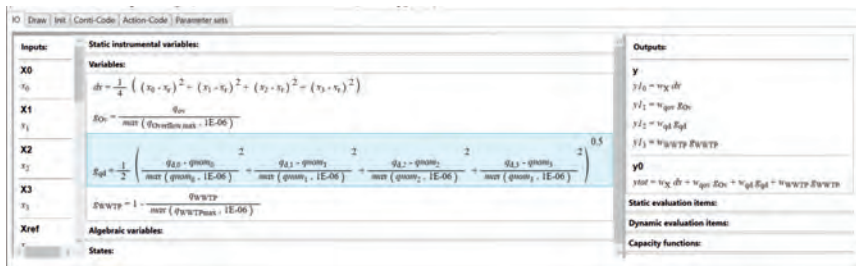


Bild 6: Programmierung der anwendungsspezifischen Gütefunktion

Diese modell-prädiktive Regelung wird nun für einen Testzeitraum von einem Jahr angewendet. In vielen Veröffentlichungen zur Kanalnetzsteuerung erfolgt ein Test lediglich einzelner Niederschlagsereignisse. Diese vermögen nicht das gesamte Spektrum möglicher Regenerereignisse abzudecken. Daher erfolgt hier eine Simulation unter Anwendung der Steuerung über kontinuierliche Niederschlagszeitreihen an vier Regenschreibern über ein Jahr. Abbildung 7 zeigt das kumulierte Überlaufvolumen (als eines der definierten Ziele) im Vergleich mit einigen anderen Ansätzen zur Kanalnetzsteuerung.

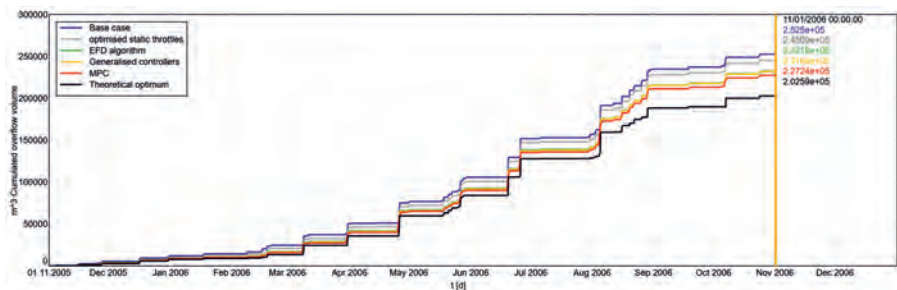


Bild 7: Vergleich verschiedener Steuerungsansätze für Astlingen – kumuliertes Gesamtüberlaufvolumen über alle Simulationszeitschritte

Hier zeigen die einzelnen Kurven (vgl. Tabelle 1), mit welchem Betriebsführungsansatz sich welches Ergebnis für das gegebene Kanalnetz erreichen lässt.

Tabelle 1: Simulationsergebnisse

Base case	Der Basisfall: Das Kanalnetz wird nicht gesteuert – Situation in den allermeisten Netzen. Die Drosseln der Regenüberlaufbecken sind einmalig auf Basis von Entwurfsüberlegungen fest eingestellt.
Optimised static throttles	Verbesserte Drosseleinstellungen: Auch ohne Steuerung kann das Systemverhalten verbessert werden. Mit Auswertung von Betriebserfahrungen und Analyse mit Simulationsuntersuchungen lassen sich die statischen Drosseleinstellung feinjustieren und ein besseres Ergebnis erreichen.
EFD	Equal-Filling-Degree-Algorithmus: Diese Regelung ist ein einfacher Algorithmus aus der Familie der „Verallgemeinerten vorkonfektionierten Regler“, Steuerungsziel ist eine gleichmäßige Füllung aller Speicherbecken im System. die Performance ist in etwa gleichwertig zu dem zweiten Algorithmus aus der Familie der „Verallgemeinerten vorkonfektionierten Regler“, die grüne Kurve liegt hinter der gelben Kurve.
Generalised Controller	Verallgemeinerter Kanalnetzregler (nach Alex <i>et al.</i> 2008 [8]): Diese Regelung ist ein weiterer einfacher Algorithmus aus der Familie der „Verallgemeinerten vorkonfektionierten Regler“. Hier ein ähnlicher und signifikanter Performancegewinn wie EFD gegenüber dem Basisfall.
MPC	Die vorgestellte MBPC: nochmals eine kleinere Verbesserung der Regelgüte (Überlaufvolumens) gegenüber den einfachen Regelungskonzepten. Diese Performance dokumentiert, dass MBPC ein sinnvoll auf Kanalnetze anwendbares Regelungsprinzip darstellt und von allen bekannten Regelungsansätzen hier die beste Performance erreicht.
Theoretical Optimum	Eine Abschätzung der theoretisch möglichen Performance einer Kanalnetzregelung, abgeschätzt nach dem Zentralbeckenansatz

6. Unterwegs zur praktischen Anwendung

Wie oben dargestellt, ist das verwendete Simulationssystem nicht nur zur Planung von MBPC geeignet, sondern ist auch als Laufzeitumgebung für modellbasierte Regler vorgesehen.

Für diesen Einsatz ist eine Verbindung zwischen dem simulierten Prozess und dem realen Automatisierungssystem (SPS + SCADA) erforderlich. Die offene Middleware-Plattform ifak-FAST (Framework for integrated automation and simulation technologies) kann generell als Laufzeitsystem für die Organisation der Echtzeitsimulation und des Datenaustausches mit dem Automatisierungssystem eingesetzt werden. ifakFAST wurde in den letzten Jahren im Rahmen mehrerer Forschungs- und Industrieprojekte entwickelt (Hübner *et al.* 2013 [26]). Ziel dieser Entwicklung ist es, die Umsetzung von Lösungen zur Überwachung und Optimierung

des Betriebs verfahrenstechnischer Prozesse zu unterstützen. Der Anwendungsschwerpunkt liegt im Bereich wasserwirtschaftlicher Anlagen und Prozesse, insbesondere Abwasserreinigungsanlagen. Der Datenaustausch zwischen den Eingängen des Beobachter- und des MBPC-Modells und dem realen System kann mittels der Plattform ifakFAST realisiert werden. Zur Ankopplung an das Automatisierungssystem wurden in den bisher umgesetzten Projekten OPC und OPC UA verwendet. Alternative Schnittstellen sind verfügbar bzw. können leicht durch das modulare Daten-Adapterkonzept bereitgestellt werden. Darüber hinaus werden Optionen zur Datenspeicherung von Zeitreihendaten, Dashboard, Nutzerverwaltung und Alarmierungsfunktionen bereitgestellt; wenn möglich sollten hierfür jedoch die im existierenden SCADA vorhandenen Funktionen genutzt werden. Für eine größere Kläranlage ist eine mit diesem Werkzeug realisierte überlagerte Ammoniumregelung seit mehreren Jahren im Einsatz.

7. Schlussfolgerungen

In diesem Beitrag wurde die Integration von Modulen zur modell-prädiktiven Kanalnetzsteuerung in ein weltweit im abwassertechnischen Praxiseinsatz befindlichen Simulationssystem dargestellt und ihre Anwendung auf das Benchmark-Kanalnetz Astlingen demonstriert. Somit steht das Instrumentarium der modell-prädiktiven Kanalnetzsteuerung, einschließlich ihrer Anbindung an reale Automatisierungssysteme nun auch Praktikern und Endanwendern zur Verfügung, ohne dafür auf akademische Spezialsoftware oder Spezialschnittstellen zugreifen zu müssen.

Weitere Anwendungen, wie z. B. die Einbeziehung von Kläranlage und Fließgewässer in die Simulation und Steuerung (vgl. Schütze *et al.*, 2017 [27]) sind nun unmittelbar möglich. Des Weiteren ist die Anwendung des Simulators und der dargestellten Steuerungsmodule nun auch für andere wasserwirtschaftliche Aufgabenstellungen, wie z. B. der Regenwassernutzung (*rainwater harvesting*) (vgl. Gilboa *et al.* [28]) denkbar.

Die bequeme Verfügbarkeit modellprädiktiver Steuerungsansätze ermöglicht es nun auch, Fragen bzgl. der konkreten Ausgestaltung eines Steuerungssystems (z. B.: Einbeziehung von Regenvorhersagen in welchem Umfang? Ideale Anzahl und Orte von Steuerelementen?) simulativ zu analysieren.

Insgesamt ermöglichen die erarbeiteten Module und Werkzeuge auf komfortable und praxisgerechte Weise eine verbesserte Nutzung kritischer Infrastruktur sowie einen verbesserten Gewässerschutz.

Literatur

- [1] DWA (2018): (2018): DWA-Themen: Abflusssteuerung in Kanalnetzen – Anwendungsbeispiel. ISBN 978-3-88721-697-9, DWA-Arbeitsgruppe ES-2.4 "Integrale Abflusssteuerung"; Hennef/Sieg, November 2018
- [2] Alex, J., Hübner, C., Förster, L. (2020): Planning, testing and commissioning of automation solutions for waste water treatment plants using simulation, 21st IFAC World Congress, 11.-17.07.2020
- [3] Schütze, M.; Butler, D.; Beck, M.B. (1999): Optimisation of control strategies for the urban wastewater system – an integrated approach; Water Science and Technology 39, 9, 209-216
- [4] Khelil, A., Heinemann, A., Müller, D. (1993): Learning algorithms in a rule based system for control of UDS. 6th Int. Conf. on Urban Storm Drainage, Niagara Falls, 1401-1408
- [5] Alex, J. (2015): Simulationsplattform zum integrierten Prozess- und Automatisierungsentwurf von Abwassersystemen. at – Automatisierungstechnik 63, 7, 553-563
- [6] Fuchs, L., Beeneken, T. (2005): Development and implementation of a real-time control strategy for the sewer system of the city of Vienna, Water Science and Technology, 52. 5. 187–194.
- [7] Fuchs, L., Beeneken, T., Scheffer, C., Spönmann, P. (1997): Model based real-time control of sewer systems using fuzzy logic; Water Science and Technology 36, 8-9, 343-347.
- [8] Alex, J., Schütze, M., Ogurek, M., Jumar, U. (2008): Systematic Design of Distributed Controllers for Sewer Networks; IFAC World Congress, Seoul 2008
- [9] Pabst, M., Alex, A., Ogurek, M., Niclas, C., Peikert, D., Schütze, M., Bär, T. (2011): ADESBA – A new general global control system applied to the Hildesheim sewage system, 12th International Conference on Urban Drainage, 11.-16.09.2011.
- [10] Dirckx, G., Schütze, M., Kroll, S., Thoeve, C., De Guedre, G., Van De Steene, B. (2011): Cost-efficiency of RTC for CSO impact mitigation. Urban Water Journal, 8, 6, 367-377
- [11] Gresch, M. (2018): Praxistaugliche Regelung von Entwässerungsnetzen – das Projekt INKA: 3. Expertenforum Regenüberlaufbecken Baden-Württemberg, DWA, Stuttgart, 27.02.2018
- [12] Kroll, S. (2019): Design of Real-Time Control Strategies for Combined Sewer Networks, Katholieke Universiteit Leuven.
- [13] Eulogi, M. Ostojin, S. Skipworth, P., Shucksmith, J.D., Schellart, A. (2020): Hydraulic optimisation of multiple flow control locations for the design of local real time control systems, Urban Water Journal, doi.org/10.1080/1573062X.2020.1860238

- [14] Marinaki, M., Papageorgiou, M. (2005): Optimal real-time control of sewer networks; Springer; London, Berlin, Heidelberg
- [15] Ocampo Martínez, C. (2010): Model-based Predictive Control of Wastewater Systems.
- [16] Pleau, M., Colas, H., Lavallée, P., Pelletier, G., Bonin, R. (2005): Global optimal real-time control of the Quebec urban drainage system, *Environmental Modelling and Software* 20, 4, 401-413
- [17] Rohlfing, R. (1993): Echtzeitsteuerung von Entwässerungssystemen mit Optimierungsverfahren: Durchführbarkeitsanalyse mit hydrodynamischer Simulation der Abflussvorgänge; Dissertation; Universität Hannover
- [18] Lund, N.S.V., Falk, A.K.V., Borup, M., Madsen, H., Mikkelsen, P.S. (2018): Model predictive control of urban drainage systems: A review and perspective towards smart real-time water management, *Critical Reviews in Env. Sci. and Techn.* 48, 3, 279-339
- [19] Stentoft, P.A., Vezzaro, L., Mikkelsen, P.S., Grum, M., Tychsen, P., Madsen, H., Halvgaard, R. (2020): Integrated model predictive control of water resource recovery facilities and sewer systems in a smart grid: example of full-scale implementation in Kolding, *Wat. Sci. Tech.* 81, 8, 1766–1777
- [20] Powell, M. J. D. (1998): Direct search algorithms for optimization calculations, *Acta Numerica* 7, 287-336
- [21] Powell, M. J. D. (2009): The BOBYQA algorithm for bound constrained optimization without derivatives, Department of Applied Mathematics and Theoretical Physics, Cambridge, England, Technical Report NA2009/06
- [22] Johnson, S. G.: The NLOpt nonlinear-optimization package, <http://github.com/stevengj/nlopt>
- [23] ifak (2021): Simba#water, Version 5. Manual, Institut für Automation und Kommunikation e. V. (ifak), Magdeburg
- [24] Förster, L.; Alex, J. (2021): Werkzeug für die modellgestützte Betriebsführung von Kläranlagen. *at - Automatisierungstechnik* 69(3): 256-267, 2021
- [25] Schütze, M., Lange, M., Pabst, M., Haas, U. (2017): Astlingen – ein Benchmark-Beispiel zur Kanalnetzsteuerung. DWA-GMA Gemeinschaftstagung „Mess- und Regelungstechnik in abwassertechnischen Anlagen“, Niedernhausen/Ts., 30./31.05.2017
- [26] Hübner, C.; Thron, M.; Alex, J.; Bangemann, T. (2013). Akteur-basierte Middleware-Plattform für fehlertolerante, verteilte SCADA-Systeme. *AUTOMATION* 2013, 25.-26.06.2013, Baden-Baden, VDI Wissensforum GmbH, 2013, ISBN 978-3-18-092209-6

- [27] Schütze, M., Lange, M., Pabst, M., Haas, U. (2017): Real-time control of Astlingen – considering sewer system and WWTP, 14th Int. Conf. on Urban Drainage, Prag, 10. – 15.09.2017
- [28] Gilboa, Y., Friedler, E., Schütze, M. (eingereicht): Assessing water use and reuse options - A holistic analysis of a Model City, coupling dynamic system modelling with Life-Cycle Assessment, zur Veröffentlichung eingereicht

Nonlinear Model Predictive Control: from Digital Twin to Real-World Application in Polymerization Process

Dr.-Ing. **Bernd-Markus Pfeiffer**, **Chris Leingang**,
Dr.-Ing. **Corinna Busse**, Dr.-Ing. **Mathias Oppelt**,
Solutions Process Automation, Simulation & Digitalization,
Siemens AG, Erlangen and London, Great Britain;
Dr.-Ing. **Otmar Lorenz**, Siemens AG, Karlsruhe,
Dr. Alejandro Cano, Siemens PSE, Parsippany, USA

Abstract

This paper gives an overview and recent results of an ongoing pilot application of a nonlinear MPC (Model Predictive Controller) to an industrial polymerization process. Both MPC and soft sensor are using the same digital process twin, a high-fidelity dynamic flowsheet model based on first principles, which is also the heart of an operator training system.

1. Introduction

Combining technologies from automation, simulation and digitalization in a seamless way enables the next step of digital transformation, towards a more resource efficient and sustainable industrial production.

Model Predictive Control (MPC) is established as the most powerful and most successful method for multivariable control in process industries. Most industrial applications of MPC are still based on linear dynamic process models that are identified from step tests or other active experiments in the plant. If however a first-principles dynamic model of the process unit to be controlled already exists, it is efficient to use this model inside an MPC algorithm. This is a decisive advantage for processes where a linear control concept is not sufficient, e.g. batch processes or semi-continuous processes with multiple products and frequent grade changes, like e.g. polymerization processes (see [6.] for an extensive comparison of linear and nonlinear MPC). gPROMS nonlinear MPC [9.] by Siemens PSE [7.] provides such a functionality. The digital process twin can be used inside nonlinear MPC, but also for soft-sensing of product quality attributes that up to now are only known from lab samples, pathing the way to closed-loop quality control.

2. Process Description

In a polymerization reactor for high-density polyethylene (HDPE), a mixture of gaseous monomer and chain transfer agent, liquid co-monomer and solvent, and a solid catalyst in suspension is fed to the reactor. Inside the reactor polymer macro-molecules are growing from catalyst particles. The specific example is a low-pressure slurry process for the production of high-performance multimodal HDPE grades. The liquid phase is hexane as the medium in which the reaction mixture is suspended. Afterwards, the polymer is separated from the slurry in the downstream section and dried. The final polymer powder is transported to the addition and extrusion section.

The strongly exothermic chemical reaction takes place in a continuously stirred tank reactor in continuous operation. The following chemical compounds are used in this process:

- Ethylene C_2H_4 : Monomer
- Butene C_4H_8 : Co-Monomer
- Hydrogen H_2 : Chain termination
- Nitrogen N_2 : Inert gas
- Hexane C_6H_{14} : Solvent, mother liquor

Currently, the plant operator specifies setpoints for C_2 -Feed, partial pressure C_2 and reactor temperature. C_2 feed as the main raw material determines reactor throughput. Partial pressure C_2 is controlled by a PID controller manipulating catalyst feed. Reactor temperature is controlled by a PID controller manipulating cooling water flowrate. Temperature setpoint and feed flows for co-monomer and hydrogen are specified in a value range according to recipe specifications that are provided on a sheet of paper in the control room. Product quality is mainly defined by polymer properties melt flow rate (MFR) and density. These properties cannot be measured online in the reactor, but only by taking lab samples in a cycle of 2 to 4 hours. Lab results are available half an hour after extraction of samples, which is too late for closed loop feedback control.

Different polymer grades with very specific properties are produced in a continuous process. The changeover from one polymer grade to the next takes place by simultaneously manipulating several reaction conditions (e.g., feed flows, temperature) while the chemical reaction is running. Today such grade changes are controlled by experienced plant operators, and the output created during grade changes is typically off-spec material.

3. Process Model

The first-principles [5.] flowsheet model [5.] of the polymerization reactor includes the existing PID controllers in the DCS (see Fig. 1), controlling e.g. feed flows, reactor level and tempera-

ture, because they all will be slave controllers of the MPC master controller in a cascade structure. The model contains mass balances for all chemical species, energy balances and detailed reaction kinetics. Besides the polymerization reactor, auxiliary process units like coolers, flash vessels, condensers, decanter and mother liquor collecting tank are part of the model. The first-principles model has more than 6000 equations including more than 400 differential equations. Model parameters are fitted to steady state process data for the different polymer grades produced alternately in the same reactor. PID controller parameters are taken from the original DCS, considering gain scaling factors. Dynamics of closed PID control loops e.g. temperature control are checked against dynamic process data before developing the soft-sensor application.

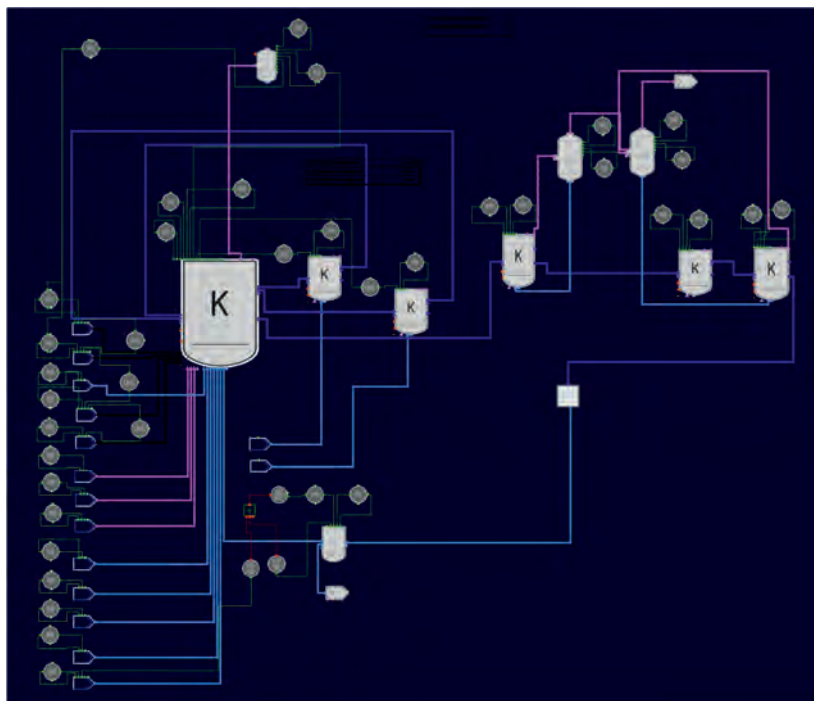


Fig. 1: gPROMS flowsheet model of polymerisation process: reactor with coolers, flash vessel, condensers, decanter and mother liquor collecting tank, including PID control loops in existing DCS

The process model is implemented using gPROMS software by Siemens PSE. In 2019, Siemens has acquired PSE (Process Systems Enterprise, [7.]), which was a spin-off from

Imperial College in London. PSE is active in the area of process simulation for more than 20 years. The acquisition of PSE by Siemens opens the path to integrate PSE software into distributed control systems like SIMATIC PCS 7 and exploit gPROMS models online during production.



Fig. 2: gPROMS (“general PROcess Modelling Suite”) by Siemens PSE. On top of the general modeling platform, there are two software suites: “gPROMS Process” is used for continuous production processes handling fluids and gases in different industry branches like chemicals, oil&gas, water/wastewater; “gPROMS FormulatedProducts” is used for processes that involve not only fluids, but also solids, with a special focus on pharmaceutical processes

The model of the HDPE polymerization reactor is build using components of the standard “Process” library and a dedicated “Polyolefins Library” (see Fig. 2), including custom models for different types of polymerization reactors such as fluidized bed reactor, slurry reactor, loop reactor and solution polymerization reactor. The following chemical reactions are considered in the model: activation, initiation, chain propagation, deactivation, chain transfer, site transformation, chain termination. Different reaction mechanisms are taken into account: spontaneous reaction or reaction kinetics driven by cocatalyst, transfer agents, impurities or monomers. Multiple types of active sites on catalyst grains are distinguished.

The overall gPROMS model describes the behavior of the real process very well, both with respect to steady states and with respect to transients like load changes. It can be used for all known grades produced in this reactor, and also for dynamics of grade changes. The model includes some simplifications (e.g. assumption of perfect mixing, method of moments to represent polymerization reactions) to enable the model to solve rapidly enough for online optimization. It is not intended for designing the chemical process, i.e. detailed equipment design or design of new grades. Models for such purposes need to be even more complex.

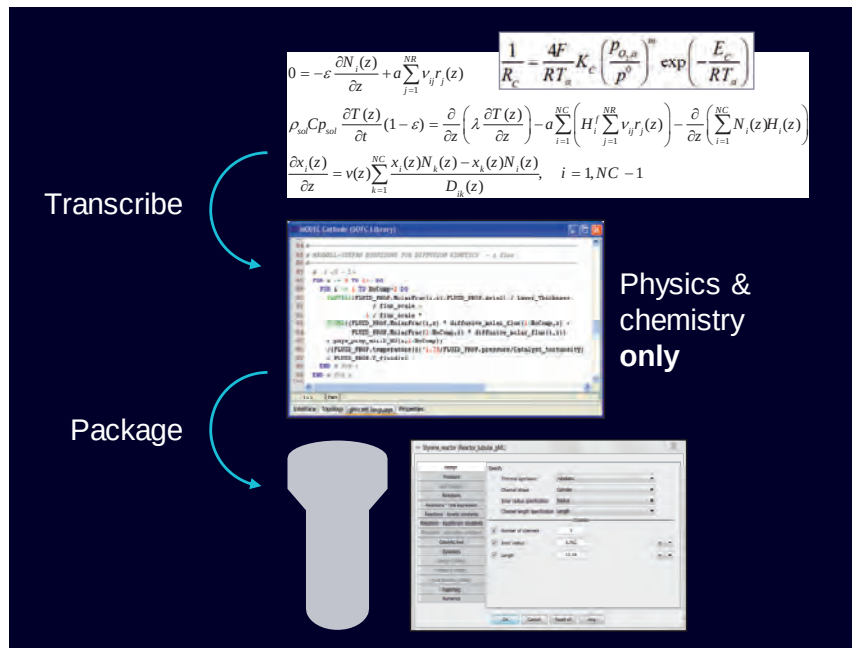


Fig. 3: The gProms custom modelling platform allows to create high-fidelity models of virtually any type of process, using a purely declarative description with focus on physics, not on solution algorithm. The solvers for systems of algebraic, ordinary/partial differential and integral equations are provided by the platform.

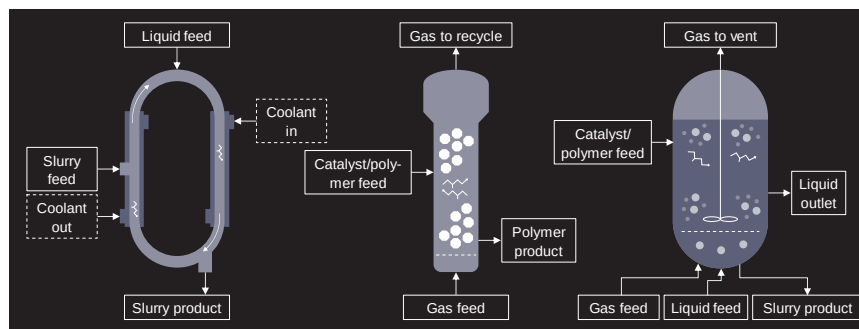


Fig. 4: Different types of reactor models in gProms Polyolefines Library [10.]: loop reactor, fluidized bed reactor and solution polymerization reactor

4. Soft Sensor

Any concept for automatic quality control of polymerization will rely on the ability to estimate the polymer properties online, especially MFR, based on current reaction conditions in reactor. This is the task of a so called "soft sensor", a piece of software that replaces a real sensor device ("hardware") by a model-based online calculation. Sporadic lab samples are used for development of soft-sensor and online adaptation of model parameters. In principle, there are two different options to create a soft sensor for such a polymerization process:

- Purely data-driven black-box models, e.g. based on artificial intelligence algorithms (artificial neural networks). Black-box models can only provide interpolation between data-points contained in training data, but no extrapolation, and they do not contribute to process understanding.
- White-box models based on first principles, as described in section "process model". In this project, the white-box approach is preferred, to exploit deep process understanding and achieve maximal process transparency.

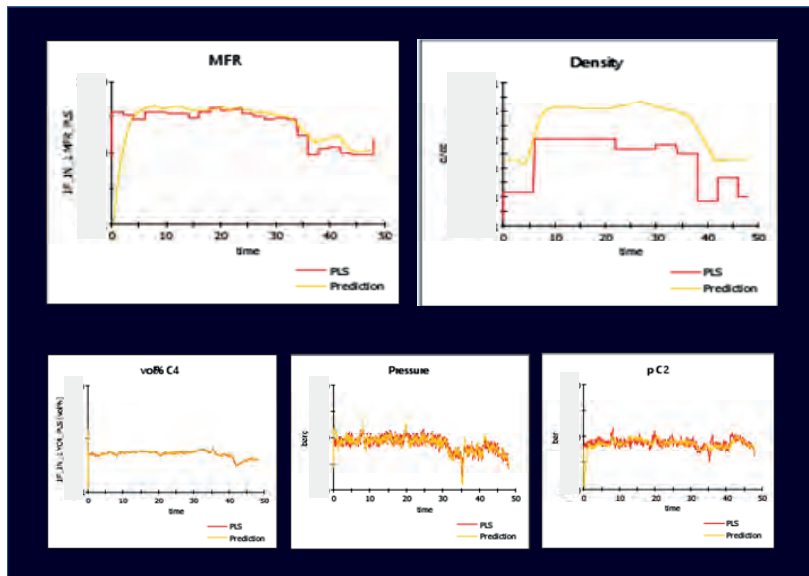


Fig. 5: Test run with EKF on historical plant data including a grade change. Time axis is scaled in hours. Yellow curves are estimated by soft sensor, red curves are measured values from DCS. The red curves for product quality (MFR and density) are lab samples plotted in staircase form. Polymer density for reactor content is calculated by soft sensor as well, but cannot be directly compared to lab samples due to mixing after reactor. (Axis scalings have to be hidden for confidentiality reasons.)

Inside soft sensor, the first-principles process model is running in real time in parallel to the real process, receiving the same input signals as the real process. Measured variables like partial pressure of monomer, co-monomer and hydrogen are compared to respective model predictions. Differences between model outputs and measurement values are used to adapt model parameters and internal state variables using an EKF algorithm (Extended Kalman Filter). Model fit including dynamic state estimation by EKF algorithm is tuned and validated with dynamic process data from real plant, including all relevant grade changes (see Figure 5).

5. Nonlinear Model Predictive Control

Many of the basic MPC principles are valid both for linear and nonlinear MPC ([1.] , [3.]):

- IMC principle (Internal Model Control [4.]): a dynamic process model is not only used for offline controller design but constitutes an integral part of the online control algorithm, where it is used to predict future process behaviour along a specified finite prediction horizon.
- Formulation of the feedback control problem as an optimization problem: minimize the squared sum of “control deviations” and other “costs” of control like MV (Manipulated Variable) moves over a finite time horizon.
- Moving horizon principle: although the optimization results contain a planning of future MV moves along the complete control horizon, only the MV values for the next sample step are actually implemented on the plant. In the next sample step, new measured values for the CVs (Controlled Variables) will arrive, the predictions will be updated and new MV values will be calculated by optimization.

However, the improved predictive power delivered by a nonlinear first-principles model comes with an increased computational burden for nonlinear MPC. The superposition principle to calculate CV movements as a sum of contributions from effects of different MVs on a CV is not valid. Some mathematical properties like the convexity of the MPC inherent optimization problem for linear process models and quadratic performance index are no longer applicable for the nonlinear case. This can make the solution of the optimization problem in real time much more challenging.

The requirements for an advanced process control solution for polymerisation reactors [2.] can be summarized in two categories:

- Closed loop quality control during production of a specific polymer grade. This involves feedback control of conditions inside reactor (partial pressure of monomer, co-monomer and hydrogen) and product quality (namely MFR calculated by soft sensor) according to recipe specification, by providing optimal setpoints for existing PID control loops for reactor temperature, catalyst flow, hydrogen and co-monomer feed. During constant production, fluctuations of product quality have to be minimized and all MV constraints have to be fulfilled.
- Automatic grade change: manipulate reaction conditions while the chemical reaction is running, in such a way that the reactor will produce a new grade with different MFR value. The time needed for grade transition shall be minimized, to minimize off-spec production in between two grades. During grade transition, MV constraints allow some more space for movement than during constant production, but at the end of

the grade transition, MV values must arrive inside constraints defined by recipe of target grade.

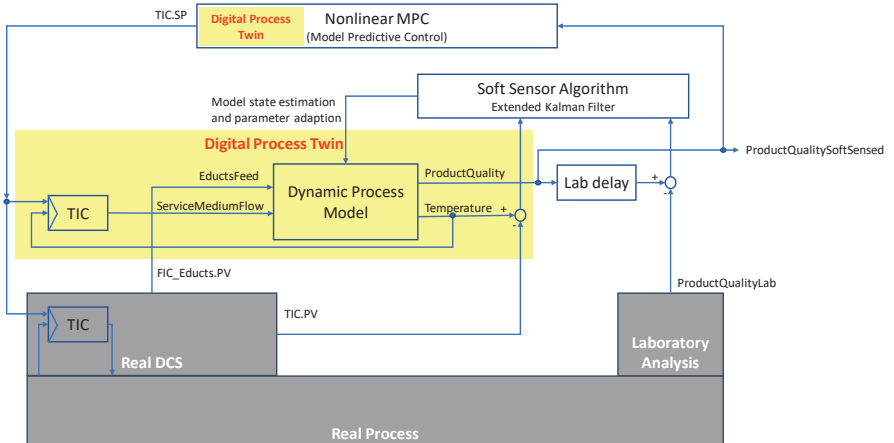


Fig. 6: Signal flow of model-based APC solution, where Digital Process Twin (yellow) is used both as soft sensor and as predictor inside nonlinear MPC. Only a representative subset of signals and PID controllers is displayed for reasons of clarity.

Intensive discussions with process experts of plant operating company yield the following decisions on control concept for MPC:

- Monomer feed is specified by plant operators to define plant throughput. MPC is not allowed to manipulate monomer feed, but it can use the information on current monomer feed as measurable disturbance variable.
- The existing PID controller for reactor temperature via cooling water flow into reactor jacket is kept as it is, but the temperature setpoint is provided by MPC.

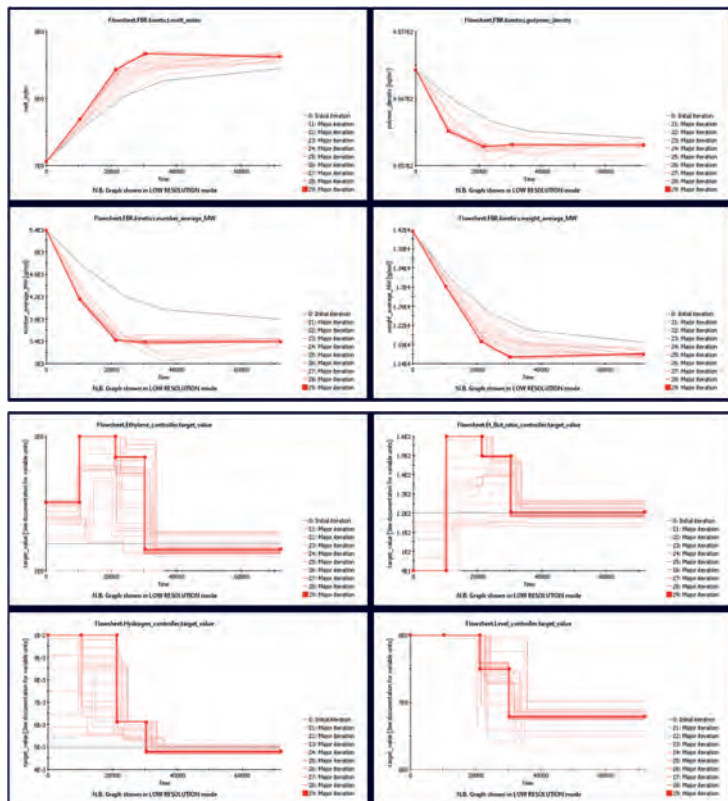


Fig. 7: Iterative solution of dynamic optimization problem for grade change in a polymerization reactor, bottom figure: possible trajectories of decision variables (MVs), top: resulting predicted trajectories of controlled variables (CVs), optimal solution marked by bold line. (These plots are intended to show the MPC working principle, they are not created with the current reactor model.)

The following extensions of control concept are planned:

- Reaction conditions are best described by ratio of partial pressure hydrogen and partial pressure co-monomer to partial pressure monomer. These ratios can be used as controlled variables for MPC instead of the absolute partial pressures.
- In contrast to temperature control, the existing PID controller for partial pressure monomer via catalyst flow can be replaced by MPC, i.e. partial pressure monomer is

a CV for MPC, and catalyst flow is an MV. The existing PID controller for partial pressure monomer had to be switched to manual mode during grade changes, and MPC needs the handle on catalysts flow to move the process to a new grade.

- Subordinate decisions which of several pipes is used for hydrogen feed, or which of several catalyst pumps shall be activated, are automated with elementary new functions inside DCS. MPC will only provide sum of catalyst flow, no matter which pumps are used.

The MPC is running in a cycle time of 10 min, using a control horizon of 50 min and a prediction horizon of nearly 22 h.

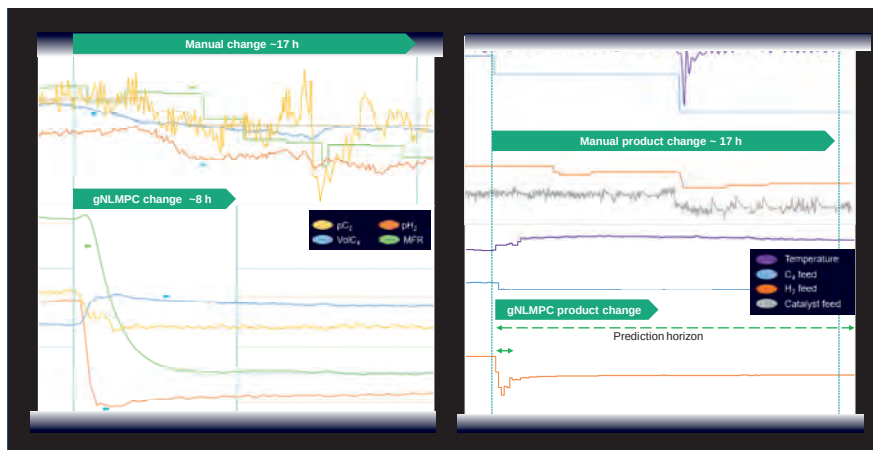


Fig. 8: Grade change, performed by human operators in real plant (upper part of figures), compared to automated grade change run by MPC (lower part of figures); CV trends shown in left figure, MV trends in right figure. (Axis scalings have to be hidden for confidentiality reasons.)

Fig. 8 shows an example of a grade change run by human operators in real plant. They change setpoints of several PID controllers related to reactor temperature and reactor feeds in several steps and need about 17 hours to complete the grade change. They heavily rely on lab results which do no longer represent the true state of the process during a transition if they arrive with the delay of half an hour, and for the two hours between lab sampling, plant operators are running in “blind flight”. Using the new APC solution, the soft sensor provides a continuous picture of actual process state even during transitions. For the operators, a grade change just means to specify a new combination of CV setpoints (e.g. MFR) according to

recipe. The MPC will perform the setpoint step in automatic mode with minimal overshoots and will need less than 8 hours to complete the grade change, according to simulation results with high fidelity process model and real plant constraints. MPC starts and accelerates the transition by “oversteering” H_2 feed to a value which is further away from old operating point than final steady state value in new grade.

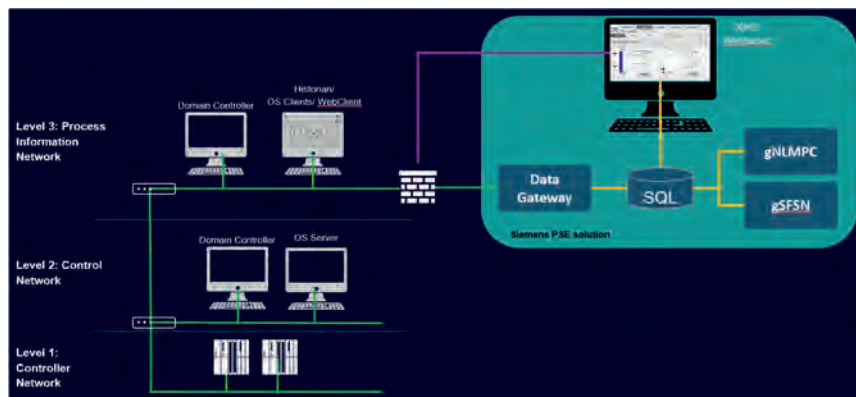


Fig. 9: Solution architecture using gDAP (gPROMS Digital Application Platform) connected via OPC UA to existing DCS. GUI for APC application is developed in XHQ reading data from SQL database of gDAP.

The MPC is currently tested in the production plant of a customer in Germany, and the results up to now are promising. The whole APC solution is running in a virtual machine connected to process information network of existing DCS. gDataGateway [8.] reads/writes data via OPC UA from/to OS server of DCS and performs data validation. All data relevant data for APC solution are collected in a SQL database. Nonlinear MPC and soft sensor are computational modules inside gDAP. A dedicated web-enabled GUI for APC solution is implemented in XHQ [11.], which is also connected to SQL server.

Some practical issues still need to be resolved, e.g. what to do if plant operators close manual valves in the plant that are not connected to DCS, or if they perform maintenance activities at gas analysers, and invalid signals arrive in model-based APC software that are not marked as invalid.

6. Outlook

The goal of model-based solution for HDPE polymerization consisting of soft sensor and nonlinear MPC using the same high-fidelity process model is to minimize fluctuations of

product quality during constant production, and to achieve time-optimal grade transitions with minimal give-away. This is one exemplary way how to improve resource efficiency of industrial production by means of digitalization, in this case by a digital process twin in online operation. The same digital process twin is also used offline in an operator training system for this plant.

Literature

- [1.] Allgöwer, F., Badgwell, T. A., Qin, J. S., Rawlings, J. B., Wright, S. J. (1999). Nonlinear predictive control and moving horizon estimation - An introductory overview. In P. M. Frank, editor, *Advances in Control, Highlights of ECC'99*, pages 391-449. Springer.
- [2.] Bartusiak, R.D. (2007): NLMPC: A Platform for optimal control of feed- or product-flexible manufacturing. In Findeisen, R., Allgöwer, F., Biegler, L.T. (ed.): *Assessment and future directions of nonlinear model predictive control. Lecture Notes in Control and Information Sciences*, pp. 367-381. Springer, Berlin.
- [3.] Allgöwer, F., Findeisen, R., Ebenbauer, C. (2008): Nonlinear model predictive control. *Encyclopedia Of Life Support Systems*. http://ifatwww.et.uni-magdeburg.de/syst/about_us/people/findeisen/papers/EOLSS.pdf
- [4.] Garcia, C.E., Prett, M., Morari, M. (1989): Model predictive control: theory and practice – a survey. *Automatica* 25 Nr. 3, pp. 335-348.
- [5.] Pantelides, C.C.; Renfro, J.G. (2013): The online use of first-principles models in process operations: Review, current status and future needs. *Computers & Chemical Engineering*, Volume 51, April 2013, pp. 136-148.
- [6.] Pfeiffer, B-M., Oppelt, M., Leingang, Ch., Pantelides, C., Pereira, F. (2020): Nonlinear Model Predictive Control based on Existing Mechanistic Models of Polymerisation Reactors. IFAC World Congress 2020, Berlin, Jul. 2020.
<https://www.sciencedirect.com/science/article/pii/S2405896320322825>
- [7.] PSE – A siemens business: <https://www.psenterprise.com/company/pse-a-siemens-business>
- [8.] gPROMS Digital Application Platform (gDAP):
<https://www.psenterprise.com/products/gproms/digital-applications/platform>
- [9.] gPROMS Nonlinear MPC (gNLMPC):
<https://www.psenterprise.com/products/gproms/digital-applications/gnlmpc/>
- [10.] gPROMS Process libraries:
<https://www.psenterprise.com/products/gproms/process/libraries>
- [11.] XHQ Operations Intelligence:
<https://new.siemens.com/global/en/products/automation/industry-software/xhq-operations-intelligence-software.html>

Kognitive Automations-Architektur für Cyber-Physische-Produktionssysteme (CPPS)

B.Sc. **André Mankowski**, M.Sc. **Daniel Antonow**,
Dipl.-Math. **Natalia Moriz**, Prof. Dr.-Ing. **Henning Trsek**,
inIT – TH OWL, Lemgo

Kurzfassung

CPPS eröffnen neue vielfältige Möglichkeiten Mehrwerte wie bspw. Datengetriebene Selbst-diagnose von Produktionsanlagen, Self-X-Szenarien, intelligente Service-Planungen oder vereinfachte (Prozess)Daten-Aufbereitung und Darstellung hochgradig modularer und damit auch komplexer Systeme. Die Konfiguration und Bedienung solch moderner Produktionssysteme wie CPPS sind daher komplex und erfordern oftmals speziell geschultes Personal, um diese Komplexität beherrschen zu können. Der Markt verlangt jedoch eine einfachere Einrichtung, Bedienung und Überwachung dieser Systeme, um den Anforderungen der Industrie 4.0 gerecht zu werden. Dieser Beitrag beschreibt daher zum einen den ersten prototypischen Ansatz einer ganzheitlichen kognitiven Automations-Architektur für CPPS, die den Umgang mit derartigen Systemen vereinfacht und durch moderne Sprachsteuerung und NLP intuitiv und interaktiv gestaltet. Zum anderen können durch analytische intelligente Services und einen standardisierten Datenzugriff mittels OPC UA auf die Prozessebene neue Mehrwerte generiert und dabei die Komplexität von CPPS für den Menschen - in erster Linie für das Bedien- oder Wartungspersonal - reduziert oder weitestgehend verborgen werden. Die aus drei Schichten bestehende Architektur wurde dabei exemplarisch an einem CPPS in Form einer Montagezelle für Fidget-Spinner (VDMA OPC UA Demonstrator) umgesetzt und erprobt.

1. Einleitung und Motivation

Der zunehmende Einsatz von Cyber-Physischen-Systemen in der produzierenden Industrie erhöht die Möglichkeiten und Fähigkeiten heutiger Produktionssysteme, so entstehen sogenannte Cyber-Physische-Produktionssysteme (CPPS). In CPPS stehen heute umfangreiche Daten wie z.B. Sensorwerte, Energiemessungen und Produktinformationen zur Verfügung. Datengetriebene Analysemethoden nutzen diese Daten, um automatisch neues Wissen über das System zu gewinnen. Dieses Wissen trägt u.a. dazu bei, bspw. Verfahren zur Selbstdiagnose zu entwickeln, die zukünftig die Basis intelligenter Automatisierungstechnik bilden, womit die im Rahmen von Industrie 4.0 (I4.0) anvisierten Ziele einer wandlungsfähigen, ressourcen-effizienten und ergonomischen Produktion erreicht werden können. Der Einsatz von I4.0-Ansätzen z.B. zu Diagnosezwecken oder kognitiver Systeme erfordert umfangreiches Know-how über die verwendeten Methoden sowie über die benötigten (Prozess)Daten, die bspw. semantisch beschrieben und standardisiert zur Verfügung gestellt werden müssen. Dies stellt oftmals ein Hindernis dar, da es an gesamtheitlichen Lösungsansätzen mangelt, die einerseits den Umgang mit CPPS erleichtern und andererseits eine Möglichkeit schaffen vom Mehrwert, den diese Systeme bieten, einfacher und vor allem intuitiver profitieren zu können.

Um dieser Herausforderung zu begegnen, haben die Projektpartner ISI Automation GmbH und das Institut für industrielle Informationstechnik (inIT) der Technischen Hochschule Ostwestfalen-Lippe (TH OWL) in dem Forschungsvorhaben „Kognitive Automation in der Praxis“ (KoAP) eine kognitive Automations-Architektur (s. Bild 1) realisiert, die mit Hilfe moderner Technologien wie Natural Language Processing (NLP), OPC UA und intelligenter Services ein intuitiv und interaktiv bedienbares CPPS ermöglicht. In diesem Beitrag wird eine Architektur bestehend aus drei Schichten beschrieben: (i) Ein-/Ausgabeschicht bzw. HMI, welches natürlich sprachliche Eingaben durch den Benutzer entgegen nimmt, analysiert, Abfragen an das CPPS stellt und entsprechende Systemantworten dem Benutzer in einfach verständlicher Form übermittelt. (ii) Kognitive Schicht, die mittels einer einfachen Ontologie basierend auf den Eingabeparametern geeignete intelligente Services ermittelt, ausführt und die entsprechenden Systemantworten generiert. (iii) Schicht drei beinhaltet den standardisierten Anlagenzugriff und die semantische Beschreibung der grundlegenden Prozesswerte, welche die Basis für die intelligenten Services der kognitiven Schicht darstellen.

2. Stand der Technik

Kognitive Automation

Die Kognitive Automation stellt einen weitreichenden Entwicklungsschritt in der Automation von Produktionsanlagen dar [1], [2], [3]. Neuartige, intelligente Assistenzsysteme helfen zunehmend dem menschlichen Experten komplexe Systeme beherrschbar zu halten, z.B. bei: Anpassen des Automatisierungssystems an neue Anlagenkonfigurationen (Selbstkonfiguration), Identifikation von Fehlern und Verschleißerscheinungen (Selbstdiagnose) [4]. Einzelne Assistenzsysteme müssen in eine integrierende, interdisziplinäre Systemarchitektur eingebunden sein, um eine Vernetzung unterschiedlicher Komponenten, Systeme und Prozesse abzubilden. Nur eine Systemarchitektur, welche in produzierenden Unternehmen als Basis für eine selbstoptimierende Wertschöpfungskette umgesetzt werden kann, wird den Anforderungen nach der I4.0. Referenzarchitektur gerecht werden können [5]. In den letzten Jahren erzeugen umfangreiche Aktivitäten in Deutschland einen Schub für die Entwicklung diverser Referenzarchitekturen für CPPS-Lösungen, wie bspw. das Spitzencluster it's OWL [6] oder ERANET [7]. Bei diesen Ansätzen zeigt sich ein wesentlicher Unterschied zwischen der konventionellen Automation und neu zu entwickelnden Systemen: die Verwendung von explizitem, formalisiertem Domänenwissen durch intelligente Produktionssysteme, z.B. Domänenwissen über die Anlage, das Automatisierungssystem oder die Umgebung. Dieses Wissen wird durch Data-Mining-Ansätze maschinell verarbeitet. Trotz großem Interesse am Thema „Kognitive Automation“ müssen noch diverse offene Forschungsfragen gelöst werden, die z.B. den direkten Einsatz CPPS-basierter Lösungen/Architekturen in laufenden Produktionsanlagen ermöglichen [8], [9].

Cloud-Lösungen/ Intelligente Services

Die Komplexität von CPPS und die zunehmende Verlagerung von Steuerungsintelligenz in Feldgeräte [10] erfordern neue Architekturen, die sich von der konventionellen hierarchischen Automatisierungspyramide unterscheiden. So wurde z.B. in [11] eine neue globale Informationsarchitektur für industrielle Informationen vorgeschlagen, in der die zeitkritischen Ebenen (Steuerungs- und Feldebene) traditionell und die oberen Ebenen durch die Cloud gelöst sind. Aktuell gibt es aber nur wenige Cloud-basierte Lösungen auf Steuerungs- und Feldebene der Automation [12]. So wurde in [13] ein neuer, auf [11] basierender Ansatz für die Cloud-Lösung präsentiert, welcher Control-Module für Produktionsmodule als Services anbietet.

OPC UA als Schlüsseltechnologie der Industrie 4.0

OPC UA ist eine plattformunabhängige und serviceorientierte Architektur (SOA), die durch eine Reihe von Normen unter der IEC 62541 [13] spezifiziert ist. OPC UA ermöglicht eine konsistente horizontale und vertikale Kommunikation zwischen Feldebene und höheren Ebenen der klassischen Automatisierungspyramide. Für den Datentransport bietet OPC UA zwei allgemeine Paradigmen: Das Client/Server-Modell, das sich strikt an ein Request/Response-Pattern hält, und das Pub/Sub-Modell nach einem Publish/Subscribe-Ansatz. Für die Entwicklung von OPC UA Anwendungen existiert eine Vielzahl sowohl kommerzieller als auch open-source SDKs (z.B. open62541 [15]) für jede gängige Programmiersprache und Plattform. OPC UA liefert die Grundlage für die in diesem Beitrag vorgestellte kognitive Automations-Architektur.

3. Konzept kognitive Automations-Architektur:

Im Folgenden wird das Konzept der kognitiven Automations-Architektur näher beschrieben. Die Architektur ist dabei wie folgt als 3-Schicht-Architektur realisiert.

Schicht I - HMI: Diese Schicht ermöglicht die Interaktion des Menschen (z.B. Bedien- oder Wartungspersonal) mit dem Automationssystem über ein entsprechendes Mensch-Maschine-Interface (MMI/HMI). Das HMI nimmt dabei. natürlichsprachliche Eingaben (Sprachbedienung) oder ausformulierte Texteingaben entgegen, analysiert diese nach bestimmten Kriterien und extrahiert Schlüsselwörter, die im Anschluss an die nächste Verarbeitungsschicht der Architektur über eine definierte Schnittstelle weitergereicht werden. Parallel dazu nimmt das HMI die vom System ermittelten und angefragten Informationen in einem definierten Format entgegen und stellt diese anschaulich und einfach verständlich dem Bedien- bzw. Wartungspersonal dar.

Schicht II – Kognitive Schicht: Die zweite Schicht der Architektur befasst sich mit der Interpretation und dem Mapping von Schlüsselwörtern und deren Synonymen (Wortontologie) auf definierte analytische Services (Intelligent Services). Das Mapping zwischen Schlüsselwörtern und der zur Verfügung stehenden Services erfolgt durch einen sogenannten. „Discovery Service“. Dieser wählt anhand der Schlüsselwörter den passenden Service (z.B. Sensor abfragen oder Wartungsinformationen ermitteln) aus und ruft diesen über eine definierte Schnittstelle mit entsprechenden Eingabeparametern auf. Der aufgerufene Service wird ausgeführt und

greift dabei über eine Prozessschnittstelle auf die zugrundeliegenden Prozessdaten des Systems zu, um die angefragten Informationen zu ermitteln. Das Ergebnis der Anfrage bzw. des Serviceaufrufs wird anschließend wieder an das HMI kommuniziert und ausgegeben.

Schicht III – Anlagenzugriff: Die dritte Schicht stellt alle benötigten Meta- und Prozessdaten des Systems über definierte Schnittstellen wie bspw. OPC UA (Server-Informationsmodelle + Companion-Spezifikationen) und/oder zusätzlicher Datenbankschnittstellen für historische Prozesswerte zur Verfügung.

Die Schnittstellen 1-3 aus Bild 1 sind dabei REST-kompatibel. Der Datenaustausch geschieht über das JSON-Format und entsprechend definierter JSON-Strukturen. Die Schnittstellen 4 - 6 sind OPC-UA-basiert und erlauben den Zugriff auf statische und dynamische sowie auf historische Prozesswerte.

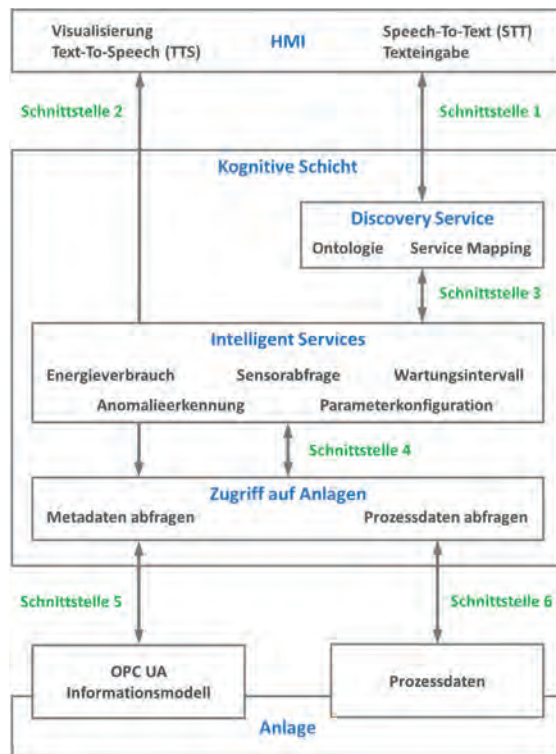


Bild 1: Kognitive Automations-Architektur als 3-Schicht-Architektur

4. Implementierung und Umsetzungsbeispiel

In diesem Abschnitt wird eine prototypische Implementierung der kognitiven Automationsarchitektur vorgestellt, welche die aktuellen technologischen Möglichkeiten zur Umsetzung einer kognitiven Architektur anhand eines Demonstrators aufzeigt. Als Basis für die Architektur dient der VDMA OPC UA Demonstrator (*Bild 2: VDMA OPC UA Demonstrator*), der im Folgenden als CPPS betrachtet wird und über einheitliche Informationsschnittstellen in Form von OPC UA Informationsmodellen (Companion Spezifikationen) verfügt. Somit stellt dieser eine ideale Grundlage für die angestrebte kognitive Architektur dar.

VDMA OPC UA Demonstrator: Mit dem erstmals auf der AUTOMATICA 2018 vorgestellten VDMA OPC UA Demonstrator [14] (*Bild 2*) wurde u.a. eine Neuheit in Bezug auf Interoperabilität und durchgängiger sowohl vertikaler als auch horizontaler Kommunikation auf Feld-ebene Controller-To-Controller-Kommunikation über OPC UA vorgestellt. Der Demonstrator veranschaulicht einen Produktionsprozess für die Montage von Fidget Spinnern und verfügt dabei über eine vollständig OPC UA-basierte



Bild 2: VDMA OPC UA Demonstrator

Maschinenarchitektur mit hierarchischer Struktur bestehend aus Zellen-, Stations-, Substations- und Komponentenebene. Eine detailliertere Beschreibung des Demonstrators findet sich in [16]. Der Demonstrator besteht aus sechs Hauptmontagestationen, die wiederum in Unterstationen unterteilt sind. Jede Hauptstation, Unterstation sowie jede Komponente besitzt dabei eine eigene Steuerung inklusive OPC UA Server mit einheitlichem Informationsmodell.

Als Hauptaugenmerk für die erste prototypische Umsetzung einer ganzheitlichen kognitiven Architektur – d.h. alle Schichten umfassend – dient zunächst lediglich eine Station des Demonstrators. In diesem Fall die Einpressstation der Kugellager für die Fidget-Spinner Montage. Diese Station verfügt über eine Steuerung mit integriertem OPC UA Server und bietet statische (z.B. Hersteller- oder Wartungsinformationen) wie dynamische Prozessinformationen (z.B. aktueller Luftdruck oder Energieverbrauch) über ein standardisiertes Informationsmodell ihrer Umwelt an. Auf diese Weise ist die erste Grundlage eines einheitlichen Anlagenzugriffs (Schicht 3 der Automationsarchitektur aus **Fehler! Verweisquelle konnte nicht gefunden werden.**) gewährleistet.

Implementierung HMI & Discovery Service:

Die prototypische Implementierung des HMI (Schicht 1) umfasst eine Weboberfläche, die als Benutzerschnittstelle zur Verfügung steht und Sprach- sowie Texteingaben des Benutzers entgegennimmt. Getätigte Spracheingaben werden analog zu dem aus dem Consumer-Bereich bekannten Sprachassistenten wie Amazon Alexa oder Google Home-Assist durch ein angeführtes Schlüsselwort (in diesem Fall „ISI“) aktiviert und anschließend an einen Cloud-basierten Sprachanalyseservice (z.B. Google Natural Language API) übermittelt. Via Speech-To-Text (STT) wird die Spracheingabe in Text umgewandelt und das Resultat zurück an das HMI übertragen. Durch den Benutzer direkt getätigte Texteingaben sowie die STT-Resultate werden anschließend durch den Discovery Service weiter analysiert. Dieser lokale Dienst stellt das NLP für die weitere Analyse der Texteingaben bereit. Die Eingaben werden zunächst thematisch analysiert und nach zuvor definierten Schlüsselwörter zerlegt und kategorisiert. Etwa-

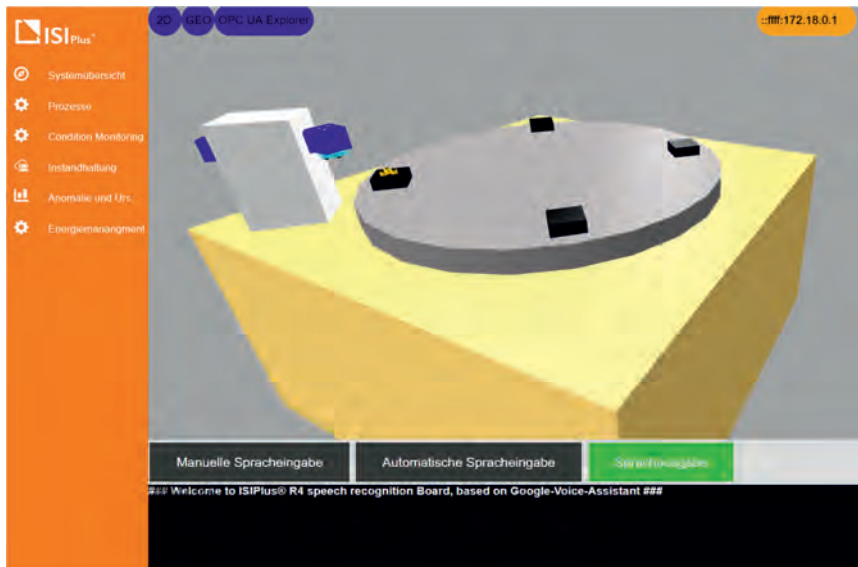


Bild 3: HMI der Schicht 1 - Ein-/Ausgabemaske der Sprach- und Texteingaben sowie Anlagen-Dashboard

ige Synonyme der Schlüsselwörter werden dabei durch eine einfach gehaltene Ontologie berücksichtigt. Hierdurch wird eine korrekte Auswahl der später aufzurufenden Intelligent Services sichergestellt (Service-Mapping). Des Weiteren werden die in den Eingaben enthaltenen Eingangsparameter (z.B. Sensor einer bestimmten Station oder Zeitraum für die Abfrage einer

Zeitreihe) für die Intelligent Services extrahiert. Eingaben im HMI können so direkt den entsprechenden Services zugeordnet werden. Zum Abschluss der Verarbeitung ruft der Discovery Service den entsprechenden Intelligent Service der Schicht zwei mit den ermittelten Eingangsparametern auf. Das NLP des Discovery Service wurde hierbei mit Hilfe des Python-basierten NLP-Frameworks spaCy umgesetzt [17].

Implementierung Intelligent Services: Für die erste prototypische Implementierung der Schicht zwei der Architektur sowie eines geeigneten Umsetzungsbeispiels auf Basis eines CPPS wurden die folgenden Intelligent Services zur Datenermittlung definiert und implementiert:

- (i) Sensorabfrage: Dieser Service ermittelt anhand der Eingabeparameter die gewünschten Prozessdaten einer Komponente, Station oder gesamten Maschine. Dies umfasst dabei zum einen die Ermittlung von live- bzw. aktuellen Ist-Werten wie z.B. der aktuelle Luftdruck einer Station des zugrundeliegenden Systems (hier: Einpressstation Fidget-Spinner). Zum anderen ermöglicht dieser Service eine Aggregation und Mittelung mehrerer Prozesswerte über einen bestimmten Zeitraum (z.B. der durchschnittliche Luftdruck oder Energieverbrauch einer Station oder Komponente der letzten drei Wochen). Die hierfür benötigten historischen Daten sind dabei in einer Zeitreihen-Datenbank hinterlegt, auf die über OPC UA zugegriffen werden kann.
- (ii) Wartungsinformation: Dieser Service ermittelt den gewünschten Wartungsbedarf einer Komponente oder Station auf Basis statischer Wartungsinformationen des Herstellers. Diese Werte stehen bspw. direkt über das OPC UA Informationsmodell zur Verfügung und können semantisch eindeutig ermittelt bzw. direkt abgerufen werden. Darüber hinaus lassen sich hierüber auch dynamische Wartungsinformationen generieren, die zum Zwecke von bspw. Condition Monitoring (CM) und Predictive Maintenance (PM) Szenarien auf dem tatsächlichen Systemzustand beruhen.
- (iii) Customer Handover: Dieser Service übermittelt. aktuelle Betriebs- und/oder Wartungshandbücher, die für das Bedien- und Wartungspersonal relevant sind. Diese liegen. auf einem File-Server bereit. Die benötigten Daten für den Abruf sind in OPC UA modelliert und standardisiert abrufbar.

Alle Intelligent Services sind dabei als eigenständiger Webservice implementiert und werden entsprechend des Discovery Service des HMI mit den gewünschten Parametern aufgerufen.

Jeder Intelligent Service verfügt zudem über einen integrierten OPC UA Client, der den Anlagenzugriff über die offerierten OPC UA Server ermöglicht. Der gezielte Datenabruf erfolgt dabei entweder über vorkonfigurierte NodeSets bzw. Node-IDs, die den entsprechenden Services zugeordnet sind oder dynamisch über die semantische Notation von Schlüsselwörtern bspw. enthalten in einem Server-Dictionary mit entsprechender Referenz auf die Ziel-Knoten im Informationsmodell, die den oder die gewünschten Prozesswerte oder Metadaten repräsentieren.

Jeder Intelligent Service generiert gemäß seinem Zweck eine geeignete Service- bzw. Systemantwort mit dem Ergebnis der Anfrage, die direkt an das HMI zurück übermittelt und dem Personal entsprechend ausgegeben oder im Fall von Spracheingabe vorgelesen wird (Text-To-Speech).

Implementierung Anlagenzugriff: Um einen standardisierten Zugriff auf den Prozess und dessen Werte zu ermöglichen, werden sämtliche Prozessinformationen einer Komponente, Station oder der gesamten Maschine über mehrere OPC UA Server mit einheitlichen Informationsmodellen zur Verfügung gestellt. Die Einpresstation des Umsetzungsbeispiels stellt einen OPC UA Server mit den relevanten Informationen zur Verfügung. Dies umfasst statische Wartungsinformationen als auch die dynamischen Prozesswerte wie bspw. der aktuelle Luftdruck der Station. Die Implementierung der Server-Informationsmodelle basiert dabei auf den OPC UA Companion-Spezifikationen „Device Integration (DI)“ [18] und „Industrie 4.0 Asset Administration Shell (I4AAS)“ [19]. Bild 4 zeigt einen groben Auszug des Informationsmodells der Einpresstation. Zum Zwecke eines einheitlichen Datenabrufs wurde mit Hilfe der I4AAS Companion-Spezifikation exempla-

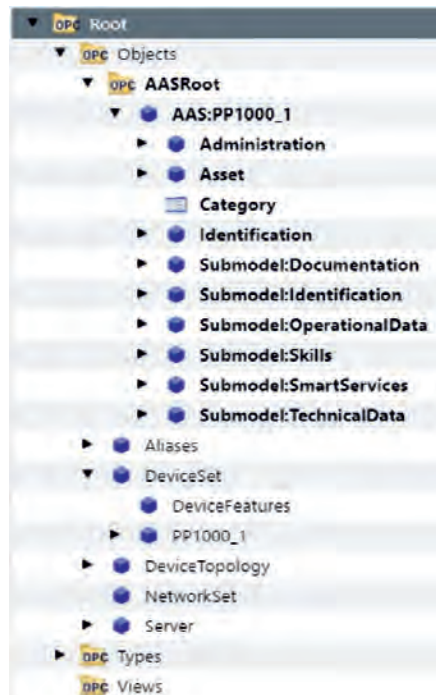


Bild 4: Exemplarisches OPC UA Server-Informationsmodell der Einpresstation mit Companion-Spezifikation DI und I4AAS

risch eine I4.0-Komponente bzw. Verwaltungsschale (VWS/AAS) auf Basis von OPC UA abgebildet. Hierfür wurden prototypisch einige Submodelle (SM) der AAS definiert, die die benötigten Meta- sowie Prozessdaten bereitstellen. Bspw. „Documentation“ für das Bereitstellen von Dokumenten wie z.B. Bedienungs- und Wartungsanleitungen oder „OperationalData“ für die Live- bzw. aktuellen Prozesswerte wie z.B. aktueller Luftdruck der Station.

5. Zusammenfassung

Die in diesem Beitrag vorgestellte prototypische Implementierung einer kognitiven Automations-Architektur für CPPS zeigt das Potential und den Mehrwert neuer Industrie 4.0 Technologien auf. OPC UA als einer der Schlüsseltechnologien der Industrie 4.0 ermöglicht einen einfachen und einheitlichen Anlagenzugriff durch standardisierte Informationsmodelle (Companion-Spezifikationen). Intelligente analytische Services können hierauf aufsetzen und gezielt dazu genutzt werden, Mehrwerte wie z.B. Condition-Monitoring oder Predictive Maintenance zu generieren. Informationsaggregation, Auswertung und Darstellung können durch analytische Services vereinfacht werden, indem genaue Antworten auf konkrete Fragen an ein CPPS generiert und verständlich dargestellt werden können. Die Umsetzung eines innovativen HMI durch die Erweiterung um Interaktionsmöglichkeiten in natürlicher Sprache zeigt, dass die Bedienung und Überwachung komplexer Systeme auch für ungeschultes oder weniger geschultes Personal möglich ist. Das Zusammenspiel und das Integrieren dieser einzelnen Bestandteile zu einer kognitiven 3-Schicht-Architektur zeigt, dass auch künftig immer komplexer werdende Systeme (CPPS) überschaubar und beherrschbar bleiben.

6. Quellen

- [1] M. Broy, E. Geisberger, et al. agendaCPS - Integrierte Forschungsagenda. Cyber-Physical Systems, 2011
- [2] European Commission. Factories of the Future PPP Strategic Multi-annual Roadmap, 2010
- [3] ZVEI Integrierte Technologie-Roadmap Automation 2015+, 2006
- [4] acatech - Abschlussbericht des Arbeitskreises I4.0 „Umsetzungsempfehlungen für das Zukunftsprojekt I4.0“, 2013
- [5] plattform-i40.de, AG Referenzarchitekturen, Standards und Normung, 2019
- [6] Technologie-Netzwerk „Intelligente Technische Systeme OstWestfalenLippe“ (it's OWL), www.its-owl.de/technologie-netzwerk
- [7] BMBF Hightech-Strategie „Industrie 4.0. Innovationen für die Produktion von morgen“, www.bundesregierung.de, 2014
- [8] O. Niggemann: Industrie 4.0 ohne modellbasierte Softwareentwicklung. In: atp 5/2014.
- [9] M. Weyrich, C. Diedrich, A. Fay, M. Wollschlaeger, S. Kowalewski, P. Göhner, B. Vogel-Heuser: Industrie 4.0 am Beispiel einer Verbundanlage. In: atp 7-8/2014.
- [10] Westerholt, J.: Feldgeräte mit Intelligenz, www.prozesstechnik-online.de, Mai 2012
- [11] B. Vogel-Heuser, G. Kegel, K. Bender, and K. Wucherer: Global information architecture for industrial automation. Automatisierungs-technische Praxis (atp), Oldenbourg-Verlag, München, 2009
- [12] O. Givehchi, H. Trsek, and J. Jasperneite: Cloud computing for industrial automation systems, a comprehensive overview. In Proceeding: IEEE 18th Conference on Emerging Technologies & Factory Automation (ETFA), Italy, 2013
- [13] O. Givehchi, J. Imtiaz, J. Jasperneite, H. Trsek: Control-as-a-Service from the Cloud: A Case Study for using Virtualized PLCs. In Proceeding: IEEE Workshop on Factory Communication Systems (WFCS'2014) Toulouse, France, May 2014
- [13] International Electrotechnical Commission, IEC TR 62541-1:2016: OPC Unified Architecture -Part 1: Overview and concepts, 2016
- [14] VDMA OPC UA Demonstrator, <https://robotik.vdma.org/viewer/-/v2article/render/25230733>

- [15] open62541, <https://open62541.org>, abgerufen Mai 2022
- [16] Brandenbourger B., Durand, F.: Design Pattern for Decomposition or Aggregation of Automation Systems into Hierarchy Levels; In: ETFA, Turin, 2018
- [17] spaCy NLP Framework, <https://spacy.io/>, abgerufen Mai 2022
- [18] OPC Foundation: OPC 10000-100 – Device Integration
- [19] OPC Foundation: OPC 30270 - OPC UA for I4 Asset Administration Shell

Data Mining in der Prozesstechnik als Key-Enabler intelligenter Digitaler Zwillinge für eine datengetriebene Optimierung der Prozessführung

Marius Krüger M. Sc., RWTH,
Univ.-Prof. Dr.-Ing. **Birgit Vogel-Heuser**, **Kathrin Land** M. Sc.,
Lehrstuhl für Automatisierung und Informationssysteme,
Technische Universität München;
Gunnar Grim, **Josef Lorenzer**, **Markus Hanf**
Andritz Separation GmbH, Vierkirchen

Kurzfassung

In diesem Beitrag wird aufgezeigt, wie durch Data Mining auf historischen Prozessdaten Ähnlichkeiten zwischen Prozessvariablen beschrieben werden können. Als Anwendungsbeispiel wird die Dynamische Crossflow Filtration (DCF) betrachtet, welche für die Klärung von Wein oder Herstellung von Proteinpulver aus proteinhaltigen Suspensionen eingesetzt wird. Charakteristisch bestehen die Filtrationsprozesse bei diesen Anlagen aus einzelnen Produktionsphasen (Batchprozesse), zwischen denen Reinigungsprogramme von Anlage und Filter ablaufen. Um die Stillstandzeit der Anlage, die Produktqualität sowie den Ressourceneinsatz zu optimieren, müssen die Kosten einer ausführlichen Reinigung gegen den Nutzen für die darauffolgenden Separationsphasen abgewogen werden. Der vorgestellte Data-Mining-Ansatz kann Abhängigkeiten in verkoppelten, aufeinanderfolgenden Batch- und Reinigungsprozessen identifizieren.

Abstract

This paper shows how similarities between process variables can be described by data mining on historical process data. As an application example Dynamic Crossflow Filtration (DCF) is considered, which is used to clarify wine or to produce protein powder from protein-containing suspensions. Characteristically, the filtration processes in these plants consist of individual production phases (batch processes) between which cleaning programs of the plant and the filter run. The cost of extensive cleaning must be weighed against the benefit of subsequent separation phases to optimize plant downtime, product quality and resource utilization. The presented data mining approach can be used to identify dependencies in coupled, sequential batch and cleaning processes.

1. Digitale Transformation von Produktionsprozessen in der Lebensmittelindustrie

Produzierende Unternehmen in der Prozessindustrie werden fortlaufend mit neuen Herausforderungen konfrontiert. Der globale Wettbewerbsdruck steigt kontinuierlich und immer ambitioniertere Nachhaltigkeitsziele bei der Produktion rücken zunehmend stärker in den Fokus. Die Unternehmen sind folglich gezwungen, die Produktionsprozesse ökonomisch effizienter und effektiver zu gestalten, um international konkurrenzfähig zu bleiben. Diese Entwicklungen prägen die umsatzstarke Prozessindustrie am Standort Deutschland maßgeblich. Insbesondere Maschinen- und Anlagenbauer sind bestrebt, die Produktionsanlagen auf ein höheres Automatisierungslevel zu heben und mit Funktionen zur intelligenten Prozessführung auszustatten. Eine durchgängige Prozessdigitalisierung und Prozesstransformation hin zu Cyber-physischen Produktionsprozessen bildet dabei das technologische Fundament für Produktionsprozesse nach den Industrie 4.0-Prinzipien [1]. Infolgedessen wurden bereits zahlreiche Prozessanlagen mit Sensoren und Industriesteuerungen ausgerüstet, um eine stringente Erfassung eines vollständigen Prozessabbildes zu ermöglichen. Obwohl zahlreiche historische Datensätze aus den Produktionsprozessen vorliegen, werden diese kaum für die Prozessoptimierung verwertet [2]. Zumeist erfolgt lediglich eine manuelle Analyse einzelner ausgewählter Datenpunkte oder Prozessabschnitte u.a. zur Regleroptimierung. Die wertvollen Informationen über das Prozessverhalten und Ursache-Wirkungsabhängigkeiten, welche in den Datensätzen enthalten sind, werden in der Regel nicht erschlossen oder liegen nur implizit im Operatorwissen vor. Die immensen Potentiale einer industriellen Nutzbarmachung der Daten werden nur partiell genutzt [3]. Hardwaretechnisch und regelungstechnisch sind die Optimierungspotentiale von Prozessanlagen bereits nahezu ausgeschöpft, sodass die Overall-Equipment-Efficiency (OEE) der Anlagen zumeist nur noch durch eine optimierte Prozessführung erhöht werden kann. Ein vielversprechender Ansatz zur Entwicklung intelligenter Prozessführungsmechanismen ist die Kombination eines Digitalen Zwillings mit Methoden der künstlichen Intelligenz. Hierdurch können Prozesse in Echtzeit überwacht, Optimierungspotentiale erkannt und geeignete Entscheidungsregeln zur Anpassung der Prozessführungsstrategie abgeleitet werden. Eine systematische Erfassung und strukturelle Analyse von historischen Prozessdaten bilden dabei das technologische Fundament von KI-Anwendungen. Durch Data Mining auf historischen Prozessdaten können unbekannte und sich ändernde Wirkungszusammenhänge zwischen Prozessvariablen erkannt und Optimierungspotentiale aufgedeckt werden. Speziell bei den Produktionsprozessen in der Lebensmittelindustrie offenbaren sich die immensen Hürden aber auch Vorteile einer datengetriebenen Strategie einer intelligenten Anpassung der Prozessführung. Separationsprozesse zur Herstellung von Molkereiprodukten, Klärung von Wein

oder industriellen Fermentation, wie z.B. die Dynamische Crossflow Filtration (DCF), sind aufgrund einer Vielzahl an untereinander abhängigen, dynamischen Prozessvariablen anspruchsvoll zu regeln und erfordern Erfahrungswissen qualifizierter Maschinenführer:innen. Eine nur geringfügige OEE-Steigerung führt aufgrund der großen Produktionsmengen bereits zu erheblichen Kosten- und Nachhaltigkeitssteigerungen. Im BMEL-Projekt InSeLDiP (Intensivierung von Separationsprozessen in der Lebensmittelindustrie durch Digitale Zwillinge und intelligente Prozesskontrolle) werden Ansätze zur Online-Optimierung der Prozessführung während der Produktionszeit in einem Konsortium aus produzierenden Unternehmen, Maschinen- und Anlagenbauer sowie universitären Forschungseinrichtungen erforscht. Dazu werden mechanistische Modelle mit datengetriebenen Verfahren kombiniert. Der vorliegende Beitrag stellt erste Konzepte und Ansätze in diesem Projekt heraus, inwiefern Data Mining Methoden auf Datensätze von DCF-Anlagen angewandt werden können, um Prozessabhängigkeiten zu identifizieren.

2. Stand der Technik bei Dynamischen Crossflow Filtrationsanlagen

Bei der DCF handelt es sich um ein industrielles Separationsverfahren zur Trennung von hoch viskosen Suspensionen. Dabei strömt eine Suspension (Feed) turbulent an rotierenden Filterscheiben entlang, sodass sich Retentat (wird am Filter zurückgehalten und verbleibt im Feed/Suspension) und Permeat (passiert die Filtermembran) trennen. Diese Technologie wird beispielsweise zur industriellen Herstellung von Proteinpulver (Permeat) aus proteinhaltigen Suspensionen (Feed) angewandt. Ein wesentlicher Vorteil von DCF-Anlagen gegenüber herkömmlichen Filtrationsanlagen liegt in der Art und Weise, wie die Deckschichtbildung minimiert wird. Während bei konventioneller Crossflow Filtration das Retentat mit hoher Fließgeschwindigkeit zirkuliert werden muss, kann bei der DCF durch die rotierenden Filterscheiben hoch konzentriertes Retentat ohne Zirkulation abgetrennt werden. Eine Verblockung der Membran ist charakteristisch für Filtrationsprozesse und wächst mit zunehmender Produktionszeit [4]. Bestandteile der Suspension setzen sich dabei an der Oberseite der Filtermembran ab, sodass sich die Filtereigenschaften kontinuierlich verschlechtern [5]. Die Anlage und insbesondere die Filtermembranen müssen zwischen mehreren Batches teilweise aufwendig gereinigt werden, um eine gleichbleibend hohe Produktqualität im Permeat sowie eine ressourceneffiziente Separation der Bestandteile zu gewährleisten.

3. Aktuelle Praxis in der Prozessführung

DCF besitzen eine Vielzahl an untereinander gekoppelten sowie zeitlich abhängigen Prozessvariablen und externen Stell- und Einflussgrößen auf den Prozess. Die Auslegung einer optimalen Prozessführungsstrategie in Hinblick auf Ressourceneffizienz, Filtrationsausbeute, Energieaufwand, Anlagenverschleiß und Ertrag führt somit zu einem multikriteriellen Optimierungsproblem. Die Lösung einer solchen Optimierungsaufgabe ist teuer. Ob eine statische Lösung innerhalb annehmbarer Toleranzen bei der unbekannten Dynamik der zeitlich aufgelösten Prozessvariablen überhaupt existiert, bleibt zumeist unklar. In der Praxis wird eine Prozessführungsstrategie durch Auswahl einzelner Stellgrößen und Mehrpunktregler umgesetzt: Im Folgenden werden die Prozesszusammenhänge exemplarisch unter Bezug auf die produzierte Gesamtmenge an Permeat im Batch, den Reinigungsaufwand der Anlage sowie die Flussrate an Permeat skizziert. Ziel ist es, die Filtrationszeit und den Gesamtertrag bei Minimierung des Reinigungsaufwandes zu maximieren. Eine Ertragssteigerung führt zu einer höheren Flussrate und einem erhöhten Reinigungsaufwand der Anlage. Für eine Ertragsreduzierung wird hingegen die Flussrate gesenkt und die notwendige Reinigungsintensität fällt geringer aus. Die optimale Abwägung von Gesamtertrag, Flussrate und Reinigungsaufwand gegeneinander wird in der Praxis durch einen Zweipunktregler mit unterschiedlichen Regelgrößen approximiert. Zu Beginn des Filtrationsprozesses erfolgt eine Regelung der Permeatflussrate auf ein konstantes Niveau. Ab einem gewissen Schwellenwert folgt eine Regelung auf die abfallende Druckdifferenz an der Filtrationsmembran (Transmembrandruck). Die Stellgröße bildet in beiden Fällen die Drehzahl der Feedpumpe.

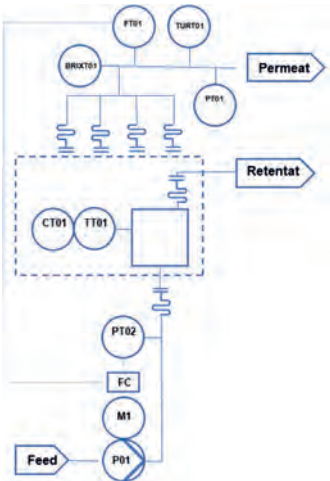
4. Herausforderungen einer datengetriebenen Optimierung der Prozessführung

Die Eigenschaften des DCF-Prozesses legen die Randbedingungen einer Strategie zur Optimierung der Prozessführung maßgeblich fest. Bei der DCF wirken eine Vielzahl an zeitlich veränderlichen Prozessvariablen in unbekannten Zusammenhängen aufeinander. Eine Betrachtung der DCF als Blackbox und Anwendung einer Trial-and-Error-Optimierungsmethode ist in Hinblick auf qualitäts- und sicherheitskritische Anlagenzustände ungeeignet. Infolgedessen müssen Informationen über den Anlagenzustand aus historischen Prozessdaten oder analytischen Modellen des I/O-Verhaltens der Anlage gewonnen und mit Machine-Learning-Ansätzen kombiniert werden [6]. Stand der Technik existiert **kein hinreichendes rigoroses Prozessmodell** von DCF-Anlagen. Ob eine analytische Modellierung vor dem Hintergrund von turbulenten Strömungen und Suspensionen mit zeitlich veränderlichen chemischen und physikalischen Eigenschaften in hinreichender Genauigkeit möglich ist, bleibt fraglich. Ein Ansatz

zur Optimierung von Prozessen ohne rigoroses Prozessmodell muss folglich datenbasiert arbeiten, um Erkenntnisse über die qualitativen und quantitativen Wirkungszusammenhänge berücksichtigen zu können. Um das Anlagenverhalten anhand des Prozessabbildes identifizieren zu können, müssen die vorliegenden historischen Prozessdaten mit Kontextinformationen interpretierbar gemacht werden. Eine zentrale Hürde bei der Kontextualisierung bilden nicht vorhandene oder nur unzureichend dokumentierte **manuelle Eingriffe unmittelbar in den Prozess**. Während manuelle Prozesseingriffe über das HMI grundsätzlich automatisiert dokumentiert werden können, ist die Informationsgrundlage für unmittelbare Veränderungen an der Filtrationssuspension oder an der hardwaretechnischen Beschaffenheit der Prozessanlage oftmals unzureichend. Einzelne Filtrationszyklen innerhalb und zwischen mehreren Batches sind zudem **zeitlich nicht mehr unmittelbar zusammenhängend**, da die Anlage zwecks Wartungsarbeiten, manueller Prozesseingriffe oder aufgrund der Ablaufsteuerung zeitweilig stillsteht. Abnormales Verhalten des Prozesses kann auf Basis der Prozessdaten folglich nicht uneindeutig einem manuellen Prozesseingriff, einer implementierten Logik innerhalb der Ablaufsteuerung oder einer dem Prozess selbst inhärenten Ursache zugeordnet werden. Eine datengetriebene Identifikation von Ursache-Wirkungsbeziehungen und Optimierungspotentialen erfordert, dass sämtliche Phänomene im Prozess durch die Daten erfasst und Prozessvariablen zugeordnet werden können. Aufgrund der unbekannten Prozesszusammenhänge und Dynamiken ist eine **zeitsynchrone Zuordnung von Prozessvariablen** herausfordernd.

5. Data Mining zur Beschreibung von Prozesszusammenhängen

Nachfolgend wird schrittweise ein Vorgehen vorgestellt, um für DCF-Anlagen Abhängigkeiten zwischen Prozessvariablen unter Berücksichtigung der in Kapitel 4 diskutierten Herausforderungen zu beschreiben. Exemplarisch werden die Methoden auf Prozessdaten eines Filtrationsprozesses angewandt und die Erkenntnisse qualitativ diskutiert. Im vorliegenden Beitrag werden exemplarisch die in Bild 1 (rechts) grau hinterlegten Prozessgrößen betrachtet. Es wird ein zeitlicher Teilverlauf mit Volumenstromregelung (Regelgröße: konstanter Volumenstrom des Permeats, Stellgröße: Pumpendrehzahl am Motor M1, vgl. Bild 1) analysiert. Der Regler ist derart ausgelegt, dass sich ein konstanter Volumenstrom im Permeat mit konstantem Transmembrandruck-Anstieg einstellen soll. Die Datensätze in diesem Beitrag werden aufgrund von Geheimhaltungsvereinbarungen anonymisiert und daher ohne Einheiten und Dimensionen dargestellt.



Mess- stelle	Prozessgröße	Prozessteil	Ursprung
TT01	Temperatur	Retentat	I/O, Messstelle
CT01	Leitfähigkeit	Retentat	I/O, Messstelle
BRIXT01	Dichte	Permeat	I/O, Messstelle
PT02	Druck	Feed	I/O, Messstelle
PT01	Druck	Permeat	I/O, Messstelle
-	Transmembran- druck (TMP)	Filter	PT02 – PT01
-	Drehzahl	Pumpe P01	I/O, SPS

Bild 1: Vereinfachtes R&I-Fließbild des Filtrationsprozesses mit Durchflussregelung (links) und Prozessgrößen (rechts).

5.1 Datenerhebung und –vorverarbeitung

Industriell etablierte Austauschformate für zeitlich aufgelöste Prozessdaten existieren in Form von csv oder ASCII, bei welchen die Werte der Variablen einem textuellen Zeitstempel tabellarisch zugeordnet werden. Um die zeitliche Dimension maschinell bearbeiten und interpretieren zu können, werden alle historischen Daten ins performante Hierarchical Data Format (HDF5) unter Umrechnung der Zeitstempel in die zugehörige Unixzeit nach POSIX-Standard überführt. Durch die Zählweise der vergangenen Zeit im Unix-Zeitsystem im Sekundentakt zu einem festen Referenzzeitpunkt werden die Zeitstempel mathematisch leicht händelbar. Die Daten im HDF5-Format werden anhand von logikbasierten Regeln auf doppelte Einträge und Plausibilität geprüft und vorverarbeitet. Jede Prozessvariable erhält hierdurch eine separate zeitliche Zuordnung, sodass im csv-Format entstehende Lücken in der Datenaufnahme aufgrund leicht verschiedener Zeitstempel für verschieden Prozessvariablen vermieden werden können und die Datenlast reduziert werden kann. Große Datenmenge werden hierdurch händelbar, sodass Big Data Methoden anwendbar sind.

5.2 Segmentierung der Zeitreihen nach Anlagenmodifikationen

Es wird angenommen, dass die Beschaffenheit und die Sauberkeit der Membran einen wesentlichen Einfluss auf die Filtrationseigenschaften haben. Anlagen mit verschiedener anfänglicher Verblockung und Verunreinigung der Membran besitzen logischerweise eine unterschiedliche Filtrationsgüte. Folgerichtig werden die zeitlich aufgelösten Prozessvariablen nach den Reinigungs- und Wartungszuständen in der Ablaufsteuerung unterteilt, um die über die jeweilige Produktionszeit gemittelte Prozessgüte zwischen zwei Anlagenmodifikationen betrachten zu können. Basierend auf dieser Segmentierung können pro Anlage und pro Jahr ca. 100 – 200 Zeitreihen extrahiert werden. In Bild 2 wird exemplarisch der durch obiges Vorgehen segmentierte Druckabfall an der Filtrationsmembran (TMP) zwischen zwei Reinigungsvorgängen an der DCF abgebildet.

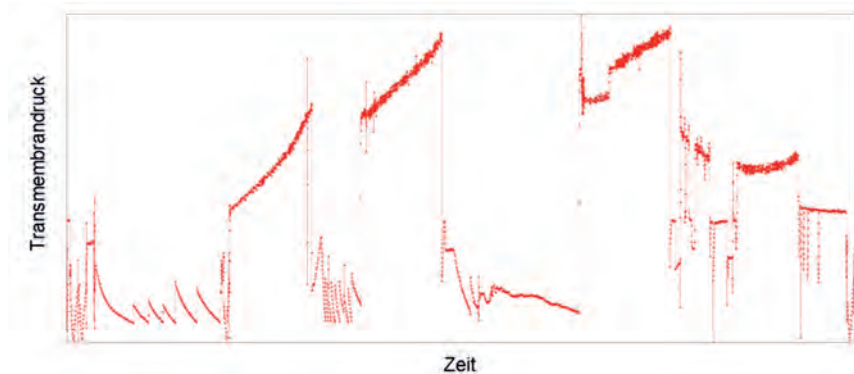


Bild 2: Zeitlicher Verlauf des Transmembrandrucks zwischen zwei Reinigungsvorgängen einer DCF-Anlage.

Unter Anwendung von Expertenwissen und Betrachtung der Engineeringdaten der DCF-Anlage kann der Verlauf in Bild 2 kontextualisiert werden. Die zahlreichen Sprünge zwischen abschnittsweise ähnlichen zeitlichen Abschnitten können auf Stillstandzeiten der Anlage, manuelle Eingriffe und die einzelnen Zustände in der Ablaufsteuerung zurückgeführt werden. Nach jeder Stillstandzeit und Wiederanfahren der Anlage kommt es zudem zu Einschwingvorgängen in den Regelkreisen. Innerhalb der in Bild 2 exemplarisch veranschaulichten Sequenzen werden außerdem mehrere Batch-Filtrationsprozesse abgebildet. Eine weitere Segmentierung der Zeitreihen nach Filtrationszyklen mit eingeschwungenem Regler ist notwendig.

5.3 Segmentierung der Zeitreihen nach Filtrationszyklen

Aus den nach den Anlagenmodifikationen unterteilten Zeitreihen werden anhand des Verlaufes der Regelgrößen und der Informationen aus der Ablaufsteuerung jene Zeitbereiche extrahiert, bei welchen mit eingeschwungenem Regler filtriert wird. Die einzelnen Filtrationsbereiche werden zeitlich fortlaufend aneinandergeschnitten, um mögliche Sprünge von Prozessgrößen während der Filtration detektieren und die Zeitreihentrends qualitativ beschreiben zu können. In Bild 3 werden exemplarisch die aus Bild 2 extrahierten und zeitlich aneinander gereihten Filtrationsphasen dargestellt.

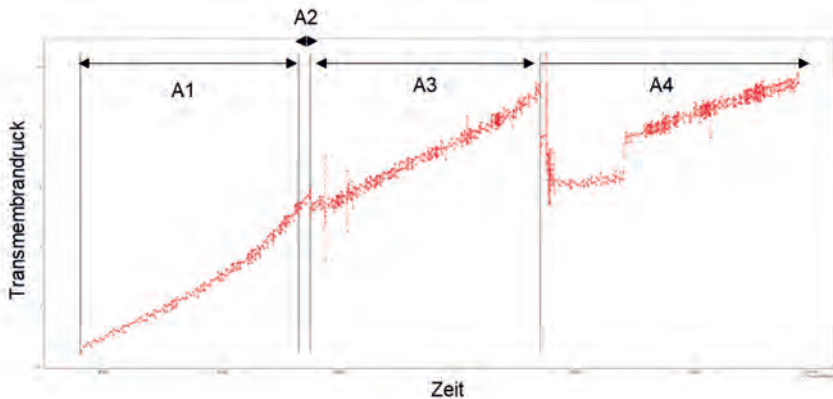


Bild 3: Zusammengesetzter zeitlicher Verlauf des Transmembrandrucks von einzelnen, unterbrochenen Filtrationszyklen zwischen zwei Reinigungsvorgängen der DCF. Unterteilung in zeitlich zusammengehörende Abschnitte (A) A1 bis A4.

Im vorliegenden Beispiel können 4 Filtrationszyklen mit oben beschriebenen Eigenschaften extrahiert und anhand der Ablaufsteuerung interpretiert werden. Abschnitt (A) A1 folgt unmittelbar auf eine Reinigung der Anlage und bildet den Verlauf einer Batch-Filtration ab. Der Batch-Vorgang wird beim Übergang zu A2 unterbrochen und in A2 abgeschlossen. A3 und A4 bilden jeweils einen vollständigen Batchzyklus ohne Unterbrechungen ab. Für A1, A3 und A4 wurden jeweils neue Filtrationssuspensionen in die DCF-Anlage eingefüllt oder Überdruck aus den Rohrleitungen entlassen.

5.4 Qualitative Beschreibung von Trends

Der in Kapitel 5.3 diskutierte Segmentierungsansatz der Zeitreihen in 4 Abschnitte wird analog auf die in Bild 1 (rechts) grau hinterlegten Variablen angewandt. In den nachfolgenden Abbildungen werden die zugehörigen zeitlichen Verläufe dargestellt und Trends sowie Anomalien in Form von Prozesssprüngen beschrieben. Wie in Bild 3 dargestellt, steigt in A1, A2 und A3 der TMP annähernd linear mit konstanter Steigung. In A4 verläuft der TMP stufenweise linear mit geringerer Steigung. Da die Sequenzen in der Ablaufsteuerung bei der Segmentierung der Zeitreihen berücksichtigt werden, kann der in Abschnitt 4 erkennbare sprunghafte Anstieg des TMP entweder auf manuelle, nicht dokumentierte Prozesseingriffe oder als Antwort des TMP auf sprunghafte Änderungen in der Stellgröße (Pumpendrehzahl) zurückgeführt werden. In Bild 4 ist die Drehzahl der Pumpe zur Förderung des Feed an den Filter dargestellt. TMP und Pumpendrehzahl zeigen in den segmentierten Zeitverläufen ein ähnliches Verhalten hinsichtlich der zeitlichen Trends. Der sprunghafte Anstieg in A4, Bild 3, ist qualitativ im zugehörigen Abschnitt A4 in Bild 4 wiederzufinden und liegt damit in einer Anpassung des Sollwertes der Regelgröße begründet. Die Zeitverläufe in Bild 3 und Bild 4 weisen folglich kein abnormales Prozessverhalten auf und sind zeitlich korreliert.

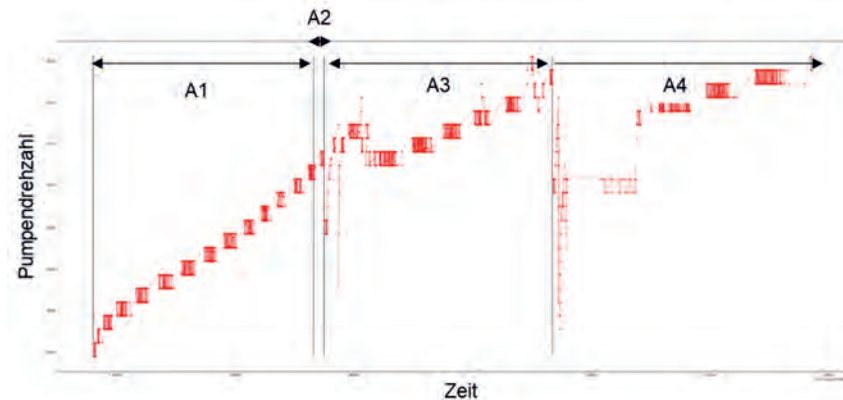


Bild 4: Zusammengesetzter zeitlicher Verlauf der Pumpendrehzahl von einzelnen, unterbrochenen Filtrationszyklen zwischen zwei Reinigungsvorgängen der DCF. Unterteilung in zeitlich zusammengehörende Abschnitte (A) A1 bis A4.

In Bild 5 wird der zeitliche Verlauf der Dichte im Permeat abgebildet. Es können visuell insgesamt zwei verschiedene Kurvenverläufe extrahiert werden. Im zusammengesetzten Zeitbereich aus A1/A2 steigt die Permeatsdichte steil an und läuft dann in Sättigung. In den separaten Abschnitten A3 und A4 sind entgegengesetzte Trendverläufe zu beobachten, bei welchen die gemessene Permeatsdichte zunächst abnimmt und dann unverändert auf konstantem Niveau verbleibt. Das Sättigungsniveau von A3 und A4 liegt unterhalb dessen von A1/A2. Obwohl A1/A2 sowie A3 und A4 separate Batchprozesse darstellen, ist der Kurvenverlauf qualitativ unterschiedlich.

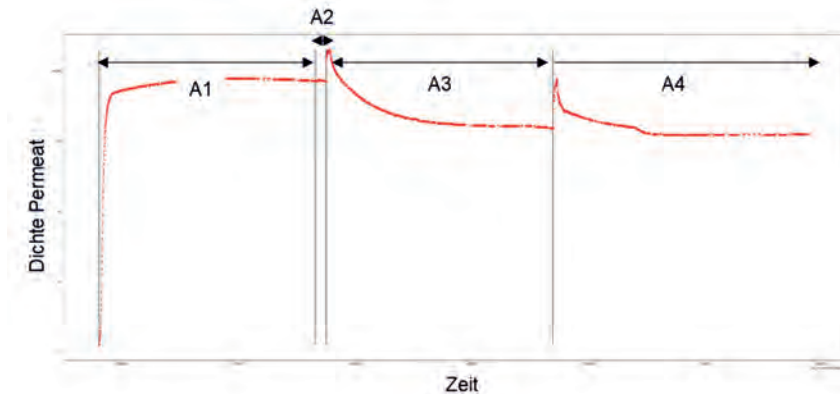


Bild 5: Zusammengesetzter zeitlicher Verlauf der Permeatsdichte von einzelnen, unterbrochenen Filtrationszyklen zwischen zwei Reinigungsvorgängen der DCF. Unterteilung in zeitlich zusammengehörende Abschnitte (A) A1 bis A4.

Aufgrund der analysierten Ähnlichkeiten innerhalb der Zeitreihen kann als erste Erkenntnis gefolgert werden, dass die Änderungsgeschwindigkeit der Dichte im Permeat sowie die Endniveaus die Filtergüte beeinflussen können. Bild 6 zeigt den zeitlichen Verlauf der Leitfähigkeit im Retentat. Die Leitfähigkeit im Retentat dient Stand der Technik zur Dosierung des Reinigungsmediums während der Reinigungsvorgänge. Die Klassifikation der Zeitreihen kann hier analog zu Bild 5 ebenfalls vorgenommen werden. A3 und A4 zeigen ähnliches Trendverhalten jedoch mit verschiedenen Ausgangswerten zu Beginn des jeweiligen Batchs. Beim Übergang des Batchs von A2 zu A3 zu A4 ändert sich die Leitfähigkeit nicht sprunghaft. In A1/A2 wächst die Leitfähigkeit zunächst an und fällt dann wieder ab. Dieses Verhalten bedarf weiterer Untersuchungen. Die Leitfähigkeit sinkt in A3 und A4 konstant, sodass unter Berücksichtigung

des linear wachsenden TMP ein korrelativer Zusammenhang zwischen diesen beiden Prozessgrößen gefolgert werden kann. Die Prozessgröße Leitfähigkeit ist folglich neben der Berechnung der Reinigungsgüte auch für die Optimierung der Prozessgüte von Bedeutung. Der Temperaturverlauf im Retentat wird in Bild 7 dargestellt. Alle Abschnitte zeigen gegen Ende des Batches ein ähnliches Verhalten, da die Temperatur auf einem konstanten Niveau verläuft. In A1/A2 unterscheidet sich der anfängliche Verlauf von A3 und A4, bei welchen die Temperatur zunächst sinkt und dann ähnlich zu A1/A2 steigt. Aufgrund der Ähnlichkeiten der Kurven innerhalb dieser Abschnitte und des Sättigungsverhaltens korreliert die Temperatur im Retentat

wenig mit dem Transmembrandruck.

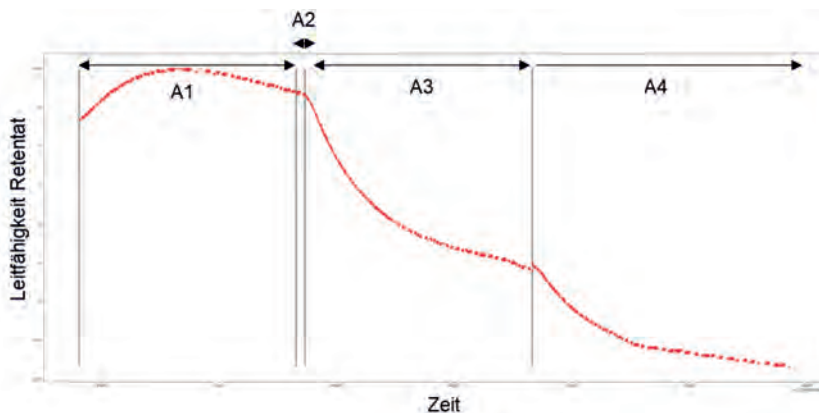


Bild 6: Zusammengesetzter zeitlicher Verlauf der Retentatleitfähigkeit von einzelnen, unterbrochenen Filtrationszyklen zwischen zwei Reinigungsvorgängen der DCF. Unterteilung in zeitlich zusammengehörende Abschnitte (A) A1 bis A4.

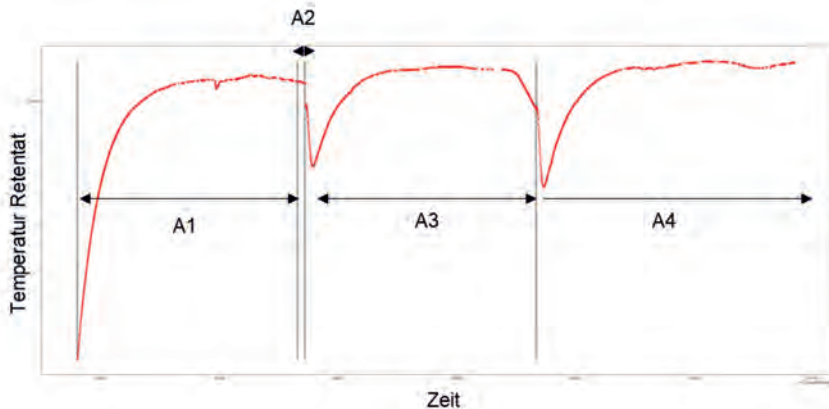


Bild 7: Zusammengesetzter zeitlicher Verlauf der Retentattemperatur von einzelnen, unterbrochenen Filtrationszyklen zwischen zwei Reinigungsvorgängen der DCF. Unterteilung in zeitlich zusammengehörende Abschnitte (A) A1 bis A4.

6. Zusammenfassung und Ausblick

In diesem Beitrag wurde ein Ansatz zur Verarbeitung von Rohdaten und zur Analyse von Abhängigkeiten zwischen Prozessvariablen bei DCF-Anlagen vorgestellt. Anhand der aufgezeigten Methodik können große Datenmengen für die qualitative Analyse und Beschreibung von Ähnlichkeiten in Zeitverläufen von Prozessvariablen aufbereitet und verwertet werden. Durch die qualitative, kontextuale Beschreibung des Trendverhaltens können Empfehlungen gegeben werden, welche Prozessgrößen und welche Kriterien zur Charakteristik der Zeitreihen für die Optimierung der Prozessführung herangezogen werden können. In einem weiteren Schritt sollen die Daten derart verarbeitet werden, dass eine Parametrisierung der Kurvenverläufe und quantitative Beschreibung von Abhängigkeiten ermöglicht werden. Die parametrisierten Kurven, welche basierend auf der vorliegenden Analyse stark korrelieren, können anschließend als Variablen in einem multikriteriellen Optimierungsproblem verwendet werden.

Danksagung

Wir bedanken uns für die Förderung (Fördernummer: 281A503A19) beim Bundesministerium für Ernährung und Landwirtschaft, dem Projektträger Bundesanstalt für Landwirtschaft und Ernährung sowie bei allen Konsortialpartnern.

Literatur

- [1] B. Vogel-Heuser und D. Hess, „Guest Editorial Industry 4.0–Prerequisites and Visions“, *IEEE Trans. Automat. Sci. Eng.*, Jg. 13, Nr. 2, S. 411–413, 2016, doi: 10.1109/TASE.2016.2523639.
- [2] J. Folmer *et al.*, „Big und Smart Data“, *atp*, Jg. 59, 01-02, 2017, doi: 10.17560/atp.v59i01-02.623.
- [3] M. Krüger, B. Vogel-Heuser, I. Weiß und E. Trunzer, „Data-Driven Product Quality Monitoring in Quality-Critical Forming Processes“, *IFAC-PapersOnLine*, Jg. 54, Nr. 4, S. 220–225, 2021, doi: 10.1016/j.ifacol.2021.10.037.
- [4] R. Bott, T. Langeloh und E. Ehrfeld, „Dynamic cross flow filtration“, *Chemical Engineering Journal*, Jg. 80, 1-3, S. 245–249, 2000, doi: 10.1016/S1383-5866(00)00097-6.
- [5] M.-T. Nguyen, „Deckschichtbildung in Kapillarmembranen bei der Querstrom-Mikrofiltration und ihre Beeinflussung durch polymere Flockungsmittel“. Dresden, Techn. Univ., Diss., 2004.
- [6] T. Bismukhametov und J. Jäschke, „Combining machine learning and process engineering physics towards enhanced accuracy and explainability of data-driven models“, *Computers & Chemical Engineering*, Jg. 138, S. 106834, 2020, doi: 10.1016/j.compchemeng.2020.106834.

Künstliche Neuronale Netze (KNN) zur Plausibilisierung verfahrenstechnischer Prozessdaten

Erfahrungsbericht aus der verfahrenstechnischen Industrie

T. Unglaube, Bayer AG, Haan;

L. Däubler, A. Aljandali, Ostfalia Hochschule, Salzgitter

Kurzfassung

Die Prozessleittechnik hat neben Bedien- und Anzeigefunktionen insbesondere auch die Aufgabe, den sicheren Betrieb verfahrenstechnischer Anlagen zu gewährleisten. Dazu ist die Überwachung von Sensor-, Aktor- und Prozesskomponenten notwendig. Der folgende Beitrag zeigt exemplarisch auf Basis realer betrieblicher Prozessdaten, wie mit Künstlichen Neuronalen Netzen (KNN) prozessleittechnische Komponenten während des Betriebs der Anlage plausibilisiert und Anomalien entdeckt werden können. Ausgehend von einer vergleichenden Gegenüberstellung unterschiedlicher Netztypen (MLP, RBF, GRN, LSTM etc.) erfolgte die Auswahl geeigneter Netze zur Plausibilisierung von Echtzeiten. Dabei wurden auch unterschiedliche Ansätze zur Abweichungsbestimmung und Entscheidungsfindung über die Funktionsfähigkeit der zu überprüfenden Komponenten untersucht. Da das Anwendungsgebiet kontinuierlich und batch-betriebene verfahrenstechnische Vielzweckanlagen mit unterschiedlich konfigurierten Produktionsprozessen sind, liegt der Schwerpunkt der vorgestellten Untersuchung in der automatisierten Generierung geeigneter Netzwerkarchitekturen und deren Parametrierung (Training) und Validierung auf Basis von Genauigkeits-, Komplexitäts- und Performancekriterien unter Verwendung der TensorFlow-Bibliothek. In einer abschließenden Bewertung wird gezeigt, ob und wie praxistauglich die gewählte KNN-basierte Plausibilitätsprüfung ist und wie sich dieser Deep-Learning-Ansatz auf Basis generalisierter stochastischer Petrinetze (GSPN) in eine betriebliche Safety-Modellierung einbinden lässt.

1. Einleitung und Motivation

Die Prozessleittechnik hat neben Bedien- und Anzeigefunktionen insbesondere auch die Aufgabe, den sicheren Betrieb verfahrenstechnischer Anlagen zu gewährleisten. Dazu ist die Überwachung von Sensor-, Aktor- und Prozesskomponenten notwendig.

Der folgende Beitrag zeigt exemplarisch auf Basis realer betrieblicher Prozessdaten, wie mit Künstlichen Neuronalen Netzen (KNN) prozessleittechnische Komponenten während des Betriebs der Anlage nicht-invasiv plausibilisiert und Anomalien entdeckt werden können.

1.1 Nicht-invasive Plausibilisierung

Nicht-invasive Plausibilisierungen beschreiben einen theoretischen Ansatz zur online Validierung von PLT-Sicherheitseinrichtungen basierend auf unterschiedlichen Modellierungsansätzen. Für die Modellierung von PLT-Sicherheitseinrichtungen werden folgende Rahmenbedingungen zugrunde gelegt:

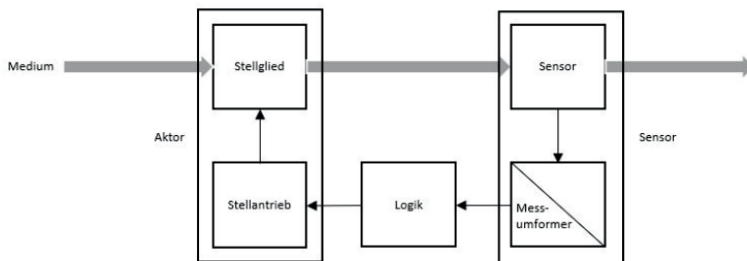


Bild 1: Prinzipschema einer PLT-Sicherheitseinrichtung

- Es werden ausschließlich gefährliche Fehler (λ_{dd} und λ_{du}) berücksichtigt, da diese die sicherheitstechnische Verfügbarkeit beeinflussen.
- Fehler gemeinsamer Ursache werden vernachlässigt.
- Unter Berücksichtigung der Anforderungen aus DIN EN 61522 Abschnitt 11.2.10 werden die Sicherheitseinrichtungen für betriebliche Funktionen mitbenutzt (VDE, DIN EN 61508, 2011).

- Für die Sicherheitseinrichtungen werden ausschließlich betriebsbewährte Geräte eingesetzt, wodurch frühzeitige Ausfälle ausgeschlossen werden können.
- Die Sicherheitseinrichtungen werden vor dem Eintritt von Alterungserscheinungen außer Betrieb genommen, wodurch die Fehlerrate als konstant angenommen werden kann.

Für die Modellierung einer komplexen, funktionalen Systemstruktur wie es bei Prozessleittechnischen Sicherheitseinrichtungen der Fall ist, müssen entsprechende Instanzen unter Berücksichtigung der Gesamtstruktur des Systems verknüpft werden. PLT-Sicherheitseinrichtungen bestehen wie in Bild 1: Prinzipschema einer PLT-Sicherheitseinrichtung aus einem Sensor, einer Speicherprogrammierbaren Steuerung (SPS) und einem Aktor, welches in Bild 2: Petri-netz einer PLT-Sicherheitseinrichtung überführt wurde. Aus Gründen der Veranschaulichung werden diese drei Komponenten im weiteren Vorgehen der Modellierung als ein System definiert, welches zwei Zuständen annehmen kann:

- System in Betrieb
- System in Fehlzustand

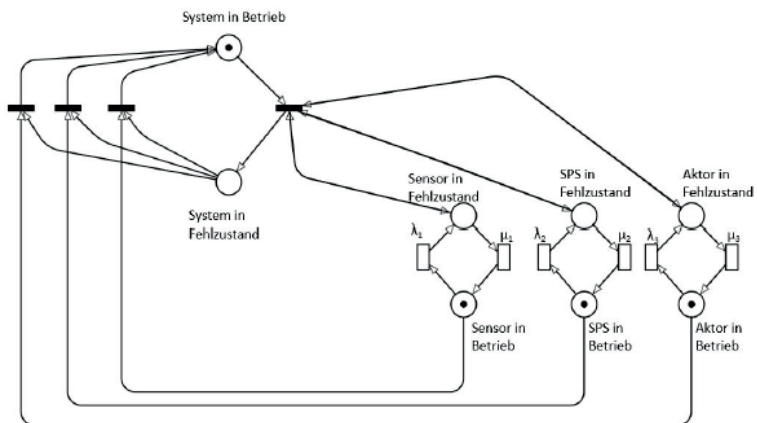


Bild 2: Petri-Netz einer PLT-Sicherheitseinrichtung

Laut DIN EN 61508 wird eine Diagnose als automatische diagnostische Online-Prüfung beschrieben, was es ermöglicht die nicht-invasiven Plausibilisierungen als Diagnose und somit ebenfalls als Teilprüfung aufzufassen (VDE, DIN EN 61508, 2011). Nicht-invasive Plausibilisierungen werden in die automatisierte Produktion integriert und identisch wie Diagnosen automatisch ausgelöst und ausgewertet. Stellt eine nicht-invasive Plausibilisierung eine Abweichung fest, kann dies visualisiert werden und darauf aufbauend ein Alarm ausgelöst werden, was wiederum eine manuelle Reparatur veranlasst. Dies bedeutet, dass eine nicht-invasive Plausibilisierung eine reaktive online Überwachung darstellt und die manuelle Instandsetzung eine Bedarfsorientierte Funktionsprüfung im Sinne der DIN EN 61511 (VDE, Funktionale Sicherheit: Sicherheitstechnische Systeme für die Prozessindustrie, 2012). Reaktiv wird in diesem Zusammenhang interpretiert als, Mitbenutzung der Sicherheitseinrichtung für betriebliche Zwecke, was sich vom derzeitigen Stand der Technik unterscheidet. Durch diesen Ansatz können nicht-invasive Plausibilisierungen online die Sicherheitseinrichtung während des gesamten Produktionsprozesses überwachen, ohne diesen zu beeinflussen. Zusätzlich kann die online Plausibilisierung wesentlich häufiger durchgeführt werden als eine wiederkehrende Funktionsprüfung. Eine „Nicht-invasiven Plausibilisierung“ (NIP) wird definiert als:

Nicht-invasive Plausibilisierung (en: non-invasive plausibility check): Nicht-invasive Plausibilisierungen sind Rückschlüsse auf die sicherheitstechnische Zuverlässigkeit der Funktion eines oder mehrerer Komponenten, durch die gleichzeitige Betrachtung von zwei oder mehreren Informationen über getrennten Informationskanäle und deren Korrelation.

NIP bedeutet demnach die gleichzeitige (also im Sinne Δt) Plausibilisierung von zwei oder mehreren Komponenten, was Bild 3: Definition nicht-invasiver Plausibilisierungen durch Korrelation der Informationen verdeutlicht und exemplarisch in Bild 4: NIP für das Teilmodell Explosion in ein Petrinetz überführt wurde.

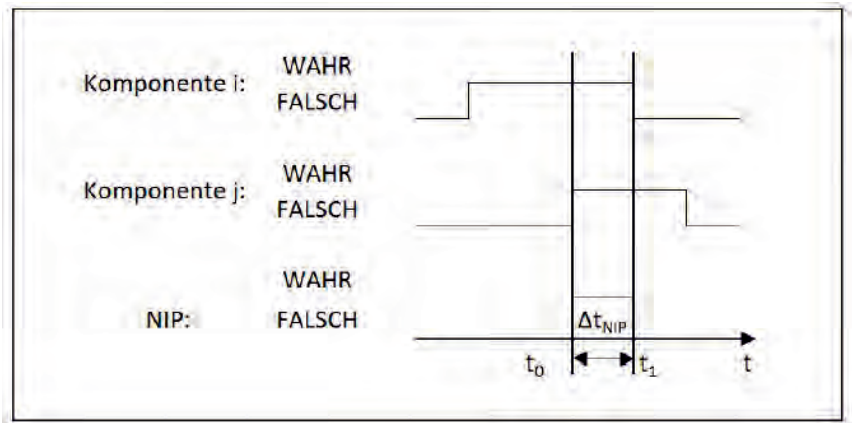


Bild 3: Definition nicht-invasiver Plausibilisierungen

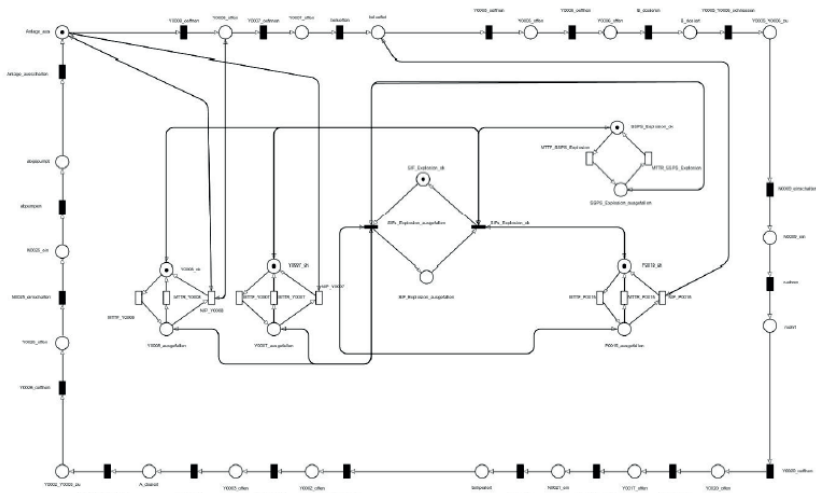


Bild 4: NIP für das Teilmodell Explosion

1.2 Physikalische und Datengetriebene Modellierungsansätze

Für die nicht-invasive Plausibilisierung kommen mit physikalischen und datengetriebenen Verfahren im wesentlichen zwei Modellierungsansätze zur Bereitstellung von Vergleichswerten in Frage (T. Unglaube, Digitale Prüfmethode für die PLT-Wiederholprüfung, 2019). Erste Untersuchungen ergaben, dass im gegebenen Anwendungsfall die zu geringe Sensorausstattung einen physikalischen Plausibilisierungsansatz verhindert, weswegen ein datengetriebener Ansatz mit Künstliche Neuronale Netzen erfolgt.

1.3 Künstliche Neuronale Netze

Der Begriff der künstlichen Intelligenz (KI) beschreibt die Erschaffung intelligenter Maschinen, die lernen, denken, entscheiden und arbeiten können. Als KI werden demnach informationsverarbeitende Systeme bezeichnet, welche in Struktur und Funktionsweise dem Gehirn von Tieren und Menschen nachempfunden sind. Maschinelles Lernen (ML) wird als eine Teilmenge der KI definiert. Ein Aspekt des ML wird als künstliche Neuronale Netze (KNN) beschrieben. KNNs wurden erstmals 1943 von dem Neurophysiologen Warren McCulloch und dem Mathematiker Walter Pitts eingeführt (Abraham, 2005). Aufgrund der erforderlichen Rechenleistung, welche für das Trainieren der Netze benötigt wurde, ließ das Interesse an dem

Forschungsgebiet von KNN bis 1990 nach. Erst Fortschritte auf dem Gebiet der Rechenleistung von Computern brachten das Interesse an KNN zurück (Rashid, 2016).

Elektrische Aktivitäten werden in einem Neuron gesammelt und weitergeleitet, wodurch mittels Interaktionen von Neuronen Informationen verarbeiten, was als grundlegende Recheneinheit von neuronalen Netzen definiert ist (Rudolf Kruse, 2011). In Abhängigkeit eines ausgeschütteten Neurotransmitters kann ein elektrisches Potenzial reduziert oder erhöht werden in einem Neuron. Somit werden Änderungen des elektrischen Potenzials am Zellkörper akkumuliert, und sobald ein Schwellwert erreicht, ist entlang des Axons weitergegeben. Im Nervensystem werden dadurch Informationen im Wesentlichen durch zwei sich ständige ändernde Größen dargestellt. Welches erstens das elektrische Potenzial ist und zweitens die Anzahl der Nervimpulse, was auch als Feuerrate bezeichnet wird. Auf diese Weise repräsentiert KNN eine Verallgemeinerung mathematischer Modelle des biologischen Nervensystems.

In erster Linie kann ein Neuron durch das in Abbildung 2: Mathematische Betrachtung von KNN gezeigte Modell reproduziert werden. Dies ist die einfachste Darstellung von KNNs, bei der eine Reihe von Informationen den Vektor $x=(x_1,x_2,...,x_n)$ zugeführt werden. Jede eingehende Information an den Eingängen zum Neuron ist durch ein Gewicht w gekennzeichnet und wird durch den Vektor $w = (w_1, w_2, \dots, w_n)$ beschrieben. Zusätzlich wird dem Neuron ein Bias-Term hinzugefügt. Manchmal werden konstante Fehler erzeugt, weshalb der Bias-Term dem Neuron hilft diesen Fehler zu kompensieren. Am Ende wird eine Aktivierungsfunktion $g(\cdot)$ an den Ausgang des Neurons angehängt, um dem Ergebnis eine gewisse Nichtlinearität hinzuzufügen.

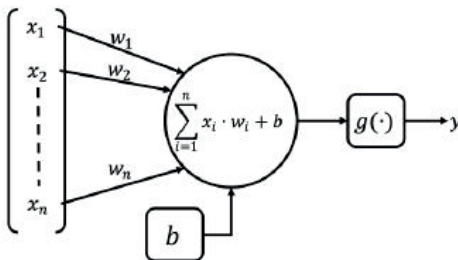


Bild 2: Mathematische Betrachtung von KNN

Die in Abbildung 2: Mathematische Betrachtung von KNN durchgeführte Berechnung wird wie folgt beschrieben:

$$y = g\left(\sum_{i=1}^n x_i * w_i + b\right)$$

2. Plausibilisierungskonzept

Ziel des Plausibilisierungskonzeptes ist es, allgemein gültige Modelle für nicht-invasive Plausibilisierungen für eine verfahrenstechnische Anlagen zu erstellen. Für das sichere Betreiben verfahrenstechnischer Anlagen werden Mittel der Prozessleittechnik herangezogen, welche die Aufgabe haben die Betriebsbedingungen zu erfassen, Steuern und zu Regeln, um eine automatisierte Produktion zu gewährleisten. Wenn die Prozessleittechnik versagt, stellen Sicherheitseinrichtungen sicher, dass die verfahrenstechnische Anlage in einen sicheren Zustand überführt wird und keine Gefahr für Mensch, Umwelt oder Maschine ausgeht. Um den sicheren Zustand der Verfahrensstechnischen Anlage zu gewährleisten, ist die Zuverlässigkeit der Sicherheitseinrichtungen maßgebend. Für eine korrekte Modellierung der Sicherheitseinrichtungen werden im weiteren Verlauf verschiedene nicht-invasive Plausibilisierungen modelliert und untersucht. Eine nicht-invasive Plausibilisierung wird wie bereits erwähnt als Teil einer Funktionsprüfung einer Sicherheitseinrichtung verstanden, welche spezifischen Fehler durch die gleichzeitige Betrachtung von mehreren Informationen und deren Korrelation aufdeckt. Darauf aufbauend wird auf diesen Ansatz untersucht, inwieweit Neuronale Netzwerke und Korrelationsanalysen die Reparaturrate und somit die Prüftiefe präziser beschreiben und somit eine noch genauere Aussage über die Zuverlässigkeit von Sicherheitseinrichtungen getroffen werden kann.

2.1 Verfahrenstechnische Gefahrenfelder

Für das Modell der verfahrenstechnischen Anlage wurde wie in Bild 5: RI-Fließbild eines Rührkessels dargestellt, ein Rührkessel gewählt. Rührkessel sind die am häufigsten eingesetzten Apparaten in verfahrenstechnischen Anlagen, da diese sich für flüssige und flüssig-feste Stoffmischungen eignen und sowohl in Batch-, als auch in Konti-Fahrweise betrieben werden können.

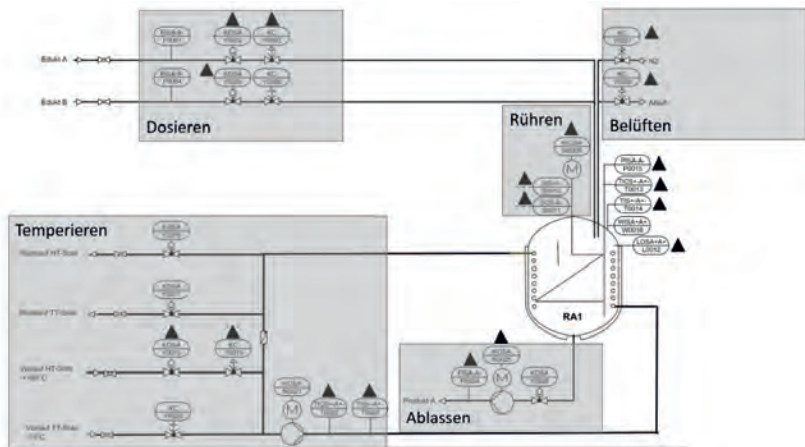


Bild 5: RI-Fließbild eines Rührkessels

Um die Anwendung von nicht-invasiven Plausibilisierungen im Rahmen von Zuverlässigkeitsmodellierungen zu betrachten, wird im Kontext eines Rührkessels die Verfügbarkeit unterschiedlicher PLT Sicherheitseinrichtungen für zuvor definierten Gefahrenfelder betrachtet. Bild 6: Hauptbestandteile des Rührkessel zeigt das durch die Steuerungsfunktionen entsprechende Gefahrenfelder mitigiert werden. So ist der Belüftungsprozess so ausgelegt, dass zündfähige Atmosphäre mittels Zugabe von Inertgas aus dem Rührkessel verdrängt wird und somit das Risiko einer Explosion verringert wird. Ebenso sorgt die Überfüllsonde L0012 dafür, dass während eines Befüll Vorgangs es zu keiner Überfüllung kommt. Durch kontinuierliches Rühren mittels Rührer N0009 wird ein homogenes Stoffgemisch im Rührkessel sichergestellt und eine Akkumulation verhindert. Durch den Dosiervorgang von Edukten wird eine exothermische Reaktion gestartet, welche durch die Mantelkühlung und Regelung mittels T0022, T0023 kontrolliert wird. F0024 erfasst den aktuellen Durchfluss beim Ablassen und schaltet die Ablasspumpe N0025 beim Unterschreiten des Mindestdurchflusses aus um eine Thermal-expansion zu verhindern.

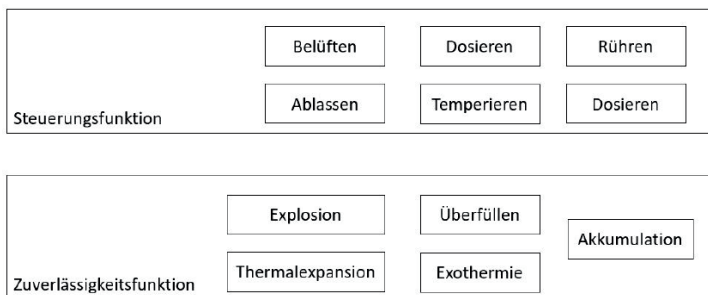


Bild 6: Hauptbestandteile des Rührkessel

2.2 Korrelation und Signalauswahl

Die beschriebenen Gefahrenfelder sind die Grundlage der experimentellen Untersuchung, weshalb in einem ersten Schritt die Korrelation der Sicherheitseinrichtungen mit vorhandenen Sensorinformationen untersucht wurde. Die Intention einer Korrelationsanalyse ist es, mögliche Beziehungen zwischen Variablen zu identifizieren. Aus diesem Zweck werden die zu untersuchenden Variablen in einer Heat-Map grafisch dargestellt. Diese Darstellungsform erleichtert es, Beziehungen zu Erfassen und Extremwerte zu identifizieren. Als Grundlage der Heat-Map wurde die Pearson Methode genutzt, um Aussagen zur Korrelation zu treffen. Die Pearson Methode drückt den Stärkegrad linearer Beziehung zwischen zwei stetigen, zweidimensional normal verteilt streuenden Zufallsvariablen X und Y aus, was beschrieben wird durch:

$$p = \frac{Cov(X,Y)}{\sqrt{var(X) \cdot Var(Y)}}$$

Mit

- Cov (X:Y) = Kovarianz der Zufallsvariablen X und Y
- Var (X) = Varianz der Zufallsvariablen X
- Var (Y) = Varianz der Zufallsvariablen Y

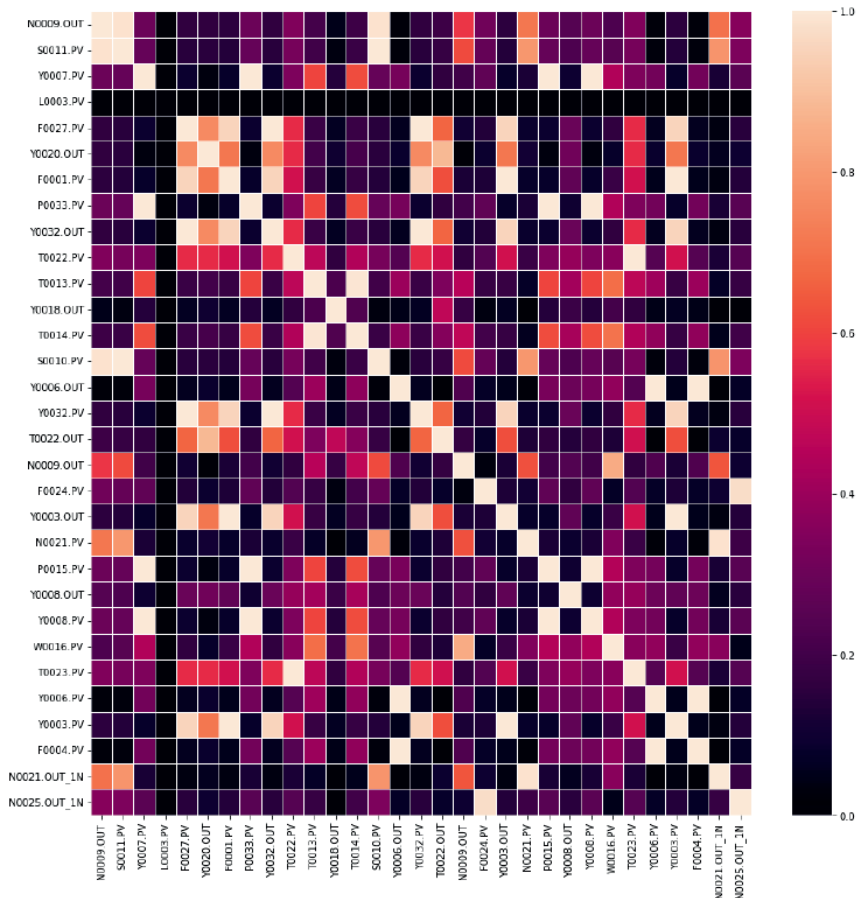


Bild 7: Heat-Map des Rührkessels

Es gilt zu berücksichtigen, dass nicht alle Korrelationen aus Bild 7: Heat-Map des Rührkessels für eine nicht-invasive Plausibilisierung herangezogen werden können, da einige Korrelationen lediglich auf den Steuerungstechnischen Zusammenhang beruhen, ohne einen physikalischen Zusammenhang zu haben. So ist zum Beispiel die Korrelation zwischen Y003 und F0027 nur dadurch gegeben, dass gleichzeitig dosiert wird, die Einlaufstrecken jedoch komplett unabhängig voneinander sind. Die Korrelation zwischen F0001 und Y0003 hingegen ist physika-

lich begründet und korrekt. Durch Expertenwissen wurden für jedes Gefahrenfeld die vorhandenen Korrelationen betrachtet und entsprechende Diagnosemodelle überführt, was Bild 8: Diagnosemodell Exothermie exemplarisch darstellt.

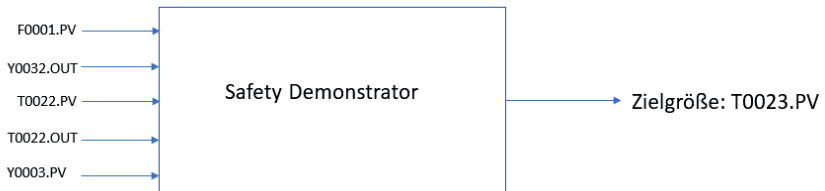


Bild 8: Diagnosemodell Exothermie

3 Experimentelle Untersuchungen

Das Konzept der Nicht-invasiven Plausibilisierung wurde mit der Software-Umgebung "Safety-Demonstrator" umgesetzt und auf Tauglichkeit hin untersucht. Die Umsetzung erfolgte mit Python-Bibliotheken aus dem Bereich Datenanalyse und KNN und mit steuerungstechnischen Tools von Siemens.

3.1 Architektur des Safety-Demonstrators

Zur Umsetzung des Konzeptes zur nicht-invasiven Plausibilisierung von prozessleittechnischen Sicherheitseinrichtung mit künstlichen neuronalen Netzen wurde ein Demonstrator mit der folgenden Architektur und funktionalen Ausstattung entwickelt.

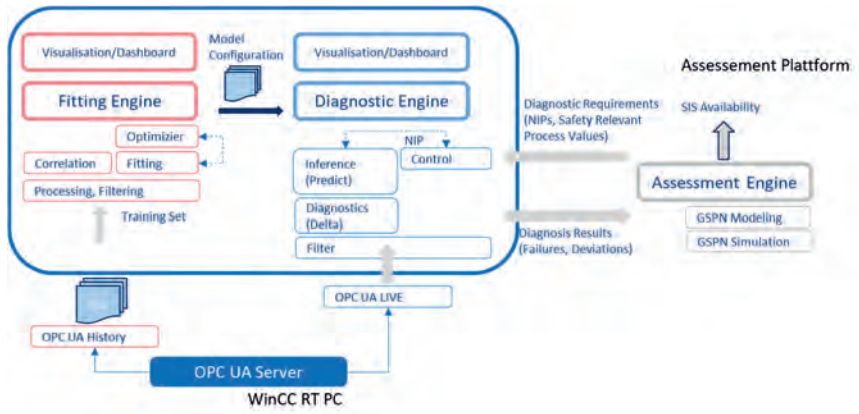


Bild 9: Safety Demonstrator, Architektur und Hauptfunktionen

Der Demonstrator teilt sich in die Assessment-Plattform und die Diagnostic Plattform mit Fitting Engine sowie Diagnostic Engine auf. Die Assessment-Plattform besteht aus generalisierten stochastisch/determinierten Petrinetzen zur Modellierung der Ausfall- und Reparaturerscheinungen in der betrachteten verfahrenstechnischen Anlage, um zu ermitteln, welche Verfügbarkeit eine nicht-invasive überwachte Anlage hat Bild 9: Safety Demonstrator, Architektur und Hauptfunktionen. Die Tatsache, dass Sicherheitseinrichtungen durch Plausibilisierung überwacht werden, wird mit der Reparaturrate μ berücksichtigt. Hierzu sind auch die Rezepte modelliert, da die verfahrenstechnischen Komponenten nicht in jedem Prozessschritt überwacht werden können: ein Rührer kann nur auf korrekten Betrieb überwacht werden, wenn er angesteuert ist und dabei Anomalien zeigt. Die Petrinetz-Modelle der Assessment-Plattform definieren die relevanten Prozessgrößen der Fitting Engine und stellen eine formale Spezifikation für den NIP-basierten Steuerungsalgorithm im Modul "Control" in der Diagnostic Engine dar.

Mit der Fitting Engine werden die Eingangsdaten im csv-Format eingelesen, normalisiert und gefiltert. Die gefilterten Daten dienen dem Training der neuronalen Netze. Das Training wurde mit Funktionen der TensorFlow-Bibliothek umgesetzt und findet mit unterschiedlichen Strukturen und Aktivierungsfunktionen statt, unter denen dann der Optimizier die Struktur mit der geringsten Modellabweichung auswählt. Die ausgewählte Struktur wird als Modellkonfiguration abgespeichert und wird in der Diagnostic Engine zum Prädizieren plausiblerer Prozesswerte genutzt (Inference).

Der eigentliche Diagnose-Schritt besteht in der Berechnung der Abweichung zwischen prädiiziertem und gemessenem Prozesswert durch Differenzbildung und der Bewertung der Abweichungen. Die Kopplung des Safety Demonstrators an die Prozessleittechnik erfolgt über OPC UA und ermöglicht das Auslesen aktueller Prozessdaten im Sekunden-Takt. In der aktuellen Teststufe wird die Prozessleittechnik durch einen lokalen WinCC RT-Rechner ersetzt.

3.3 Ergebnisse und Bewertung

Zu Versuchszwecken wurde für den Gefahrenbereich "Exothermie" die Kühlmitteltemperatur T0023 mit einem MLP prädiziert. Beim Vergleich unterschiedlicher Netztypen hat in diesem Fall ein 2-schichtiges MLP mit 98 Knoten und einer sigmoidalen Aktivierungsfunktion den geringsten Regressionsfehler ergeben.

Die Trainingsdaten bestehen aus den Prozessdaten mit 300 Abtastungen aus einem chemischen Batch-Prozess, die gemäß der Korrelationsuntersuchung die höchsten Korrelationen mit dem gesuchten Zeitwert T0023 aufweisen, was Bild 7: Heat-Map des Rührkessels entnommen werden kann.

- F0001.PV, Y0032.OUT, T0022.PV, T0022.OUT und Y0003.PV (siehe Abschnitt 2.2).

Die folgende Abbildung zeigt den Verlauf der Zielgröße T0023, die sich über einen Temperaturbereich von ca. 30 °C erstreckt.

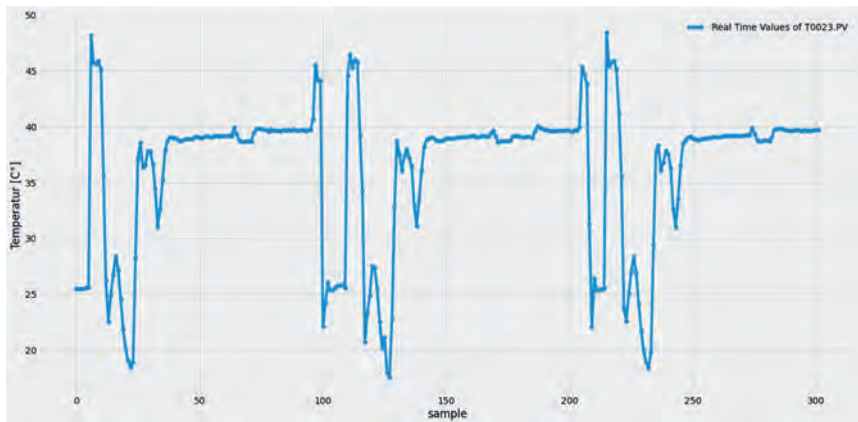


Bild 10: Kühlmitteltemperatur T0023, Gefahrenfeld "Exothermie"

Dieser Temperaturverlauf konnte mit dem trainierten MLP mit geringer Abweichung präziziert werden. Die Abweichung zwischen Trainingsdatensatz und präziziertem Wert ist Bild 11: mittlere Abweichung zwischen präzizierten Werten und Trainingsdaten zu entnehmen.

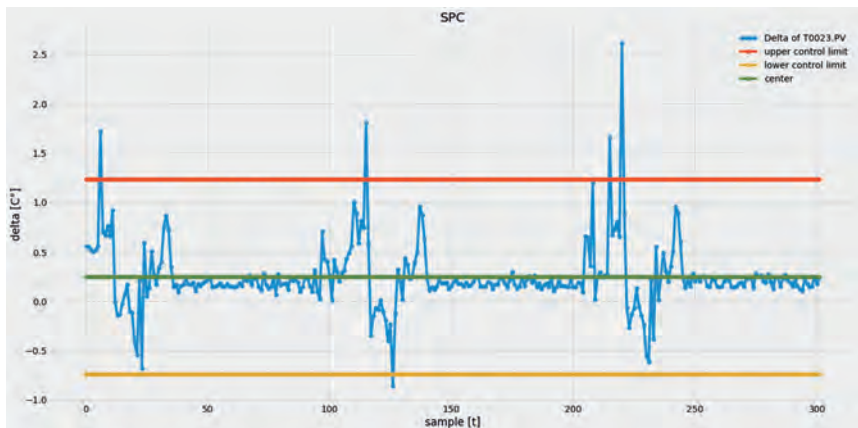


Bild 11: mittlere Abweichung zwischen präzizierten Werten und Trainingsdaten

Die mittlere Abweichung liegt bei ca. 0,25 °C mit einer dreifachen Standardabweichung (99,7% Vertrauensbereich) von ca. 1 °C. Dieser Vertrauensbereich entspricht den oberen und unteren Grenzen von Qualitätsregelkarten der Statistischen Prozesskontrolle (SPC). Vier der

300 Werten liegen außerhalb des Vertrauensbereiches, womit der rechnerisch ermittelte Vertrauensbereich plausibel erscheint. Ab einer Abweichung von ca. 2,5 °C kann mit 99,7%iger Wahrscheinlichkeit ein unplausibler Prozesswert diagnostiziert werden. In der aktuellen Umsetzung zur Abweichungserkennung wird die 7-malige Überschreitung der oberen/unteren Grenzen als Prozessanomalie erkannt.

Vor dem Einsatz des Diagnoseverfahrens mit realen Anlagendaten wurden zu Testzwecken der zu überwachenden T0023-Sensorwert mit einem konstanten Offset verfälscht, was wie erwartet zu entsprechenden Abweichungen zwischen den Sensorwerten und den mit dem MLP prädizierten Plausibilisierungswerten führt, was Bild 12: Signalverläufe und Abweichungen bei manipulierten Sensorwerten sowie Bild 13: Signalverläufe und Abweichungen bei manipulierten Schwellwerten belegen.

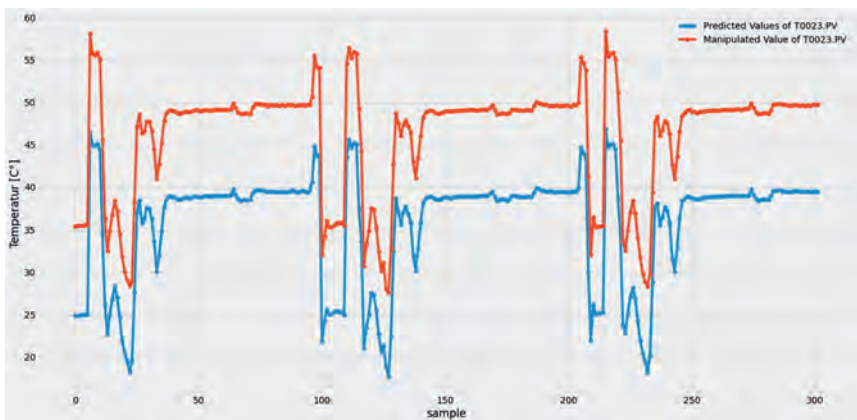


Bild 12: Signalverläufe und Abweichungen bei manipulierten Sensorwerten



Bild 13: Signalverläufe und Abweichungen bei manipulierten Sensorwerten mit Schwellwerten

4 Zusammenfassung, Ausblick

Es kann festgehalten werden, dass der Safety-Demonstrator das Training unterschiedlicher KNN für Regressionsaufgaben und die Plausibilisierung realer Prozessdaten auf Basis der trainierten KNN ermöglicht und Aussagen über die Signalgüte erlaubt. Die Prozesskopplung erfolgt über OPC UA, wodurch eine universelle Schnittstelle für unterschiedliche Leitsysteme etabliert ist. Die Tauglichkeit des Safety-Demonstrators unter realen Anlagen Bedingungen muss noch getestet werden ebenso die automatische Berechnung der Sicherheitstechnischen Zuverlässigkeit mittels Petrinetzen. Aufgrund der nicht-invasiven Plausibilisiert und dem wissenschaftlichen Ansatz zur Bestimmung von Reparaturraten zeigt der Safety-Demonstrator bereits, wie sich durch neue Technologien und den Einzug von Industrie 4.0 die Sicherheitstechnik weiter verbessern lässt und somit eine noch sichere und effizientere Produktion unterstützen.

Literaturverzeichnis

- Abraham, A. (2005). *Artificial neural networks*. Indiana University: Wiley.
- AG, S. (2013). *Einführung und begriffe zur funktionalen Sicherheit von Maschinen und Anlagen*. Siemens AG.
- Anders, U. (1995). *Neuronale Netze in der Ökonometrie*. Mannheim: AnwendungenZentrum für Europäische Wirtschaftsforschung GmbH.
- Felleisen, M. (2001). *Prozessleittechnik für die Verfahrensindustrie*. Oldenbourg: Oldenbourg Industrieverlag.
- Früh, K. F. (2009). *Handbuch der Prozessautomatisierung*. Oldenbourg: Oldenbourg Industrieverlag.
- León, F. P. (2012). *Messtechnik: Systemtheorie für Ingenieure und Informatiker*. Springer.
- Lewis, R. W. (2001). *Modelling Control Systems Using {IEC}* 61499. London: Institution of Electrical Engineers.
- Polke, M. (2009). *Prozeßleittechnik*. Oldenbourg.: Oldenbourg.
- Rashid, T. (2016). *Make Your Own Neural Network*. CreateSpace Independent Publishing Platform.
- Rudolf Kruse, C. B. (2011). *Künstliche neuronale Netze*. Wiesbaden: Vieweg+Teubner Verlag.
- T. Unglaube, L. D. (2019). *Digitale Prüfmethode für die PLT-Wiederholprüfung*. Baden-Baden: Automation 2019.
- T. Unglaube, L. D. (2019). Mit nicht-invasiven Prüfungen Anlagenstillstände vermeiden. *atp magazin*.
- T. Unglaube, L. D. (März 2020). Industrie 4.0 — die selbstüberwachende Anlage. *Technologie-Informationen*.
- VDE, D. D. (2011). *DIN EN 61508*. Frankfurt am Main: VDE VERLAG GMBH.
- VDE, D. D. (2012). *Funktionale Sicherheit: Sicherheitstechnische Systeme für die Prozessindustrie*. Berlin: Beuth Verlag.

Konzept und Realisierung für die automatisierte Erstellung von Verwaltungsschalen in Unternehmen basierend auf bestehenden Daten aus verschiedenen Informationssystemen

Concept and Realization for the automated Creation of Asset Management Shells in Companies based on existing Data from different Information Systems

Igor Garmaev, Torben Miny, Prof. Dr.-Ing. **Tobias Kleinert**,
Lehrstuhl für Informations- und Automatisierungssysteme für die
Prozess- und Werkstofftechnik (IAT), RWTH Aachen University;
Dr.-Ing. **Andreas Schüller, Philipp Bitterlich**,
Yncoris GmbH & Co. KG, Hürth

Kurzfassung

Im Rahmen von Industrie 4.0 und der damit verbundenen digitalen Transformation innerhalb der Prozessketten sollen Unternehmen zukünftig ihre Informationen interoperabel mit Hilfe von Verwaltungsschalen bereitstellen. Hierfür müssen die bereits existierenden Informationen in das neue Datenformat transformiert werden. Um dies möglichst einfach und auch für kleine und mittelständige Unternehmen zu ermöglichen, wird in diesem Beitrag eine Erweiterung des AAS Managers vorgestellt, mit Hilfe dessen diese Transformation unterstützt wird.

Abstract

In the context of Industry 4.0 and the associated digital transformation within the process chains, companies have to provide their information in the future interoperable with the help of asset administration shells. For this purpose, the already existing information must be transformed into the new data format. In order to make this as simple as possible and also possible for small and medium-sized companies, this article presents an extension of the AAS Manager that supports this transformation.

1. Einleitung

Im Rahmen von Industrie 4.0 und der damit verbundenen digitalen Transformation innerhalb der Prozessketten stehen Unternehmen vor dem Problem, wie sie ihre Informationen zukünftig interoperabel bereitstellen wollen. Hierzu gibt es das in [1] eingeführte Konzept der

Verwaltungsschale (VWS), die eine einheitliche Schnittstelle für den Informationsaustausch über Asset-Informationen ermöglicht. Ein zugehöriges Informationsmodell wurde von der Plattform Industrie 4.0 bereits veröffentlicht und wird kontinuierlich weiterentwickelt [2].

Die nächste Phase ist die Anwendung innerhalb der Unternehmen. Im ersten Schritt erstellen Unternehmen eigene VWS für ihre Assets. Um dies zu ermöglichen, wurden verschiedene Software Development Kits [3-8] und zugehörige Editoren [9,10] entwickelt, mit der händisch neue VWS erstellt und mit Informationen gefüllt werden können. Einer dieser Editoren ist der AAS Manager [10], der am Lehrstuhl IAT der RWTH Aachen entwickelt wurde. Da Unternehmen jedoch eine Vielzahl von bestehenden Assets besitzen, deren Informationen bereits in einer unternehmensspezifischen Form in digitalen Datenbanken vorliegen, muss die Erstellung der zugehörigen VWS zukünftig automatisiert mit Informationen aus diesen Systemen erfolgen. Den Autoren ist kein Tool bekannt, dass diese Aufgabe heute schon erfüllt. Aus diesem Grund wurde im Rahmen eines Kooperationsprojekts der AAS Manager um diese Funktionalität erweitert. Das Konzept ist dabei technologieneutral definiert, sodass beliebige Informationssysteme als Quelle zur Erstellung genutzt werden können. Mittels einer geführten Benutzerinteraktion werden die entsprechenden Informationssysteme ausgewählt und die enthaltenen Informationen den Elementen der zugehörigen VWS-Teilmodellen zugewiesen. Anschließend werden diese Informationen vollautomatisch eingelesen und zugehörige Verwaltungsschalen erzeugt. In einer ersten Betrachtung wurden spaltenorientierte Informationssysteme, wie z.B. Excel-Tabellen oder SQL-Datenbanken betrachtet.

Das dahinterliegende Konzept sowie die Umsetzung werden in diesem Beitrag vorgestellt. In Kapitel 2 wird dafür zunächst der Stand der Technik vorgestellt. Der hier betrachtete Use Case wird in Kapitel 3 dargestellt. In Kapitel 4 werden die Anforderungen an das Konzept definiert. Im Kapitel 5 wird das Transformationssystem und ihre Realisierung beschrieben. Kapitel 6 fasst die Validierung und Ergebnisse zusammen.

2. Stand der Technik

In diesem Abschnitt werden die benötigten Konzepte und Vorarbeiten, die für diese Arbeit notwendig sind, vorgestellt. Zunächst wird als Grundlage das Konzept der VWS beschrieben. Danach folgt eine Einführung in den AAS Manager, der im Rahmen dieser Arbeit erweitert wird. Abschließend folgt eine kurze Begriffsdefinition im Kontext von Modelltransformationen.

2.1. Verwaltungsschale

Die Verwaltungsschale (VWS) ist die digitale Repräsentanz eines Assets und wurde 2016 in der DIN SPEC 91345 [1] eingeführt. Ein Asset ist ein "physisches oder logisches Objekt, das

sich im Besitz einer Organisation befindet oder unter dem Gewahrsam einer Organisation steht, und entweder einen wahrgenommenen oder tatsächlichen Wert für die Organisation hat" [2]. Sie dient dem herstellerunabhängigen Austausch von Asset-Informationen für die interoperable Zusammenarbeit zwischen verschiedenen Stakeholdern [11]. Das Ziel ist es, den Zugriff, auf die Informationen sowie die Funktionen eines Assets in allen Phasen des Lebenszyklus zu modellieren und für den Austausch bereitzustellen. 2018 wurde ein erstes technologieneutrales Informationsmodell der VWS von der Plattform Industrie 4.0 veröffentlicht, einschließlich verschiedener Serialisierungsformate (z. B. JSON, XML) [12]. Dies wurde stetig weiterentwickelt und liegt aktuell in der Version 3.0RC01 [2] vor. Zudem ist eine technologie-unabhängige Schnittstellendefinition sowie ein zugehöriges Mapping auf HTTP definiert [13]. Für die Verbreitung und Nutzung wurden zugehörige Software Development Kits [3-8] sowie zugehörige Editoren [9,10] entwickelt.

Das Informationsmodell der VWS ist in Abbildung 1 vereinfacht dargestellt. Es besteht aus den zwei Hauptklassen *AssetAdministrationShell* und *Submodel*. Objekte der *AssetAdministrationShell*-Klasse stellen den Assetbezug her und enthalten eine beliebige Anzahl von Referenzen zu *Submodel*-Objekten, die jeweils einen Aspekt des Assets darstellen. Für den Assetbezug enthalten *AssetAdministrationShell*-Objekte *AssetInformation*-Instanzen, die Informationen über das Asset bereitstellen, z. B. den Identifizierer oder ob es sich um eine Asset-Instanz oder einen Asset-Typ handelt. Die *Submodel*-Objekte bestehen wiederum aus *SubmodelElement*-Objekten. Für verschiedene Informationsarten gibt es verschiedene Klassen, z. B. *Property* für einfache Datentypen, *File* für referenzierte Mediadateien oder *Range* für die Darstellung von Wertebereichen. Es wird weiterhin unterschieden in *Referable* und *Identifiable*-Objekte. *Referable* sind dabei lokal eindeutige Objekte, die mittels des *idShort*-Attributes eindeutig in ihrem Vaterobjekt referenziert werden können (z. B. *Property*, *File*, *SubmodelElement*). *Identifiable*-Objekte hingegen besitzen einen weltweit eindeutigen Identifizierer (z. B. *AssetAdministrationShell*, *Submodel*). Jedes *Referable*-Objekt kann entweder als Vaterobjekt wiederum ein *Referable*-Objekt oder aber ein *Identifiable*-Objekt haben. In einer verschachtelten Struktur, muss als oberstes Vaterobjekt immer ein *Identifiable*-Objekt stehen. Zusätzlich erben *Submodel*- und *SubmodelElement*-Instanzen von *HasSemantic* und erhalten dadurch die Möglichkeit für eine semantische Referenz, z. B. zu einem Wörterbucheintrag.

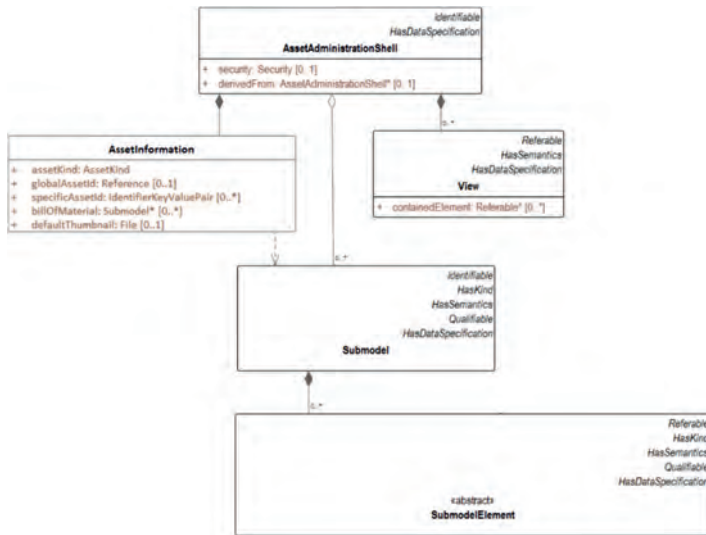


Bild 1: Informationsmodell der Verwaltungsschale nach [2]

2.3. AAS Manager

Für die visuelle Darstellung sowie die Erstellung und Bearbeitung von VWS ohne Programmierkenntnisse wurde am Lehrstuhl IAT der RWTH Aachen ein Editor mit einer grafischen Benutzeroberfläche implementiert, vergleichbar mit dem AASX Package Explorer. Der Editor verwendet dabei das pyi40aas-SDK [8].

Die wichtigsten Bauteile des Editors stellen drei Teile dar, welche in Abbildung 3 zu sehen sind: die Referable-Navigation (oben), der Detailbereich (unten) und die Dialoge (rechts). Im Referable-Navigationsbereich werden geöffnete VWS-Dateien sowie darin enthaltene Referable-Objekte als Objekt-Baum dargestellt. Die Spalten in dem Bereich repräsentieren die einzelnen Attribute der "Referable"-Objekte. In den jeweiligen Zellen werden Attributwerte entsprechender Objekte angezeigt.

Der Detailbereich stellt detaillierte Informationen (Attribute) des im Referable-Navigationsbereich ausgewählten Objekts dar. Elemente in beiden Bereichen können entweder direkt in den jeweiligen Zellen oder über spezielle Dialoge editiert werden. Das Hinzufügen der neuen Elemente erfolgt über Dialoge. Die Eingabefelder dieser Dialoge werden dynamisch während der Laufzeit konstruiert. Dies erfolgt basierend auf einer Analyse der entsprechenden pyi40aas-Klasse, die dem Elementtyp entspricht. Dabei werden

insbesondere passende Eingabe-Widgets für jeden Parameter anhand der Typ-Annotationen in pyi40aas-SDK ausgewählt. Falls ein Objekt über ein Dialog editiert wird, werden die Eingabe-Widgets in dem Dialog mit den Werten aus dem zu editierenden Objekt vorausgefüllt. Aktuell ist es möglich mit dem Editor VWS gemäß des Metamodells der Version 2.0.1 zu erstellen.

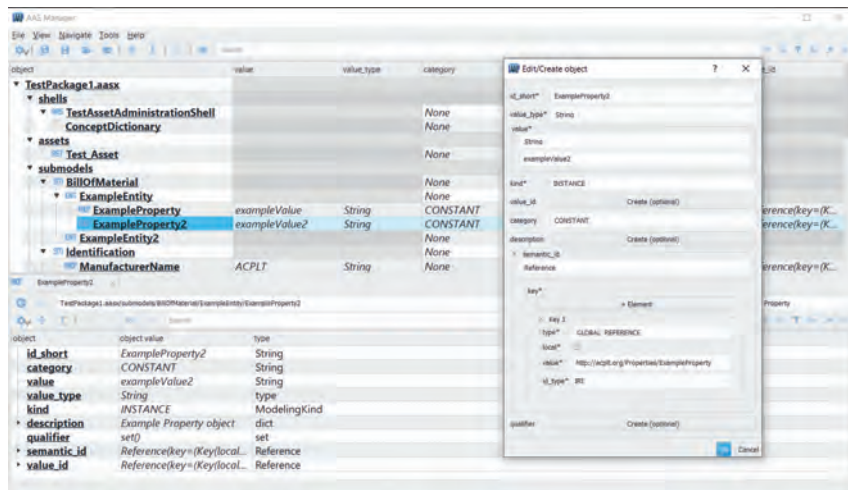


Bild 2: AAS Manager

2.2. Modelltransformation

Nach [14] ist eine Modelltransformation “the automatic generation of a target model from a source model, according to a transformation definition. A transformation definition is a set of transformation rules that together describe how a model in the source language can be transformed into a model in the target language. A transformation rule is a description of how one or more constructs in the source language can be transformed into one or more constructs in the target language“. Im Rahmen dieser Arbeit werden Datenformatssprachen betrachtet. Hierbei kann in die Syntax und die Semantik unterschieden werden. Die Bedeutung der Informationen und die Bewahrung dieser während der Transformation ist elementar [14]. Zur Durchführung der Transformation werden die Regeln auf Ebene der Metamodelle in Transformationsdefinitionen beschrieben. Ein zugehöriges Transformationssystem liest das/die gegebene/n Quell-Modell/e ein, wendet die definierten Regeln an und erstellt das/die zugehörige/n Ziel-Modell/e. Die Beziehungen zwischen den eingeführten Begriffen sind in

Abbildung 2 nach [15] dargestellt. Im Bereich der Verwaltungsschalen gibt es bereits erste Ansätze zur Verwendung von Modelltransformationen [16, 17].

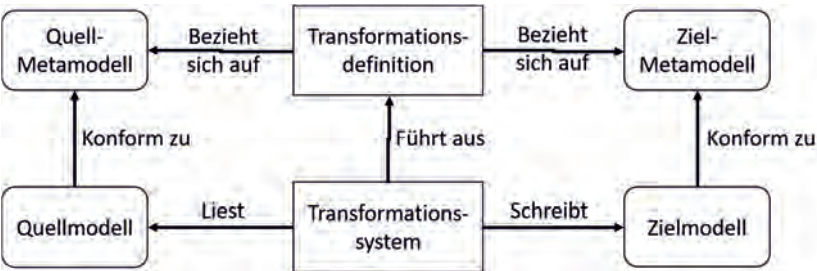


Bild 3: Begriffe der Modelltransformation und deren Beziehung nach [15]

3. Use Case

In diesem Beitrag wird eine Liste von Motoren, welche in Form von einer Excel-Tabelle vorliegt (Abbildung 4), als Anwendungsfall betrachtet. In der Tabelle repräsentieren Reihen die einzelnen Motoren. Jeder Spalteneintrag in der Zeile beschreibt eine Eigenschaft des jeweiligen Motors (z.B. Herstellername, Seriennummer, Nennleistung etc.). Die Tabelle enthält Informationen, die der Anwender momentan zur Verfügung hat.

Der Anwender möchte die aktuelle Datenverwaltung von Assets aus der Tabellenform in Form von VWS mit standardisierten Teilmodellen umwandeln. Um dies zu realisieren, erstellt der Anwender mit dem entwickelten Tool eine Transformationsdefinition und führt ebenfalls mit dem Tool die Modelltransformation für alle Motoren in der Tabelle durch. Anschließend werden aus den jeweiligen Tabellenreihen einzelne VWS erstellt, welche anschließend vom Anwender auf einen Server hochgeladen werden.

Die Nutzung der standardisierten Teilmodelle soll die Interoperabilität der Datenverwaltung sowohl im Unternehmen des Anwenders als auch in der Arbeit mit seinen Partnerunternehmen steigern.

	A	B	C	D	E
1	Hersteller	HerstellerTyp	eClass_Geraetetyp	HerstellerSpezifikation	SerienNr
2	ATB	CD 160	27-02-21-02	160 L2 Y3	789456372
3	Hermetic	Spaltrohrmotor	27-02-27-05	65YA36562	352856398
4	Siemens	1MJ6097	27-02-21-02	4CA36	537425312

Bild 4: Ausschnitt aus Excel-Tabelle mit Motoren

4. Anforderungen

Im Folgenden werden die Anforderungen an die Softwarelösungen erläutert, die erfüllt sein müssen, um die Modelltransformation im oben genannten Use-Case zu realisieren.

1. Unterstützung diverser Datenformate:

In dem beschriebenen Use-Case wird ein spaltenorientiertes Informationssystem als Quellmodell verwendet. Daher besteht die Anforderung in der Unterstützung der entsprechenden Formate dieser Systeme, wie CSV-Dateien, Excel-Tabellen, SQL-Datenbanken als Quellmodelle für Modelltransformation.

2. Struktur der Verwaltungsschale:

In dem Use-Case wird angenommen, dass die Ursprungsdatensysteme spaltenorientierte Systeme darstellen, wo die Informationen in Form von flachen Listen vorliegen. Das Zieldatensystem (VWS) weist im Gegenteil dazu eine hierarchische, verschachtelte Struktur auf. Diese Tatsache macht es notwendig, dass bei der Modelltransformation ins Zielmodell solche Strukturen angemessen behandelt werden müssen.

3. Änderungsmanagement und Wiederverwendbarkeit erstellter Transformationen:

Die Struktur des Quell- bzw. Zielmetamodells oder Informationen im Quell- bzw. Zielmodell können sich ändern, sodass ein neuer Modelltransformationsvorgang mit einer angepassten Transformationsdefinition durchgeführt werden soll. Hier besteht die Anforderung darin, dass neue angepasste Transformationsdefinition nicht von Anfang an neu definiert werden muss. Die Transformationsdefinition soll also speicherbar sein, damit sie später zur Wiederverwendung bzw. Anpassung verfügbar ist.

4. Programmierlose Modelltransformation:

Bei der Erstellung der Transformationsdefinitionen müssen Ingenieure dies auch ohne Programmierkenntnisse erstellen können, sodass sich mehr Personal in Unternehmen mit der Aufgabe befassen können und somit die Kosten für die Umwandlung der Datensysteme in Form von VWS sinkt. Außerdem soll die Transformationsdefinition-Erstellung möglichst schnell und fehlerfrei sein sowie eine geringe Einarbeitungszeit aufweisen.

5. Transformationssystem

In diesem Abschnitt wird das entwickelte Transformationssystem für den im Abschnitt 3 beschriebenen Use-Case vorgestellt. Als erstes wird der allgemeine Workflow des Systems gezeigt. Danach werden einzelne Elemente des Systems und ihre Funktionalitäten, sowie der Aufbau der Transformationsdefinition beschrieben.

5.1. Allgemeiner Workflow

Der allgemeine Workflow ist in Abbildung 5 dargestellt. Das Transformationssystem besteht aus Transformationsinterpretierer und -definierer. Der Transformationsdefinierer liest zuerst das Quellmodell ein, welches z.B. eine CSV oder Excel-Datei sein kann. Anhand der Struktur des eingelesenen Quellmodells und des Zielmetamodells wird von dem Benutzer mit dem Transformationsdefinierer eine Transformationsdefinition erstellt. Anschließend kann diese von dem Transformationsinterpretierer eingelesen und ausgeführt werden. Bei der Anwendung der Regeln aus der Transformationsdefinition werden anhand der eingelesenen Quellmodelldateien Zielmodelldateien erstellt. Die Bedienung des Transformationssystems inklusive der Ein- und Ausgabe der Informationen zur Erstellung der Transformationsdefinitionen, Auswahl der Quellmodelldateien etc. erfolgt mit Hilfe einer grafischen Oberfläche (GUI).

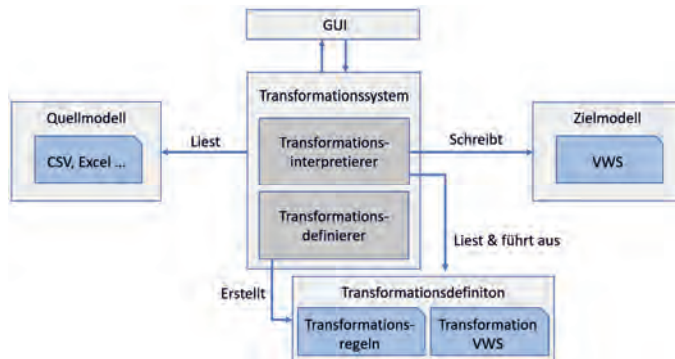


Bild 5: Allgemeine Workflow

5.2. Transformationsdefinierer

Der Transformationsdefinierer wird für die Erstellung der Transformationsdefinition verwendet. Der Nutzer wählt in der GUI die Quellmodelldatei aus, die anschließend von dem Transformationsdefinierer eingelesen wird. Danach kann der Nutzer in der GUI mit üblichen

Instrumenten des AAS Managers die einzelnen *Referable*-Elemente (z.B. Asset, Submodel, SubmodelElement) und somit auch die Struktur der Zielmodell-VWS erstellen. Der Unterschied zu einer gewöhnlichen Erstellung von VWS mit dem Editor besteht hier darin, dass neben den statischen Daten bei der Erstellung von VWS-Elementen der Benutzer für einzelne Attributen auch Spaltenreferenzen eingeben kann. In Abbildung 6 ist ein Dialog zur Erstellung eines *Referable*-Objekts vom Typ *SubmodelElement* dargestellt. Hier ist der Subtyp *Property* ausgewählt. Für das *id_short*-Attribut ist ein statischer Wert „ManufacturerName“ eingegeben, während für *value*-Attribut eine Spaltenreferenz für die Spalte „A“ eingetragen ist. Bei statischen Daten handelt es sich um Daten, die nicht aus Quellmodelldateien importiert werden und in allen erstellten VWS gleichbleiben. Bei Spaltenreferenzen handelt es sich dagegen um „individuelle“ Informationen für jedes Asset. Diese werden für jede VWS aus der referenzierten Spalte der Assetreihe in der Quellmodelldatei importiert. Zum Beispiel wird der Wert für ein „value“-Attribut von dem „Property“-Teilmodellelement für jede Assetreihe aus Spalte „A“ in der Quellmodelldatei importiert. Außerdem können Spaltenreferenzen mit statischen Daten kombiniert werden, sodass bei der Transformationsausführung der Wert teilweise aus Spalteneintrag und teilweise aus statischen Daten bestehen kann, wie „http://company.com/asset/%E%“ welches z.B. nach der Transformationsausführung für die erste Reihe aus der Tabelle in Abbildung 4 zu „http://company.com/asset/563457001H“ wird. Zur Unterscheidung zwischen statischen Daten und Spaltenreferenzen, werden diese Referenzen entsprechend gekennzeichnet und haben folgende Struktur: „\$SPALTENBEZEICHNUNG\$“.

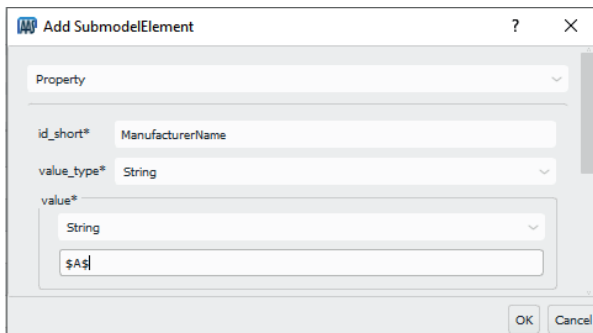


Bild 6: Dialog zur Erstellung eines SubmodelElement

Nachdem der Nutzer bei der Erstellung eines *Referable*-Elements seine Angaben im Dialog bestätigt hat, wird geprüft, ob in den Angaben auch Spaltenreferenzen vorliegen. Falls dies so ist, werden die Referenzen aufgelöst und durch die Einträge in der ersten zu importierenden

Reihe aus der Quellmodelldatei ersetzt. Mit den statischen Angaben des Nutzers sowie den Werten aus den aufgelösten Spaltenreferenzen wird ein *pyi40aas*-Objekt erstellt, welches das VWS-Element in Python repräsentiert.

Die Auflösung der Referenzen dient dazu, dass echte Werte aus den einzelnen Spalten verwendet werden um ein *pyi40aas*-Objekt mit diesen zu erstellen. Somit müssen *pyi40aas*-Klassen nicht für die Verwendung mit Spaltenreferenzen angepasst werden. Außerdem sind in *pyi40aas*-SDK diverse Einschränkungen des VWS-Metamodells implementiert, sodass falls eins solche Einschränkung beim Erstellen des Objekts verletzt wird, ein Fehler auftritt. Der Fehler wird anschließend dem Nutzer in einem Dialog angezeigt.

Da die Werte in den einzelnen Spalten normalerweise vom gleichem Typ sind, werden durch Auflösung der Spaltenreferenzen mit Werten aus der ersten Reihe viele potenzielle Constraint-Verletzungen in zu erstellenden VWS schon in der Phase der Transformationsdefinition geprüft. Ein weiterer Vorteil der Referenzauflösung besteht darin, dass der Nutzer direkt beurteilen kann, ob die Spaltenreferenzen an sinnvollen Stellen angegeben wurden.

Die vom Benutzer angegebenen Spaltenreferenzen gehen allerdings nicht verloren, sondern werden in Transformationsregeln in Form von Mappings zu Transformation-VWS Endpoints gespeichert.

5.3. Transformation-VWS

Wie in 5.2 erwähnt, unterscheidet sich der Prozess der Transformationsdefinitionserstellung von der üblichen VWS-Erstellung für den Nutzer kaum, da dabei ebenfalls eine VWS erstellt wird, welche auch gespeichert werden kann. Diese VWS wird Transformation-VWS genannt. Sie stellt eigentlich eine VWS dar, die beim Import der ersten Reiheneinträge erstellt wird und dient dazu, dass sie die Struktur der Zielmodel-VWS abbildet, sowie statische Daten beinhaltet, die in allen zu erstellenden VWS gleichbleiben.

5.4. Transformationsregeln

Die Transformationsregeln stellen das Mapping der vom Benutzer angegebenen Spaltenreferenzen in den Transformation-VWS in einer anderen Form dar. In der hier vorgestellten Realisierung ist das Mapping in den *Referable*-Elementen verschachtelt aufgebaut (Abbildung 7(a)): In der ersten Ebene stehen Referenzen auf *Referable*-Elemente, welche Spaltenreferenzen in sich haben. In den nächsten Ebenen sind Namen der Attribute aufgeführt, die referenzierte Spalten enthalten. In Abbildung 7 (b) ist dargestellt, wie die Spaltenreferenzen in den Transformationsregeln zu interpretieren sind.

Zusammen mit der Transformation-VWS stellen die Transformationsregeln die Transformationsdefinition dar.



Bild 7: Ausschnitt aus Transformationsregeln und ihre Bedeutung

5.5. Transformationsinterpretierer

Wenn die Transformationsdefinition fertig ist, kann der Nutzer mit dem Transformationsinterpretierer die VWS anhand der Daten in der Quellmodelldatei erstellen. Dabei nimmt der Transformationsinterpretierer die Quellmodelldatei und iteriert über jede Reihe. Für jede Reihe wird eine VWS erstellt, in die alle *Identifiable*-Elemente aus der Transformation-VWS kopiert werden. Beim Kopieren wird allerdings überprüft, ob das kopierte *Identifiable*-Element bzw. in dem Element enthaltene *Referable*-Elemente in Transformationsregeln vorkommen. Falls dies der Fall ist, werden für alle in den Transformationsregeln definierten Spaltenreferenzen in dem *Identifiable*-Element die vorgesehene Spaltenreferenzen für die aktuelle Reihe aufgelöst. Auf Basis der aufgelösten Werten wird ein neues *Identifiable*-Element erstellt und in die VWS übernommen.

Wenn alle *Identifiable*-Elemente aus der Transformation-VWS in die VWS kopiert wurden, stellt diese die fertige VWS für das Asset in der aktuellen Reihe dar. Sobald die Iteration durch alle Reihen in der Quellmodelldatei abgeschlossen ist, werden die individuellen VWS für jede Reihe ausgegeben.

6. Validierung und Ergebnisse

Das entwickelte Tool für die Transformation von Daten aus spaltenorientierten Systemen in Form von VWS unterstützt aktuell CSV-Dateien und Excel-Tabellen als Quellmodelldateitypen (Anforderung 1). Die Unterstützung von SQL-Datenbanken wurde aus zeitlichen Gründen nicht implementiert, kann allerdings im Weiteren realisiert werden.

Für die Validierung des Tools wurde der im Abschnitt 3 beschriebene Use Case verwendet. Hierfür wurde mit dem Tool eine Transformationsdefinition erstellt und gespeichert: die Transformation-VWS wurde in einer AASX-Datei und die Transformationsregeln in einer JSON Datei gespeichert. Anschließend wurden die Transformation-VWS und die Regeln in das Tool importiert und konnten dort weiter angepasst werden (Anforderung 3). Diese Transformationsdefinition wurde dann verwendet, um die Motordaten aus der tabellarischen Form in die Form von VWS zu transformieren. Die erstellten VWS wurden am Ende im AAS Manager geöffnet und validiert, ob sie die gewünschte Struktur aufweisen (Anforderung 2) und die erwarteten Daten enthalten. Alle Schritte sind mit Hilfe der GUI durchgeführt worden. Dabei waren keine Programmierfähigkeiten von dem Benutzer erforderlich (Anforderung 4).

Obwohl die meisten Anforderungen für das Tool erfüllt wurden, weist es einige Einschränkungen und somit Verbesserungspotential auf:

- Die für die Transformation verwendeten Tabellen mit Asset-Daten müssen eine bestimmte Form haben, wo jedes Asset durch eine Reihe in einer solchen Tabelle repräsentiert ist.
- Die Tabellen selber müssen von hoher Qualität sein. Zum Beispiel muss, wenn bei der Transformationsdefinition eine Spalte referenziert wird, wo Integer-Werte erwartet werden, dann auch diese Spalte nur Integer-Werte aufweisen.
- Die Transformationsdefinition besteht derzeit aus zwei Dateien (Transformation-VWS, Transformationsregeln). Wenn die Metamodellversion 3.0RC02 spezifiziert ist und von dem AAS Manager unterstützt wird, wäre es möglich Transformationsregeln für jedes Referable-Element der Transformation-VWS direkt im Element im Attribut *extension* zu speichern. Somit könnte man die Transformationsdefinition in einer Datei speichern.

In diesem Paper wurde eine Erweiterung des AAS Manager vorgestellt, mit der es möglich ist, Asset-Daten aus einer tabellarischen Form in die Form von VWS zu transformieren. Die Erweiterung soll kleinen und mittelständigen Unternehmen den Umstieg zur VWS erleichtern.

Referenzen

- [1] DIN SPEC 91345: Reference Architecture Model Industrie 4.0 (RAMI4.0), Standard, DIN - Deutsches Institut für Normung, 2016
- [2] Plattform Industrie 4.0: Details of the Asset Administration Shell Part 1 - The exchange of information between partners in the value chain of Industrie 4.0 (Version 3.0RC01), Bundesministerium für Wirtschaft und Energie, 2020
- [3] G. di Orio, P. Maló and J. Barata, "NOVAAS: A Reference Implementation of Industrie4.0 Asset Administration Shell with best-of-breed practices from IT engineering," IECON 2019 - 45th Annual Conference of the IEEE Industrial Electronics Society, 2019, pp. 5505-5512, doi: 10.1109/IECON.2019.8927081.
- [4] Kuhn, Thomas, Pablo Oliveira Antonio, and Frank Schnicke. "Eclipse BaSyx–Industrie 4.0 einfach machen! Das Open-Source-Betriebssystem für die Industrie 4.0." atp magazin 63.10 (2021): 54-61.
- [5] i40-aas-objects, Accessed: Apr 2022. [Online]. Available: <https://github.com/SAP/i40-aas-objects>
- [6] i40-aas, Accessed: Apr 2022. [Online]. Available: <https://github.com/SAP/i40-aas>
- [7] coreAAS, Accessed: Apr 2022. [Online]. Available: <https://github.com/OPCUAUniCT/coreAAS>
- [8] Pyl40AAS, Accessed: Apr 2022. [Online]. Available: <https://git.rwth-aachen.de/acplt/pyi40aaa>
- [9] AASX Package Explorer, Accessed: Apr 2022. [Online]. Available: <https://github.com/admin-shell/aasx-package-explorer>
- [10] AAS Manager, Accessed: Apr 2022. [Online]. Available: https://github.com/zrgt/aas_manager
- [11] Boss B., Bader S., Orzelski A., Hoffmeister M.: VWS. In: ten Hompel M., Vogel-Heuser B., Bauernhansl T. (eds) Handbuch Industrie 4.0. Springer Reference Technik. Springer Vieweg, Berlin, Heidelberg, 2018
- [12] Plattform Industrie 4.0: Details of the Asset Administration Shell Part 1 - The exchange of information between partners in the value chain of Industrie 4.0 (Version 1.0), Bundesministerium für Wirtschaft und Energie, 2018
- [13] Plattform Industrie 4.0: Details of the Asset Administration Shell Part 2 - Interoperability at Runtime - Exchanging Information via Application Programming Interfaces (Version 1.0RC01). Bundesministerium für Wirtschaft und Energie, 2020.
- [14] Kleppe A. G. et. al.: MDA explained: the model driven architecture: practice and promise, Addison-Wesley Professional, 2003

- [15] Czarnecki K., Helsen S.: Feature-based survey of model transformation approaches, IBM Systems Journal, 2006
- [16] Platenius-Mohr, Marie, et al. "File-and API-based interoperability of digital twins by model transformation: An IIoT case study using asset administration shell." Future generation computer systems 113 (2020): 94-105.
- [17] Miny, Torben: "Konzept für die semantische Interoperabilität zwischen Informationsmodellen", Dissertation, VDI Verlag GmbH, 2022

Supporting Product Lifecycle Information Management and Distributed Engineering with an Integration Model

Dr. **M. Freund**, Dr. **C. Barth**, Festo SE & Co. KG, Esslingen;
Dr. **P. Juhlin**, ABB Corporate Research Center Germany, Ladenburg;
S. Lehnhoff, EngRoTec Consulting GmbH, Esslingen;
M. Winter, **G. Fachinger**, RWTH Aachen University, Aachen

Abstract

An industrial product's lifecycle involves various interacting phases such as design, engineering, and commissioning. Each of these phases involves different domains, tools, and data models and generates a large set of interdependent artifacts. While the result of an engineering process needs to be complete, consistent, and correct, validating these properties during and at the end of this process is very difficult. This is even more as – depending on the specific target industry or company – engineering processes may vary extremely and can be based on different domains, tools, and data models.

To tackle these challenges, this paper presents an approach for a generic *Integration Model* that can be used to integrate information represented in arbitrary data modeling formats through novel link type definitions. The presented Integration Model allows tracking and automation of all relevant data interactions across different tools during the complete lifecycle of a product and can be used as a basis for the implementation of additional tooling, e.g. to visualize connections between elements from different domains and across different engineering tools. It can also be used to provide a representation of an overall system through interconnection of the various models belonging to its assets.

1. Motivation

A typical industrial product's lifecycle involves various interacting phases such as design, engineering, commissioning, operations, maintenance, and end-of-life. Each of these phases involves different domains and people, relies on different tools and data models, and generates a large set of interdependent artifacts. This is especially true for the engineering phase which can involve multiple engineering disciplines such as mechanical engineering, electrical engineering, control engineering, simulation engineering, PLC programming, robot programming, and communication engineering. While the result of an engineering process needs to be complete, consistent, and correct, validating these properties during and at the

end of this process is very difficult. Consider for example engineering an MTP¹-compliant process module for a modular process plant: engineering typically starts with the design of the process in the form of a piping and instrumentation diagram (P&ID) and results in the generation of the MTP file. In a consistent set of engineering artifacts, each equipment indicated in the P&ID (e.g. a control valve and a flow sensor) needs to be represented as *DataAssembly* of a suitable type in the MTP file (e.g. MonAnaVlv and AnaMon) as illustrated by Fig. 1.

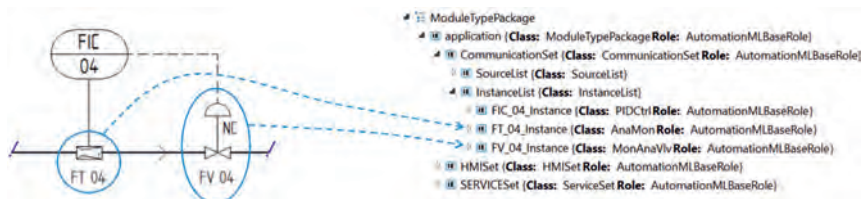


Fig. 1: Representation of P&ID elements (left) in an MTP file (right)

Validating properties like completeness, consistency, and correctness is even more complex as – depending on the specific target industry or company – engineering processes may vary significantly and can be based on different domains, tools, and data models. Considering the example above, the P&ID could, for example, be represented based on a standard like DEXPI² or Pandix³ but could also be represented in a tool- or company-specific data model which makes the definition of validation rules that are generally applicable hard to impossible. The problem is further compounded as inter-model dependencies do not necessarily need to be one-on-one as illustrated by the example above. Instead, a system-level verification might span different elements potentially located in different models: For example, the flow control element from the P&ID in Fig. 1 (FIC04) might be implemented either as a hardware or software module and this decision needs to be reflected in other domains (e.g. electrical CAD, application programming).

In this paper we present an approach for a generic meta-model which can be used to build application-specific *Integration Models*, which keep track of the interdependencies between artifacts from arbitrary data models through *links*. We then present two serializations for the

¹ *Module Type Package* – a standardized interface for process modules in modular process plants [1]

² <https://dexpi.org/>

³ <https://www.plt.rwth-aachen.de/cms/plt/Forschung/Projekte-Archiv/~ejwz/Pandix/>

practical realization of Integration Models: the first is based on AML and the second on AAS. Subsequently, three specific Integration Model use cases are presented and discussed.

2. State of the Art

Today's toolchains for automatic exchange of data during the different phases of a product lifecycle are usually tightly coupled to a specific exchange format. These exchange formats however typically only cover a certain aspect or view and do not regard all phases and domains of the product lifecycle. Examples for such exchange formats are *COLLADA*⁴ for the exchange of 3D CAD models, *DEXPI* for the exchange of piping and instrumentation diagrams in the process industry, and the *Vehicle Electric Container*⁵ (VEC) for the description of the electric sub-system of vehicles.

Currently, there is no comprehensive approach for the management, interlinking, and automatic checking of product data across different phases of an industrial product's lifecycle and the different involved domains. In fact, due to the usual lack of interoperability among the various software systems, information between and within the different phases are usually exchanged manually or informally through printouts, PDF reports, or Excel files, making automatic validation processes impossible. If automated toolchains exist, these are usually tightly coupled to specific tools/data models and/or are restricted to the ecosystem/tool chain of a single company.

There are already some approaches to cover the aggregation, integration, and linking of (engineering) data from different domains across the engineering process such as AutomationML (AML) [2] and the Asset Administration Shell (AAS) [3] which serve as intermediate data formats with modeling mechanisms for exchanging (engineering) information between different software systems and connecting to arbitrary file formats [4, 5]. However, there is no general means of modeling links and dependencies between data from arbitrary domain artifacts. Instead, both AML and AAS provide specialized modeling rules and sub-models, respectively, for the integration of data from specific file formats. Examples for this are the integration of MTP files as sub-model of an AAS [6] and the integration of COLLADA and PLCOpenXML models in AML [2]. Additionally, while both approaches can be used to describe links between different modeling elements, the application of such links to model dependencies between data points from different data models is currently not standardized.

⁴ <https://www.khronos.org/collada/>

⁵ <https://ecad-wiki.prostep.org/specifications/vec/>

Although different languages exist for the description of constraints, dependencies, and transformations between different data models, there is no general approach as these languages are usually limited to a certain modeling technology. For example, OCL⁶ can only be used to model constraints for UML-based models while Schematron⁷ is limited to XML-based models. Consequently, there is no unified approach to define constraints and transformations across all different data models that are relevant during the lifecycle of a product.

3. The Integration Model

As already discussed in Section 1, the modeling technologies, data models, and formats used to accompany the lifecycle of a product are extremely diverse. Furthermore, there is also a multitude of technologies used to specify model transformations and inter-model constraints (cf. Section 2). Hence, the goal of the Integration Model is a broad applicability that does not limit the domains and/or data formats involved in a certain workflow but allows to manage different versions of domain models, constraints, and transformations based on different formats. To keep track of relationships and dependencies between different models, the Integration Model shall in addition allow to describe a set of links between elements of these models. Furthermore, depending on the specific environment that the Integration Model is used in, different serialization formats can be beneficial. Thus, Section 3.1 presents the meta-model for the Integration Model in a technology-independent form by means of a UML class model – hereafter called *Integration Meta-Model*. After that, Sections 3.2 and 3.3 showcase two different possibilities to serialize and exchange the Integration Model based on AutomationML and the Asset Administration Shell. However, it should be noted that other implementations/serializations of the Integration Model – e.g. by means of a relational database or as part of a cloud-based solution (see Section 4.2) – are also possible. Additionally, the presented approach can be extended towards combining different Integration Models via referencing of the relevant elements: For example, an AML-based Integration Model for engineering can be embedded within a complete product lifecycle Integration Model represented in AAS covering documentation, engineering, operations, and others. Therefore, the AML file can be included in the corresponding AAS package file and the relative path of the AML Integration Model elements can be referenced in the top-level AAS Integration Model.

⁶ <https://www.omg.org/spec/OCL/2.4/About-OCL/>

⁷ <https://schematron.com/>

3.1 Integration Meta-Model

The Integration Meta-Model (see Figure 2) allows to combine and integrate heterogeneous data artefacts from different domains into a single model. That way, an Integration Model (i.e., an instance of the Integration Meta-Model) contains the different domains along with all relevant *DomainModelFiles* and respective *DomainModelSchemas*. With this approach, all domain model files remain untouched and are purely linked or copied to the Integration Model repository. This allows using arbitrary engineering tools and formats for the different domains.

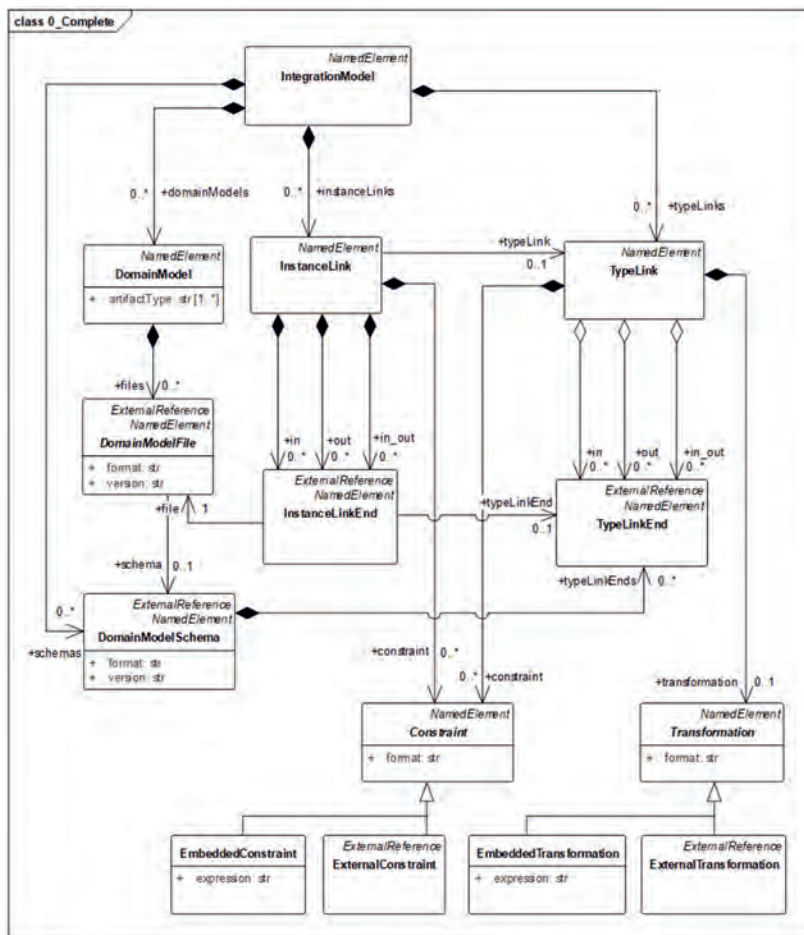


Fig. 2: UML diagram of the Integration Meta-Model

Despite the flexibility to use any given format, it can be beneficial to use standardized file formats or formats of tools which are scoring high on the openness metric as described in [7]. On top of the different domain model files, the meta-model introduces *Links* which are used to describe dependencies and relations. These links can then be used to validate and ensure the consistency of a multi-domain, multi-file engineering project and to transform information automatically from one domain to another.

TypeLinks can be used to annotate general constraints and transformations between different domain models or even within a domain model. For the use case of MTP engineering mentioned above (see Section 1), an example type link could be a connection between a *PCERequest* from the domain of P&ID engineering and a function block from the domain of application engineering resp. a *DataAssembly* from the MTP. In this case, a constraint could specify the requirement of matching names for the connected elements. In general, the complexity of these constraints can vary from simple existence checks to complex multi-parameter relations. Constraints can either be *EmbeddedConstraints* which are formulated inline or *ExternalConstraints* which reference constraints formulated in external files or libraries. The same applies to the transformations. The type links are not associated with specific instances or aspects within the domain model files. However, they can be referenced by *InstanceLinks* which are then used to formulate constraints between specific entities in the domain model files. Considering the example from Figure 1, such an instance link could describe a link between the specific *PCERequest* "FT_04" and the specific data assembly "FT_04_Instance". It is also possible to derive the instance links automatically from type links. However, not each instance link needs to have an associated type link.

In conclusion, the meta-model allows to represent a network of persisted n-ary associations pointing to entities within the domain model files, enabling instance mapping and type mapping on top of the benefits mentioned above.

3.2 Representation in AML

In AML, the Integration Model for a particular application is represented as an *Instance Hierarchy* in a top-level AML file. It is packaged together with relevant referenced files, e.g. schema files, model files, constraint or transformation files, as part of an AML container file (with extension *.amlx). The Integration Meta-Model elements shown in Fig. 2 are represented as AML Role Classes, Interface Classes, and Attribute Types: Role Classes are used for defining object semantics (e.g. *DomainModel* and *TypeLinkEnd*), Interface Classes are used for linking and referencing of objects (e.g. *InstanceLink* or *ExternalConstraint*), and Attribute Types are used for defining object properties and sub-properties (e.g. *Constraint* or *format*).

These elements serve as the basis for constructing an AML-based Integration Model for a given application following a 2-step process of (1) defining the vendor-specific *System Unit Classes* using the predefined elements and (2) instantiating the Integration Model for the application based on the defined classes. An implemented Integration Model with the main elements used for linking a sensor in a Pandix P&ID to an AnaMon in an MTP file (cf. Fig. 1) is shown in Fig. 3.



Fig. 3: Representation of an example Integration Model in AML

3.3 Representation in AAS

Representing the Integration Model as part of an AAS requires the definition of a new submodel including (1) all potential elements and sub-elements based on the AAS meta-model. In

addition, the definition of (2) *SemanticIDs* that make an element unambiguous and machine-interpretable is a crucial step.

Concerning the first aspect, a more or less direct transformation between the UML meta-model and the AAS meta-model has been used, i.e. the hierarchical structure of the submodel is directly derived from the hierarchical structure of the UML model. In addition, the following mapping rules have been applied:

- Each UML class (except the top-level container *IntegrationModel*) is represented as a *SubmodelElementCollection* (SMC).
- If a UML class contains multiple references, each reference is represented by its own SMC within the parent collection.
- Each UML attribute is represented as *Property*.
- UML associations are represented as *ReferenceElement*, *RelationshipElement*, or *AnnotatedRelationshipElement*.

An overview of the resulting AAS submodel for the example from Fig. 1 is shown in Fig. 4.

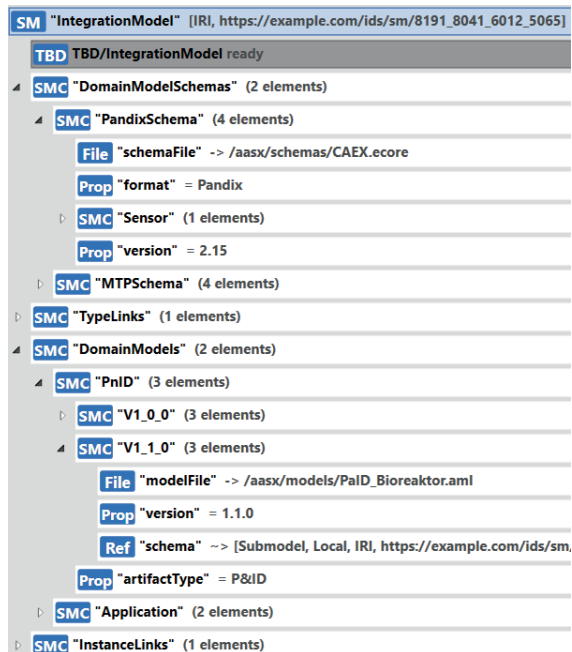


Fig. 4: Representation of an example Integration Model as AAS submodel

Concerning the second aspect, *SemanticIDs* have been derived based on the UML meta-model in the form "http://basys4.de/submodels/integrationmodel/<UMLElementName>", where *UMLElementName* is the name of either a class, attribute, or reference. Where beneficial, one semantic ID is shared for multiple UML elements, e.g. the semantic ID "http://basys4.de/submodels/integrationmodel/Link" for both *TypeLinks* and *InstanceLinks*.

4. Case Studies

This section introduces different case studies that have been carried out to demonstrate and validate the functionality and versatility of the Integration Model.

4.1 Process Module Development

As already introduced in Section 1, engineering an MTP-compliant process module needs to result in an MTP file that is consistent with the P&ID created at the beginning of the engineering process (cf. Fig. 1). Therefore, several consistency rules can be defined. An example for such a rule – implemented with the *Epsilon Validation Language*⁸ – is shown in Figure 5Fig. 3.

```
context PnIDInternalElement {
  guard: self.isAnalogSensor()

  constraint AnalogSensor2AnaMon {
    check {
      var requestID = self.getAttributeValue("requestID");
      var anaMon = getAnaMonWithTagName(requestID);
      return anaMon != null;
    }
    message: "No AnaMon FB found for Sensor " + requestID
  }
}
```

Fig. 5: Sample constraint to check the consistency between a P&ID and an MTP file

For the case study, a P&ID model (represented in Pandix) as well as the resulting MTP file have been added to the Integration Model together with a collection of type links each representing a constraint such as the one from Fig. 5. Evaluating the type links, i.e. checking the consistency between the P&ID and the MTP file, results in a set of instance links that are also persisted in the Integration Model and can for example be used to navigate from an element in the P&ID to an element in the MTP. The resulting Integration Model in two different serializations was already introduced in Figures 3 and 4. Additionally, a Python-based transformation that automatically creates the AAS MTP submodel (cf. [6]) based on an MTP file was also added to the Integration Model. If the transformation is executed, an AAS

⁸ <https://www.eclipse.org/epsilon/doc/evl>

submodel for the contained MTP file is generated and added to the Integration Model as a distinct domain model. This case study shows the versatility of the Integration Model that can handle different modeling technologies for both domain models as well as constraints and transformations.

4.2 Drive-Motor-Load Simulation

A Drive-Motor-Load Integration Model as an AAS Submodel is used to interlink different proprietary and standard models, as illustrated in Fig. 6: drive and motor type selections from ABB *DriveSize*, different load simulation models provided via AAS Simulation submodels including embedded Functional Mockup Units (FMUs), and simulation parameters for cloud-based virtual commissioning via ABB *Virtual Drive*. This Integration Model is used to establish automatic transfers of relevant engineering data and models of drive, motor, and load assets for combined drive-motor-load system simulations. Moreover, it provides a unique system-level representation of the combined drive-motor-load system through the interlinking of the contained asset models, which can be further extended to cover other phases of the system lifecycle, e.g. operations. Further details can be found in [8].

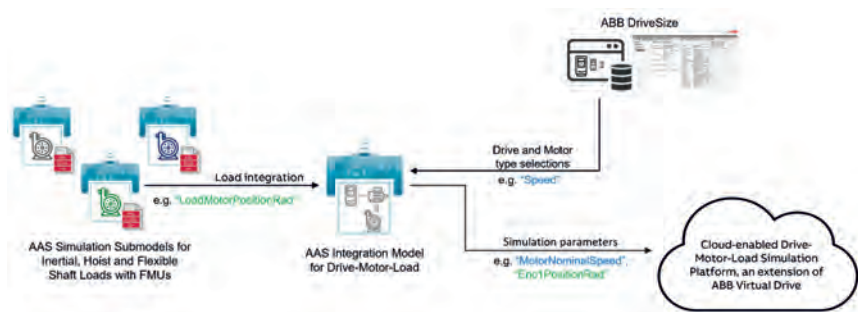


Fig. 6: AAS Integration Model for Drive-Motor-Load linking parameters for Virtual Commissioning via ABB Virtual Drive, with example source to target parameter mappings indicated in same colors

4.3 Coil car

Investigating existing data models from an old engineering project for a coil car (a cart for transporting aluminum coils) highlights the limitations/preconditions for a successful data integration across different domains: While the models for the different domains/views (e.g.

electrical, mechanical, and function view) can be easily added to and managed by an Integration Model, the definition of meaningful constraints and thus the automatic derivation of instance links turned out to be complex to impossible. The reason for this is that although elements like controllers are represented in all domains, the different domains all followed their own naming scheme, did not provide any unique identifier for model elements at all or did not make use of shared IDs across the different domains, and mostly did not contain any other elements or attributes that made them uniquely identifiable. These premises made the identification of *matching* elements from different domains impossible, which is one main premise for the definition of constraints or automatic transformations that usually act on an entity basis. While the Integration Model still allows for a manual definition of links in order to relate elements from different domains, this shows the importance of a unified database and unique, stable identifiers for elements that are (re)used across all domains. The Integration Model can act as the basis for the definition of such a system as it makes all domains involved in a project explicitly represented and can act as a means to overcome domain silos.

5. Conclusion and Outlook

In this paper, we have introduced the Integration Model as an approach to manage all relevant (engineering) data for a project including links between different models/aspects. In addition to the general concept and meta-model, we have provided two serializations for the meta-model based on AutomationML and the Asset Administration Shell. In contrast to existing solutions, the major benefit of the presented approach is that it is not limited to a single data format or a certain modeling technology. Instead, domain models, constraints, and transformations that are based on arbitrary formats/technologies can be incorporated into the domain model.

Of course, the evaluation of constraints as well as the execution of transformation still needs to be implemented in a technology-specific way. However, allowing to represent the results of such algorithms – the links between different models – is still beneficial as it allows a high-level view on the state of a project. The different case studies from Section 4 have shown the manifold ways of applying the Integration Model but have also shown some preconditions for the application of meaningful automatic constraints and transformations like the usage of stable and unique IDs in the different data models involved in a project.

The Integration Model presented in this paper is primarily a basis for additional tooling that can be implemented on top. Such tooling could for example allow to navigate between different editors/domains based on an evaluation of the links inside the Integration Model. That way, it can help communication within a project and aid the goal of consistent project data.

Additionally, it can act as a single source of truth that can for instance be used to manage the unique identifiers for elements that are used within a project. It can thereby also enable a system-level digital twin, i.e., the system's digital representation comprising of its "properties, condition and behavior by means of models, information and data" [9], through interlinking of the different asset information and simulation models across the complete lifecycle.

6. Acknowledgement

The presented work is partially supported by German Federal Ministry of Education and Research in the scope of the BaSys 4.2 project (01IS19022).

- [1] Bernshausen, J. et al.: NAMUR Modul Type Package – Definition. atp edition 2016, 58 (01-02). 2016, S. 72–81
- [2] Drath, R.; "AutomationML: A Practical Guide"; De Gruyter Oldenbourg (2021), ISBN Print 978-3-11-074622-8, EAN E-Book: 9783110746235
- [3] Bedenbender, H., et al.; "Diskussionspapier – Verwaltungsschale in der Praxis"; Federal Ministry for Economic Affairs and Energy (BMWi), Tech. Rep. (2019).
- [4] A. Lüder, N. Schmidt, and R. Drath; "Standardized Information Exchange Within Production System Engineering"; in Multi-Disciplinary Engineering for Cyber-Physical Production Systems, S. Biffl, A. Lüder, and D. Gerhard, Eds. Cham: Springer International Publishing, 2017, pp. 235–257. doi: 10.1007/978-3-319-56345-9_10.
- [5] S. Ananieva and P. Juhlin; "Integriertes Engineering"; Springer return – Magazin für Transformation und Turnaround, no. 5, 2020. [Online]. Available: <https://www.springerprofessional.de/integriertes-engineering/18485892>
- [6] Grüner, S., et al.; „Integration von Module Type Package und Industrie 4.0 Verwaltungsschale“; VDI-Berichte 2392; VDI Wissensforum GmbH (2021).
- [7] M. Barth, R. Drath, A. Fay, F. Zimmer, and K. Eckert, "Evaluation of the openness of automation tools for interoperability in engineering tool chains," in Proceedings of 2012 IEEE 17th International Conference on Emerging Technologies Factory Automation (ETFA 2012), Sep. 2012, pp. 1–8. doi: 10.1109/ETFA.2012.6489542.
- [8] P. Juhlin, A. Karaagac, J. Schlake, S. Grüner, and J. Rückert; "Cloud-enabled Drive-Motor-Load Simulation Platform using Asset Administration Shell and Functional Mockup Units"; in Proceedings of 2022 IEEE 27th International Conference on Emerging Technologies Factory Automation (ETFA 2022), Sep. 2022 (submitted).
- [9] R. Stark, S. Kind and S. Neumeyer; "Innovations in digital modelling for next generation manufacturing system design," CIRP Annals, vol. 66, no. 1, pp. 169-172, 2017.

Vergleichende Untersuchung der Orchestrierung modularer Anlagen in der Prozessindustrie, der Fertigungsindustrie und der produktionsnahen Logistik

Comparative study concerning the orchestration of modular plants in the process industry, the manufacturing industry and production-related logistics

M. Blumenstein, Prof. Dr.-Ing. **A. Fay**, **L. Beers**,
Helmut-Schmidt-Universität, Hamburg;
A. Stutz, Dr.-Ing. **M. Maurmaier**,
Siemens AG, Karlsruhe

Kurzfassung

Für die Orchestrierung modularer Anlagen bestehen in der Prozessindustrie, der Fertigungsindustrie und der produktionsnahen Logistik bereits zum Teil standardisierte Ansätze. Diese verfolgen ähnliche Ziele, unterscheiden sich aber in ihrer Ausgestaltung, was zu Systembrüchen führt. Um eine Vereinheitlichung der Orchestrierung herbeizuführen, verfolgt dieser Beitrag die Idee, das aus der Prozessindustrie bekannte Module Type Package (MTP)-Konzept branchenübergreifend weiterzuentwickeln. Es werden Forschungs- und Standardisierungslücken hinsichtlich einer branchenübergreifenden Nutzung des MTP-Konzepts mit Fokus auf die Orchestrierung beschrieben. In mehreren Experten-Workshops wurden dazu die branchenspezifischen Ansätze an einem allgemeinen Orchestrierungsmodell diskutiert und verglichen.

Abstract

For orchestrating modular plants, some partially standardized approaches already exist in the process industry, the manufacturing industry and production-related logistics. Although these approaches pursue similar goals in principle, they differ in their design, which leads to system discontinuities. In order to unify the orchestration, this paper pursues the idea of further developing the Module Type Package (MTP) concept, known from the process industry, across industries. Research and standardization gaps are described with regard to cross-industry use of the MTP concept with a focus on orchestration. In several expert workshops, industry-specific approaches were discussed and compared against a general orchestration model.

1 Motivation und Zielstellung

Die Modularisierung von Produktionssystemen und die damit einhergehende funktionale Kapselung gilt als Schlüssel, um die steigenden Anforderungen nach Flexibilität und Wandelbarkeit zu erfüllen. Produktionssysteme besitzen meist Anteile aus der Prozess- und/oder Fertigungsindustrie sowie der produktionsnahen Logistik. In jeder dieser Branchen existieren teils standardisierte Konzepte der funktionalen Kapselung und der Orchestrierung. Dabei sind u.a. die VDI/VDE/NAMUR 2658 [1] aus der Prozessindustrie, PackML [2] aus dem Bereich der Verpackungsmaschinen sowie die Forschungsanstrengungen rund um Skills und Capabilities [3] aus der Fertigungsindustrie zu erwähnen. Der Vergleich dieser Ansätze, wie z.B. in [4] durchgeführt, zeigt, dass die grundsätzlichen Zielstellungen ähnlich sind, sich jedoch im Detail unterscheiden, was letztlich in Inkompatibilitäten und Systembrüchen resultiert. Eine einheitliche Lösung für Prozess-, Fertigungs- und Logistikanlagen würde diese Inkompatibilitäten und Systembrüche vermeiden. Ein Ansatz für diese Vereinheitlichung ist, das in der VDI/VDE/NAMUR 2658 spezifizierte generische Module Type Package (MTP)-Konzept als Grundlage aufzugreifen und branchenübergreifend weiterzuentwickeln.

Um die für die Vereinheitlichung zu schließenden Forschungs- und Standardisierungslücken aufzudecken, wurden mehrere Experten-Workshops durchgeführt. Der Schwerpunkt lag dabei auf einer Vereinheitlichung des Orchestrierungsprozesses für servicebasierte Funktionen auf Basis des MTP-Konzepts über Branchengrenzen hinweg. Darüber hinaus wurden branchenspezifische Vorarbeiten auf ihre Anwendbarkeit im vereinheitlichten Ansatz einbezogen, so u.a. die Arbeiten aus dem ENPRO 2.0-Projekt „Modulare Produktionslogistik (MoProLog)“ [5] für den Bereich der produktionsnahen Logistik, aus dem Projekt „Modular Engineering Techniques for Heterogeneous Discrete Systems (METHODS)“ [6] für den Bereich der Fertigungsindustrie und verschiedenen Arbeiten aus dem Bereich der Skills und Capabilities [7, 8].

In diesem Beitrag werden zunächst anhand eines verallgemeinerten Orchestrierungsmodells (siehe Abschnitt 2) charakteristische Merkmale der Prozessindustrie (siehe Abschnitt 3), der produktionsnahen Logistik (siehe Abschnitt 4) sowie der Fertigungsindustrie (siehe Abschnitt 5) aufgezeigt. Dabei wird vor allem hervorgehoben, welche Auswirkungen auf die Orchestrierung modularer Anlagen sich aus diesen Merkmalen ergeben. Zudem werden bestehende Arbeiten zur Umsetzung der Orchestrierung sowie vorhandene Forschungs- und Standardisierungslücken beschrieben, um das MTP-Konzept branchenübergreifend einzusetzen. In Abschnitt 6 werden die erlangten Erkenntnisse und die bestehenden Lücken aus den drei Branchen vergleichend zusammengefasst. Es folgt ein Fazit und einem Ausblick in Abschnitt 7.

2 Verallgemeinertes Orchestrierungsmodell

Als Referenzrahmen zum Vergleich der Orchestrierungsansätze in der Prozessindustrie, Fertigungsindustrie und produktionsnahen Logistik wird ein verallgemeinertes Orchestrierungsmodell genutzt, welches im Rahmen von Expertenworkshops entstand (siehe Bild 1).

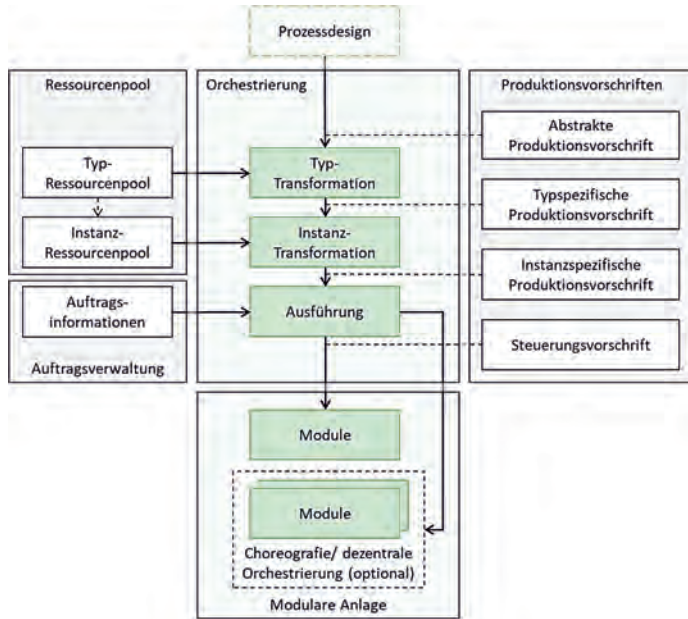


Bild 1: Branchenübergreifendes Modell für die Orchestrierung modularer Anlagen

Die Idee des MTP-Konzepts nach VDI/VDE/NAMUR 2658 besteht darin, vorprojektierte Module bereitzustellen, die schnell und einfach zu einer Gesamtanlage zusammengestellt werden können. Unter einem Modul wird eine über alle Gewerke hinweg abgeschlossene Einheit verstanden, die eine dedizierte Funktion erfüllt [9]. Nach dem MTP-Konzept automatisierte Module stellen eine standardisierte Laufzeitschnittstelle bereit und besitzen eine standardisierte Offline-Beschreibungsdatei, das sogenannte Module Type Package (MTP). Das MTP beinhaltet dabei notwendige Informationen, um das Modul möglichst automatisiert in ein Leitsystem zu integrieren. Im Rahmen dieses Beitrags wird die Orchestrierung, d.h. das Engineering und die Koordination derartiger modularer Anlagen betrachtet [10].

2.1 Prozessdesign

Im Vorfeld zur Orchestrierung muss zunächst der umzusetzende Prozess durch eine Person mit Prozesswissen (z.B. Verfahrenstechniker:in) entworfen werden. Es resultiert eine abstrakte

Produktionsvorschrift, die die prozessbedingt notwendige Sequenz von Prozessschritten inkl. deren Parameter eindeutig, vollständig und technologie-unabhängig beschreibt. Darüber hinaus sind darin Auslegungsgrenzen (z.B. Durchsatz) für den Prozess und die Anlage enthalten.

2.2 Ressourcenpool

Als Grundlage für die Orchestrierung dient eine Menge zur Verfügung stehender Module in Form eines Ressourcenpools, die für den Aufbau einer modularen Anlage vorgesehen sind. Dabei wird davon ausgegangen, dass die Module gemäß dem MTP-Konzept automatisiert sind. Jede Modulinstanz des Ressourcenpools entspricht einem Modultyp, wobei mehrere Instanzen vom gleichen Typ sein können. Von jedem Modultyp werden ein oder mehrere Serviceinstanzen bereitgestellt, die jeweils einem bestimmten Servicetyp entsprechen und zur Realisierung einer bestimmten Funktionalität verwendet werden können.

2.3 Typ-Transformation

In diesem Schritt wird die abstrakte in die typspezifische Produktionsvorschrift transformiert. Dabei muss zunächst festgelegt werden, welche Teile der Anlage prozessbedingt in einer physikalisch fest-verketteten Linientopologie und welche in einer lose verketteten Bereichstopologie umgesetzt werden sollen. Auf Basis dieser Entscheidung werden durch Experten den einzelnen Prozessschritten Funktionen, bereitgestellt durch Servicetypen, zugeordnet. Hierbei kann sowohl auf elementare als auch zusammengesetzte Servicetypen zurückgegriffen werden. Unterstützt wird dieser Vorgang durch semantische Beschreibungen und weitere abzugleichende Kriterien, die zur Prüfung der Passgenauigkeit der Servicetypen zu den Prozessschritten herangezogen werden können. Je vollständiger und genauer die semantischen Beschreibungen und Kriterien sind, desto besser kann dieser Schritt zukünftig unterstützt werden. Neben den notwendigen Servicetypen gilt es in diesem Schritt festzulegen, welche Modultypen des Ressourcenpools diese umsetzen können. Dabei ist zu beachten, dass ggfs. mehrere Servicetypen prozessbedingt durch dasselbe Modul ausgeführt werden müssen. Die geforderten Servicetypen, die ein Modul bereitstellen muss, um an einer bestimmten Stelle im Prozess eingesetzt werden zu können, werden als Modulrollen bezeichnet. Jeder Modulrolle werden anschließend ein oder mehrere Modultypen zugeordnet, die diese Rolle erfüllen können.

Auf Grundlage der definierten Modulrollen gilt es schließlich festzulegen, wie die gewählten Servicetypen zu einer Gesamtanlage assoziiert werden sollen, d.h. wie Abläufe, Prozesswertverschaltungen und Anlagenverriegelungen umgesetzt werden. Als Assoziationsmechanismen stehen die zentrale und dezentrale Orchestrierung sowie die Choreografie zur Verfügung.

2.4 Instanz-Transformation

Durch die Auswahl jeweils einer passenden Modulinstanzen für jede Modulrolle wird die typspezifische Produktionsvorschrift in die instanzspezifische Produktionsvorschrift überführt.

Diese Auswahl kann von verschiedenen Faktoren und Metriken (z.B. Ressourcenauslastung oder Kosten) beeinflusst werden. Mit diesem Schritt ist die Planung der modularen Anlage abgeschlossen.

2.5 Ausführung

Als Vorbereitung für die Ausführung kann je nach Topologie eine Inbetriebnahme der modularen Anlage notwendig sein. Hardwareseitig umfasst diese den Aufbau der gewählten Modulinstanzen und deren notwendigen Verkettung. Softwareseitig sind die geplanten Serviceassoziationen zu konfigurieren. Auf Basis der instanzspezifischen Produktionsvorschrift und einem Produktionsauftrag wird im Anschluss eine Steuerungsvorschrift abgeleitet und ausgeführt. Diese enthält Parametrierungen und Steuerbefehle, um die Services der ausgewählten Modulinstanzen anzusteuern und ggfs. notwendige Transporte zu koordinieren.

3 Prozessindustrie

3.1 Charakterisierung der Branche

Das Ziel prozesstechnischer Anlagen ist, durch physikalische, chemische, biologische oder nukleare Prozesse formlose Eingangsstoffe in formlose Ausgangsstoffe umzuwandeln. Diese Stoffumwandlung kann kontinuierlich (Konti-Anlagen) oder diskontinuierlich (Batch-Anlagen) erfolgen. Hinsichtlich der Anlagentopologie unterscheidet man in eine Produktionslinie mit starren Rohrleitungsverbindungen (z.B. Single-Use-Anlage) und in einen Produktionsbereich mit fahrbaren Modulen (z.B. fahrbare Reaktoren auf AGVs). Um in beiden Topologievarianten einen optimalen Produktionsablauf zu gewährleisten, ist ein automatisiertes Messen prozessrelevanter Größen, ein bedarfsgerechtes Steuern des Prozessablauf und ein Regeln bestimmter Prozessgrößen auf bestimmte Sollwerte notwendig. Wichtige Messgrößen in der Prozessindustrie sind beispielsweise Füllstände, Drücke, Mengen, Temperaturen, Konzentrationen usw.

3.2 Prozessdesign

Im Zuge des Prozessdesigns sind die notwendigen Prozessfunktionen, deren Reihenfolge sowie auf die Ausbringungsmenge normierte Sollwerte festzulegen und in einer abstrakten Produktionsvorschrift zu beschreiben. Eine gute Grundlage für die Beschreibung der prozeduralen Informationen bieten die Rezeptebenen der ISA 88 [11], wobei die abstrakte Produktionsvorschrift auf Ebene des General Recipe gesehen wird. In [12] wird gezeigt, wie die Rezepte der ISA 88 auf modulare Prozessanlagen übertragen werden können.

3.3 Ressourcenpool

Gemäß VDI 2776 [13] werden prozesstechnische Module als Process Equipment Assemblies (PEAs) bezeichnet. PEAs besitzen immer eine eigene Steuerung und stellen jeweils einen oder mehrere MTP-Services bereit. Das Service-Konzept geht dabei im Ursprung auf die Definitionen der ISA S88 [11] zurück. ISA 88-orientierte Services sind vor allem für Batch-Anlagen

geeignet. Für Konti-Anlagen bietet sich hingegen der Einsatz von ISA 106-orientierten Services an [14]. Mit [15] wurde bereits gezeigt, wie mit ISA 88 Funktionen auch die Funktionen nach ISA 106 umgesetzt werden können, was auf ISA 88-orientierte Services übertragbar zu sein scheint. Services bieten eine oder mehrere Prozeduren mit unterschiedlichen Parametern an, womit sie in unterschiedlichen Anlagenkontexten eingesetzt werden können [16].

3.4 Typ-Transformation

Zu Beginn dieses Schrittes wird die Anlagentopologie festgelegt. Dabei sind starr verrohrte Produktionslinien, flexibel verbundene Produktionsbereiche oder Mischformen auch unter Einbeziehung nicht-modularer Anlagenteile möglich [17]. Den einzelnen Prozessschritten werden ein oder mehrere passende Servicetypen zugeordnet und gemäß [18] zu Modulrollen zusammengefasst. Dabei können einzelne PEAs meist verschiedene Modulrollen umsetzen. So kann eine Reaktor-PEA auch als einfacher Vorlagebehälter eingesetzt werden, bei dem nur einer der angebotenen Services benötigt wird. Für den Fall einer Produktionslinie entsteht so eine typspezifische Produktionsvorschrift, die auf Basis von [11,12] beschrieben werden kann. Für die Umsetzung des Produktionsbereichs ist bisher keine standardisierte Produktionsvorschrift vorhanden, eine Anwendbarkeit von ISA 88 [11] Rezeptstrukturen gilt es zu untersuchen.

Unabhängig von der Topologie ist eine möglichst automatisierte Unterstützung der Experten bei der Auswahl möglicher Service-Typen vorzusehen. Dafür ist eine semantische Beschreibung von Prozessfunktionen notwendig. Aufgrund der Komplexität und Vielfältigkeit der Prozesse stellt die Erstellung einer solchen Semantik eine Herausforderung dar. Eine Idee ist die schrittweise Einführung der Semantik, beginnend mit grundlegenden Servicefunktionen.

Zur Assoziation von Services entsprechend der gewählten Topologie wird von der VDI/VDE/NAMUR 2658-1 [9] die Möglichkeit der zentralen Orchestrierung vorgesehen. Für die Einbindung von PEAs (wie z.B. einer Temperieren-PEA) in andere PEAs (z.B. eine Reaktor-PEA) bietet sich die dezentrale Orchestrierung an [19]. Auch der Ansatz der Choreografie kann im Umfeld der Prozessindustrie gewinnbringend eingesetzt werden [20, 21].

3.5 Instanz-Transformation

In diesem Schritt gilt es, für jede Modulrolle der typspezifischen Produktionsvorschrift eine freie Instanz mit passendem Typ auszuwählen. Diese Auswahl kann automatisiert stattfinden. Im Falle einer Produktionslinie ist die Auswahl der Modulinstanzen Teil des Engineering-Prozesses. Im Falle eines Produktionsbereichs bspw. mittels auf AGVs montierter oder manuell verfahrbarer Reaktoren, ist die Auswahl eine Scheduling-Aufgabe während des Betriebs. Die Auswahl einer zu nutzenden Instanz ist möglichst zeitnah zur Ausführung unter Berücksichtigung des Systemzustandes zu treffen.

Im Fall der Produktionslinie ist das Ergebnis eine instanzspezifische Produktionsvorschrift vergleichbar mit dem Master Recipe der ISA 88 [11]. Eine vergleichbare standardisierte instanzspezifische Produktionsvorschrift für einen Produktionsbereich ist bisher nicht bekannt.

3.6 Ausführung

Für die Ausführung sind grundsätzlich betriebsbereite PEAs notwendig. Im Falle einer Produktionslinie sind die gewählten PEA-Instanzen entsprechend der instanzspezifischen Produktionsvorschrift aufzubauen, zu verrohren und Konfigurationen für dezentrale Orchestrierungen und Choreografien zu laden. Anschließend können sie entsprechend des Produktionsauftrags gestartet werden. Im Falle eines Produktionsbereichs, ist das Ziel die flexible Nutzung von PEAs in der Produktion. Hierbei stehen die PEA-Instanzen meist bereits betriebsbereit im Produktionsbereich zur Verfügung und müssen nicht separat für den Auftrag in Betrieb genommen werden. Es gilt allerdings sicherzustellen, dass die PEAs einsetzbar und erreichbar sind.

Für die eigentliche Prozessausführung ist zwischen zwei Varianten zu unterscheiden. Bei der *Batch-Variante* werden Services bestimmter PEAs gemäß der Produktionsvorschrift ausgeführt und auf deren erfolgreiche Beendigung gewartet, bevor die Ausführung fortgeführt wird. Im Falle der *Konti-Variante* werden alle notwendigen Services zu Beginn einer Produktion einmal parametrisiert und fahren dann über eine Anfahrprozedur auf einen vorgegebenen Arbeitspunkt hoch. Alle Services sind gleichzeitig in der Ausführung mit dem Ziel, den angefahrenen Arbeitspunkt zu halten, bis ein anderer Arbeitspunkt eingestellt wird, oder der Prozess durch eine Abfahrprozedur beendet wird. In einer Linientopologie besteht die Möglichkeit der Nutzung der Batch- oder Konti-Variante. Im Produktionsbereich ist lediglich eine Batch-Variante möglich, da ein kontinuierlicher Produktfluss nur mittels einer festen Verrohrung möglich ist.

4 Produktionsnahe Logistik

4.1 Charakterisierung der Branche

Ziel produktionslogistischer Prozesse ist, sicherzustellen, dass die richtige Menge eines benötigten Materials in der richtigen Qualität am richtigen Ort zur richtigen Zeit bereitgestellt wird. Unter produktionsnaher Logistik wird dabei gemäß [22] der Teil der Produktionslogistik verstanden, der eine unmittelbare Schnittstelle zum Produktionsprozess aufweist. Der wesentliche Fokus dieses Beitrags liegt auf der Verpackungstechnik, in der ein produziertes Produkt in logistische Objekte (LO) verpackt und für den Transport vorbereitet wird. Ein Verpackungssystem kann eine Linientopologie mit starrem Materialfluss (realisiert durch Stetigförderer, wie Förderbänder), einen Logistikbereich mit flexiblem Materialfluss (realisiert durch Unstetigförderer, wie AGVs) oder eine Kombination aus beidem aufweisen. Zur Automatisierung produktionsnaher Logistiksysteme sind Mess-, Steuer- und Regelfunktionen mit Prozessgrößen vergleichbar zur Prozess- und Fertigungsindustrie notwendig, da sowohl formlose Stoffe als auch

Stückgut verarbeitet wird. Allerdings ist in produktionsnahen Logistiksystemen meist kein übergeordnetes Leitsystem vorhanden. Die Logistikmaschinen führen ihre Funktion weitestgehend autark und dezentral aus.

4.2 Prozessdesign

Das Prozessdesign in der produktionsnahen Logistik umfasst eine Festlegung der durchzuführenden logistischen Prozessschritte, deren Reihenfolge und der notwendigen Parameter. Besonders ist hierbei, in Bezug auf die Verpackungstechnik, dass ein und dasselbe Produkt aufgrund von Kundenwünschen oder bestimmten Regularien auf unterschiedliche Weise verpackt werden muss. Hierbei sind meist die gleichen Prozessschritte mit anderer Parametrierung notwendig. All diese Informationen gilt es, in einer abstrakten Produktionsvorschrift festzuhalten. Bisher existieren keine standardisierten Modelle oder Beschreibungsformen zur Planung produktionsnaher Logistiksysteme. Somit ergibt sich hier ein Forschungsbedarf.

4.3 Ressourcenpool

Analog zu den PEAs werden in diesem Beitrag logistische Module als Logistics Equipment Assemblies (LEAs) bezeichnet. Wie in [5] dargestellt lassen sich LEAs in Singleline- oder Multiline-LEAs und in Singlepurpose- oder Multipurpose-LEAs sowie Kombinationen dieser Arten unterteilen. LEAs stellen in ihrer Steuerung meist einen einzigen MTP-Service bereit. Gemäß [5] stehen dabei zwei Ausführungsarten der LEA-Services zur Verfügung, die als Prozeduren umgesetzt sind. Der Cyclic Execution Service (CES) dient zur auftragsorientierten, schnellen und gleichartigen Bearbeitung von LOen. Der Single Execution Service (SES) ermöglicht hingegen eine bedarfsorientierte, individuelle Bearbeitung einzelner LOe.

Die Konfiguration von LEA-Services kann über die Auswahl von Prozeduren und die Vorgabe von Maschinenparametern und auftragsspezifischen Parametern als Einzelparameter erfolgen. Die produktspezifische Parametrierung ist meist sehr umfangreich. Um dabei nicht jeden Parameter einzeln einstellen zu müssen, werden in LEAs häufig produktspezifische Parametersätze hinterlegt, die nach Bedarf über eine Produkt-ID referenziert und angewendet werden können. Ein entsprechendes Parametermanagement ist Gegenstand aktueller Arbeit im EN-PRO 2.0-Projekts „Modulare Produktionslogistik (MoProLog)“.

4.4 Typ-Transformation

Einzelne LEAs können verschiedene Prozessschritte ausführen (siehe Abschnitt 4.3, Multipurpose-LEA). So kann eine Stretchhaubenmaschine eine Folienhaube über eine Palette stretchen oder ein Folieninlet in einen Oktabin einlegen. Diese verschiedenen Funktionen müssen in der LEA vorgesehen sein und können mittels Parametrierung eingestellt werden.

Eine semantische Beschreibung von Logistikfunktionen und den daraus resultierenden Modultypen existiert bisher nicht. Als Grundlage könnte bspw. die Klassifikation von

Verpackungsmaschinen in ECLASS [23] verwendet werden. Bei der Definition einer Semantik sind auch Auslegungsgrößen wie der mögliche Durchsatz bzw. Prozesszeiten zu berücksichtigen.

Im Rahmen des MoProLog-Projekts wurden die Möglichkeiten der Assoziation von CES und SES zu einem modularen logistischen System auf Basis von zentraler und dezentraler Orchestrierung sowie Choreografie aufgezeigt. Dabei ist zu berücksichtigen, ob es sich um eine Logistikklinie oder einen Logistikbereich handelt. Für beide Fälle wurden im Rahmen von MoProLog bereits Lösungsansätze gefunden und evaluiert. Entsprechende Veröffentlichungen dazu sind kürzlich erschienen [24] bzw. werden folgen.

4.5 Instanz-Transformation

In Bezug auf das instanzspezifische Engineering muss in der produktionsnahen Logistik zwischen einer Logistikklinie und dem Logistikbereich unterschieden werden [5]. Im Falle von Logistikklinien ist die zu nutzende LEA-Instanz im Zuge des Engineerings festzulegen. Im Logistikbereich erfolgt die Festlegung der LEA-Instanz in der Betriebsphase des Bereichs durch eine entsprechende Koordination der Materialflüsse innerhalb des Systems.

4.6 Ausführung

Ähnlich wie in der Prozessindustrie kann auch in der produktionsnahen Logistik je nach Topologie eine Inbetriebnahme als Grundlage für die Ausführung notwendig sein. Logistikklinien dienen der kontinuierlichen Abarbeitungen gemäß eines Auftrags und einer Produktionsvorschrift. Hierfür sind die ausgewählten LEAs aufzubauen, zu verketteten und zu konfigurieren. Für die Automatisierung der LEAs werden CES eingesetzt. Diese Services werden zu Beginn einer Auftragsausführung mit einer auftragsspezifischen Parametrierung gestartet und beenden sich mit der Vollendung des Auftrags. Anders verhält es sich im Logistikbereich, da die dortigen LEAs meist bereits in Betrieb genommen wurden und auf den nächsten Auftrag warten. Diese LEAs sind mit SES automatisiert und werden gestartet, sobald sie aufgestellt und in Betrieb genommen wurden. Anschließend können sie ankommende LOe verschiedener Aufträge erkennen, beziehen für jedes LO dessen individuelle Parameter und bearbeiten es entsprechend. [5]

5 Fertigungsindustrie

5.1 Charakterisierung der Branche

Die Aufgabe der Fertigungsindustrie ist die wirtschaftliche Herstellung von geometrisch bestimmten festen Körpern sowie deren Montage zu komplexeren Baugruppen [25]. Die Körper bzw. deren Stoffe werden dabei durch Fertigungsverfahren schrittweise von einem Roh- in einen Fertigzustand überführt. Die Fertigungsverfahren lassen sich gemäß DIN 8580 [26] in Urformen, Umformen, Trennen, Fügen, Beschichten und Stoffeigenschaften ändern

unterteilen. Ein Fertigungssystem kann in Form einer Linientopologie (z.B. Fließfertigung) oder eines Fertigungsbereichs (z.B. Werkstatt- oder Matrixfertigung) umgesetzt werden. In der Linientopologie werden die Werkstücke zwischen den Maschinen mittels Stetigförderern transportiert. In einem Fertigungsbereich werden die Transporte durch Unstetigförderer (z.B. Automated Guided Vehicles, AGVs) umgesetzt. In der Automatisierung von Fertigungsprozessen kommen meist Steuerungsfunktionen zum Einsatz. Die zu erfassenden Größen sind primär diskreter Natur wie z.B. Anzahlen, Positionsinformationen oder Identifikation einzelner Werkstücke.

5.2 Prozessdesign

Zur Beschreibung des Prozessdesigns in der Fertigungsindustrie werden häufig Bill of Processes (BOP) verwendet. Diese sind nicht standardisiert und beschreiben die durchzuführenden Fertigungsschritte meist nur auf einer hohen Abstraktionsebene. Details zum Prozessablauf sind vor allem bei stoffverarbeitenden Prozessen (wie z.B. Fräsen) häufig implizit in Programmen (wie z.B. NC-Programmen) versteckt. In Montageprozessen werden in einigen Fällen Vorrangsgraphen zur Modellierung einzelner Prozessschritte eingesetzt, z.B. in [27]. Dabei besteht die Besonderheit, dass die Abarbeitungsreihenfolge der Produktionsschritte variieren kann. Transporte zwischen Maschinen sind in BOPs bisher nur implizit vorgesehen. Im Skill-Umfeld beschreiben die Autoren in [8] einen Ansatz, der Business Process Model and Notation (BPMN) zur Modellierung eines skill-basierten Fertigungsprozesses nutzt. Im MTP-Umfeld existieren noch keine Modelle zur Beschreibung von Fertigungsprozessen. Allerdings scheinen die Ansätze zur Modellierung der Produktionsvorschrift mit Skills und Capabilities eine gute Basis auch für eine MTP-basierte Automatisierung zu bilden. Dies gilt es zu prüfen.

5.3 Ressourcenpool

Analog zu den PEAs werden in diesem Beitrag fertigungstechnische Module als Manufacturing Equipment Assemblies (MEAs) bezeichnet. MEAs lassen sich nach ihrer Funktion klassifizieren, z.B. auf Basis der DIN 8580 [26] für Fertigungsverfahren oder der VDI 2860 [28] für Handhabungsverfahren. Aus automatisierungstechnischer Sicht stellen MEAs einen oder mehrere Services auf ihren Steuerungen bereit. Ein auf die Bedarfe der Fertigungsindustrie zugeschnittenes Service-Konzept existiert allerdings noch nicht. Inspiration dafür könnten die Arbeiten der Skill-Community bieten. Skills könnten dabei mit einer standardisierten Schnittstelle entsprechend der VDI/VDE/NAMUR 2658 versehen werden, wie in [7] angedacht.

Fertigungstechnische Services können auf unterschiedlichen Granularitätsebenen definiert werden. So sind spezifische Funktionen, wie z.B. ein Bohren mit den Parametern Drehzahl und Vorschub, möglich, oder aber universelle Funktionen, z.B. CNC-Fräsen, die nicht nur Parameter, sondern ein ganzes NC-Programm zur Konfiguration benötigen. Bei Letzteren kann

je nach Programm auch ein manueller oder vollautomatischer Werkzeugwechsel notwendig werden. Die genaue Ausgestaltung und Umsetzung von fertigungstechnischen Services ist noch ungeklärt und sollte untersucht werden. Dabei stellen sich u.a. Fragen nach einer sinnvollen Granularitätsebene der Services und einer geeigneten Parametrierungsmöglichkeit.

5.4 Typ-Transformation

In Bezug auf eine typspezifische Produktionsvorschrift lässt sich feststellen, dass einzelne MEAs verschiedene Prozessschritte ausführen können (z.B. kann eine Fräs-MEA auch bohren). Dies ist vor allem auf die Anpassbarkeit von Service-Funktionalitäten durch das Laden verschiedener NC-Programmen (siehe Abschnitt 5.3) zurückzuführen.

Für die notwendige semantische Beschreibung von Service-Funktionalitäten und Prozessschritten bilden Standards wie die DIN 8580 [26] und die VDI 2860 [28] bereits eine gute Grundlage. Zudem existieren bereits Systematiken für fertigungsspezifische Parameter wie z.B. die ISO 13399 [29] für Werkzeugdaten. Diese sind semantisch aufzuarbeiten und für die typspezifische Transformation nutzbar zu machen. Dabei sind auch Auslegungsgrößen wie der mögliche Durchsatz bzw. Prozesszeiten mit zu berücksichtigen.

Hinsichtlich der angestrebten Anagentopologie gilt es zunächst festzulegen, welche Teile des Fertigungssystems als Fertigungslinien und welche in Fertigungsbereichen ausgeführt werden sollen. Dabei sind entsprechende Transporte vorzusehen. Zur eigentlichen Assoziation der Services wird im Skill-Umfeld vorrangig auf eine zentrale Orchestrierung gesetzt, wie beispielsweise in [8]. Es bleibt zu prüfen, ob auch sinnvolle Anwendungsfälle für eine dezentrale Orchestrierung oder Choreografie bestehen, z.B. ähnlich wie in Abschnitt 4.4 für den Bereich der produktionsnahen Logistik beschrieben.

5.5 Instanz-Transformation

Bei der Betrachtung der Instanz-Transformation muss zwischen Fertigungslinien und Fertigungsbereichen unterschieden werden. Im Falle der Fertigungslinien muss die Auswahl der MEA-Instanzen vor der Inbetriebnahme im Zuge des Engineerings der Anlage vorgenommen werden. Im Falle der flexibel (z.B. durch AGV-Transporte) verbundenen MEAs in Fertigungsbereichen erfolgt die MEA-Auswahl durch eine entsprechende Koordination des Materialflusses flexibel innerhalb der Betriebsphase. Falls in der typspezifischen Produktionsvorschrift eine Variabilität der Reihenfolge der Prozessschritte vorgesehen ist, gilt es auf Grundlage des aktuellen Systemzustands, den nächsten auszuführenden Prozessschritt festzulegen.

5.6 Ausführung

Vor der eigentlichen Ausführung ist ähnliche wie in den zuvor betrachteten Branchen je nach Topologie eine Inbetriebnahme der MEAs notwendig sein. Im Falle einer Fertigungslinie sind die in der instanzspezifischen Produktionsvorschrift ausgewählten MEAs in der geplanten

Topologie aufzubauen, starr zu verbinden und in Betrieb zu nehmen. Hierbei werden notwendige Konfigurationen hinsichtlich existierender dezentraler Orchestrierungen oder Choreografien konfiguriert und aktiviert. Im Falle eines Fertigungsbereichs werden einzelne MEAs, unabhängig der Produktionsvorschrift in Betrieb genommen und für die Nutzung vorbereitet. In Bezug auf die Ausführung selbst gibt es im Bereich der Fertigung im MTP-Umfeld noch keine Arbeiten. Im Skill-Umfeld, z.B. in [8], scheint eine ähnliche Philosophie wie bei Batch-Anlagen vorgesehen zu sein, sodass ein Service immer so parametrisiert, gestartet und beendet wird, wie es der Produktionsablauf vorsieht. Vor allem bei schnell getakteten Massenprozessen ist dieses Vorgehen allerdings nachteilig. Daher bleibt zu untersuchen, welche Ausführungsvarianten von fertigungstechnischen Services sinnvoll sind. Ggfs. können die in Abschnitt 4.3 vorgestellten Ausführungsarten von Services dafür eine Basis sein.

6 Zusammenfassung der Erkenntnisse und Lücken

Dieser Abschnitt fasst die wesentlichen Charakteristika, Erkenntnisse und Lücken aus der vorangegangenen Analyse der drei Branchen in der Übersicht zusammen (siehe Tabelle 1).

Tabelle 1: Charakteristika, Erkenntnisse und Lücken aus der Analyse der drei Branchen

	Prozessindustrie	Produktionsnahe Logistik	Fertigungsindustrie
Prozessdesign			
Inhalt	Prozessschritte Sequenzen normierte Parameter	Prozessschritte Sequenzen Parameter(sätze) CES/SES	Prozessschritte (variable) Sequenzen Parameter NC-Programme
Standards	Produktionslinie: ISA 88 / ISA 106 Produktionsbereich: keine	keine	keine
Erkenntnisse: Die grundsätzlichen Inhalte der abstrakten Produktionsvorschriften sind in den drei betrachteten Branchen ähnlich. Allerdings ist den unterschiedlichen Konfigurationsmöglichkeiten der Services in der abstrakten Produktionsvorschrift Rechnung zu tragen.			
Konzeptionelle Lücken: Als Grundlage für die Orchestrierung ist die Existenz einer formalen Beschreibung der abstrakten Produktionsvorschrift unabdingbar. Hier zeigen sich vor allem in der Fertigungsindustrie und produktionsnahen Logistik noch Lücken. Im Bereich der Prozessindustrie gilt zu prüfen, ob die Konzepte der ISA 88 auch auf flexible Produktionsbereiche anwendbar sind.			
Ressourcenpool			
Modulbezeichnung	PEA	LEA	MEA
Servicearten	ISA 88- und ISA 106-orientierte Services	CES, SES	Bisher keine
Servicekardinalität	≥ 1 Service pro PEA	1 Service pro LEA	≥ 1 Service pro MEA
Service-Konfigurierbarkeit	Prozeduren Parameter	Prozeduren Parameter	Parameter Parametersätze

		Parametersätze	NC-Programme
Erkenntnisse: • Trotz der gezeigten Unterschiede kann die Automatisierung der Module aller drei Branchen konform zum MTP-Konzept realisiert werden. Die Modul-Schnittstellen können jeweils mittels eines MTPs beschrieben werden.			
Konzeptionelle Lücken: • Die Konformität zum MTP-Konzept in allen drei Branchen ermöglicht es, branchenübergreifende Tools zur Verwaltung des Ressourcenpools zu entwickeln. Zur Bereitstellung der Daten aus den MTPs sollten diese Tools eine offene, standardisierte API besitzen, die es zu spezifizieren gilt. • Es gilt, ein Service-Konzept zu entwickeln, welches die Bedarfe der Fertigungsindustrie bedient. Dabei ist u.a. eine Möglichkeit vorzusehen, Parametersätze und NC-Programme in die MEAs herunterzuladen.			
Typ-Transformation			
Topologien	Produktionslinien Produktionsbereich	Logistiklinie Logistikbereich	Fertigungslinien Fertigungsbereich
Modulflexibilität	Parallel-ausführbare Servicefunktionen	LEA-Services mit Pa- rametersätzen	Universelle MEA-Ser- vices mit nachladba- rem NC Programmen
Semantische Beschreibung	Keine Standards, proprietäre Lösungen	ECLASS (zu prüfen)	DIN 8580 (zu prüfen) VDI 2860 (zu prüfen)
Assoziationsmög- lichkeiten	Zentrale und dezent- rale Orchestrierung, Choreografie	Zentrale und dezent- rale Orchestrierung, Choreografie	Bisher keine
Erkenntnisse: • In allen drei Branchen können starre Linien- und flexible Bereichstopologien auftreten. • Die Module aller Branchen weisen eine gewisse Flexibilität auf, sodass sie verschiedene Rollen in einer Anlage einnehmen können. Diese Modulflexibilität basiert auf verschiedenen Mechanismen.			
Konzeptionelle Lücken: • Eine semantische Beschreibung von Servicefunktionalitäten ist zwar erforderlich, aber vor allem im Bereich der Prozessindustrie und der produktionsnahen Logistik kaum ausgeprägt. Im Bereich der Fertigungsindustrie gilt es, bestehende Systematiken für die Orchestrierung nutzbar zu machen. • Hinsichtlich der Assoziationsmöglichkeiten fehlt es an einem Leitfaden, wann welcher Mechanismus sinnvoll einzusetzen ist. Zudem gilt es für den Bereich der Fertigungsindustrie zu prüfen, ob ähnliche Möglichkeiten der Serviceassoziation wie in der Logistik sinnvoll sind. • Die semantische und formale Beschreibung von assoziierten Service-Konstrukten bedarf weiterer Standardisierung, um diese MTP-basiert in übergeordnete Systeme integrieren zu können.			
Instanz-Transformation			
Zeitpunkt der Durchführung	Produktionslinie: Im Engineering Produktionsbereich: Während des Betriebs	Logistiklinie: Im Engineering Logistikbereich: Während des Betriebs	Fertigungslinie: Im Engineering Fertigungsbereich: Während des Betriebs
Erkenntnisse: • Im Falle fest verketteter Module in einer Linientopologie zeigt sich, dass dieser Schritt vor der Inbetriebnahme im Zuge des Engineerings stattfindet, da erst danach der physikalische Aufbau der Module erfolgen kann. • Im Falle lose verketteter Module in Bereichstopologien zeigt sich, dass dieser Schritt kurz vor der Ausführung der Produktion während des Betriebs erfolgen muss, um flexibel auf Systemänderungen zu reagieren.			
Konzeptionelle Lücken:			

- Für die flexible Auswahl von Modulinstanzen in Bereichstopologien ist es sinnvoll, Laufzeiteinformationen der Instanzen (z.B. aktuelle Auslastung) mit einzubeziehen. Welche Informationen hier notwendig und sinnvoll sind, gilt es zu untersuchen.

Ausführung			
Inbetriebnahme	Produktionslinie: Aufbau und Inbetriebnahme zur Produktion Produktionsbereich: Aufbau und Inbetriebnahme unabhängig einer Produktion	Logistiklinie: Aufbau und Inbetriebnahme zur Produktion Logistikbereich: Aufbau und Inbetriebnahme unabhängig einer Produktion	Fertigungslinie: Aufbau und Inbetriebnahme zur Produktion Fertigungsbereich: Aufbau und Inbetriebnahme unabhängig einer Produktion
Auftrags-Informationen	Ausbringungsmenge Kontextinformationen	Ausbringungsmenge Kontextinformationen	Ausbringungsmenge Kontextinformationen
Ausführungs-Varianten	Konti: Service wird parametrisiert & ist für die Dauer der Produktion aktiv; Batch: Service wird parametrisiert & ist für die Dauer der Batchbearbeitung aktiv	CES: Service wird parametrisiert und läuft dann für einen Auftrag kontinuierlich durch; SES: Service wird für jedes Logistikkobjekt individuell parametrisiert	Bisher keine Angaben
Erkenntnisse: • Ausführung bedarf immer betriebsbereiter Module, wobei die Betriebsbereitschaft in einer Linie speziell für eine Produktionsvorschrift und in einem Bereich unabhängig einer definierten Produktionsvorschrift hergestellt wird. • Als auftragsspezifische Parametrierung sind in allen drei Branchen die Ausbringungsmenge und Kontextinformationen (z.B. Auftragsnummer) für die Rückverfolgbarkeit vorzusehen. • In der Batch- und SES-Ausführung sind Services nur für die augenblickliche Bearbeitung des Produkts aktiv, in der Konti- und CES-Ausführung für die gesamte Dauer der Produktion.			
Konzeptionelle Lücken: • Es gilt zu prüfen, ob die Ansätze zur Serviceausführung aus der produktionsnahen Logistik auch für den Bereich der Fertigungsindustrie sinnvoll anwendbar sind.			

7 Fazit und Ausblick

Der vorliegende Beitrag hat gezeigt, dass in der Prozessindustrie [30] und der produktionsnahen Logistik [5, 24] bereits einige Konzepte in Bezug auf die Orchestrierung modularer Anlagen im Sinne des MTP-Konzepts bestehen. Mit den Vorarbeiten zu Skills und Capabilities der Fertigungsindustrie sind Grundkonzepte für den Einsatz und die Verarbeitung funktionaler Semantik geschaffen worden. Der Vergleich hat gezeigt, dass vor allem in der Prozessindustrie diese funktionale Semantik weitestgehend fehlt, während in der Fertigungsindustrie für die Umsetzung von Skills noch kein standardisiertes Schnittstellenkonzept existiert. Daraus lassen sich zwei große Handlungsfelder ableiten. Für die Fertigungsindustrie ist verstärkt die Anwendung von MTP-Konzepten für die Realisierung von Schnittstellen und deren Beschreibung zu untersuchen. Für die Prozessindustrie und Logistik sind dagegen die Konzepte der funktionalen Semantik tiefergehend zu untersuchen. Somit können aus beiden Industriebereichen

existierende und laufende Konzepte gewinnbringend in Verbindung gebracht werden. In diesem Beitrag wurden wesentliche Aspekte hinsichtlich einer funktionalen Kapselung und Beschreibung von *Equipment Assemblies* in den drei betrachteten Branchen dargestellt. Es wurden aktuell ungelöste Fragen aufgezeigt, die in künftigen Forschungs- und Gremienarbeiten untersucht werden sollten. Die Beantwortung dieser Fragen legt die Grundlage einer Harmonisierung und Weiterentwicklung des MTP Konzepts über alle drei Branchen, womit Inkompatibilitäten und Systembrüche vermieden werden. Über diesen Beitrag hinaus zeigen sich Arbeiten zur Orchestrierung modularer Anlagen im Umfeld der Energietechnik, der Wasserstoffwirtschaft und des Schiffbaus identifiziert. All diese Arbeiten unterstreichen, dass die Potenziale des MTP-Konzepts deutlich über die Grenzen der Prozessindustrie hinaus reichen.

Literaturangaben

- [1] VDI/VDE/NAMUR 2658: Automatisierungstechnisches Engineering modularer Anlagen in der Prozessindustrie. <https://www.vdi.de/2658>
- [2] ISA-TR88.00.02 (PackML): Machine and unit states - An implementation example of ISA-88
- [3] Froschauer, R., Köcher, A., Meixner, K., Schmitt, S. u. Spitzer, F.: Capabilities and Skills in Manufacturing: A Survey Over the Last Decade of ETFA, 2022. <http://arxiv.org/pdf/2204.12908v1>
- [4] Grüner, S. et al.: Cross-Industry State of the Art Analysis of Modular Automation. In: VDI-Kongress Automation 2020
- [5] Blumenstein, M. et al.: Designprinzipien für den Modul- und Serviceentwurf in modularen Logistikanlagen. In: VDI-Kongress Automation 2021
- [6] Habiger, P. et al.: Plug & Produce im real-virtuellen Kontext fertigungstechnischer heterogener Anlagen. In: AALE 2022
- [7] Köcher, A., Beers, L. u. Fay, A.: A Mapping Approach to Convert MTPs into a Capability and Skill Ontology, 2022. <http://arxiv.org/pdf/2205.01382v1>
- [8] Köcher, A., Da Silva, L. M. V. u. Fay, A.: Modeling and Executing Production Processes with Capabilities and Skills using Ontologies and BPMN, 2022. <http://arxiv.org/pdf/2204.09472v1>
- [9] VDI/VDE/NAMUR 2658-1: Automatisierungstechnisches Engineering modularer Anlagen in der Prozessindustrie - Allgemeines Konzept und Schnittstellen, 2019.
- [10] Klose, A. et al.: Anforderungen an die Orchestrierung modularer Prozessanlagen, 2019. <https://opara.zih.tu-dresden.de/xmlui/handle/123456789/1397>
- [11] ANSI/ISA-88.01-1995: Models and terminology. Approved 23 October 1995
- [12] Menschner, A. et al.: Von der Prozessbeschreibung zur modularen Anlage. In: VDI-Kongress Automation 2019
- [13] VDI 2776: Verfahrenstechnische Anlagen - Modulare Anlagen - Grundlagen und Planung modularer Anlagen, 2020.
- [14] ISA-TR106.00.01-2013: Procedure automation for continuous process operations - Models and terminology
- [15] Rautenberg, M., Schneider, S. u. Roth, M.: Automatisierung von Sonderszenarien in kontinuierlichen verfahrenstechnischen Prozessen. In: VDI-Kongress Automation 2016
- [16] VDI/VDE/NAMUR 2658-4: Automatisierungstechnisches Engineering modularer Anlagen in der Prozessindustrie - Modellierung von Moduldiensten (V1.0.0, geplant 2022)

- [17] Markaj, A., Fay, A., Hoernicke, M., Schoch, N. u. Stark, K.: Requirements and conceptual design for hybrid process plants. In: ETFA 2021
- [18] Blumenstein, M., Stutz, A. u. Maurmaier, M.: Rollenbasierte Entwurfsmethodik für modulare Anlagen auf Basis von sogenannten Super-Services. In: VDI-Kongress Automation 2019
- [19] Stutz, A., Maurmaier, M., Spaethe, F. u. Hill, P.: Einführung, Entwicklung und Anwendung von sub-elementaren Services für modulare Anlagen. In: VDI-Kongress Automation 2020
- [20] Stutz, A., Fay, A., Barth, M. u. Maurmaier, M.: Choreographies in Microservice-Based Automation Architectures: Next Level of Flexibility for Industrial Cyber-Physical Systems. In: IEEE ICPS 2020
- [21] Stutz, A., Fay, A., Barth, M. u. Maurmaier, M.: Software Patterns for the Realization of Automation Service Choreographies. In: IEEE ETFA 2021
- [22] NE 171: Anwendung des modularen Anlagenkonzeptes in der produktionsnahen Logistik, 2020.
- [23] ECLASS e.V.: ECLASS. <https://eclass.eu/>
- [24] Stutz, A., Blumenstein, M., Fay, A., Barth, M. u. Maurmaier, M.: Anwendung von Automatisierungsdienste-Choreografien zur Koordination modularer Logistiklinien. In: EKA 2022, Magdeburg (in print)
- [25] Westkämper, E. u. Warnecke, H.-J.: Einführung in die Fertigungstechnik. Wiesbaden: Vieweg+Teubner Verlag 2002
- [26] 2020. DIN 8580: Fertigungsverfahren – Begriffe, Einteilung
- [27] Frommherz, B. J.: Ein Roboteraktionsplanungssystem, Bd. 260. Springer Berlin Heidelberg 1990
- [28] VDI 2860: Montage- und Handhabungstechnik - Handhabungsfunktionen, Handhabungseinrichtungen, Begriffe, Definitionen, Symbole, 1982.
- [29] 2006. ISO 13399-1:2006-02: Werkzeugdatendarstellung und -austausch - Teil 1: Übersicht, fundamentale Prinzipien und allgemeines Informationsmodell
- [30] Klose, A. et al.: Orchestration of modular plants - Procedure and application for orchestration engineering. atp-magazin 2022 (in print)

Orchestrierung hybrider Anlagen – auf dem Weg zur Integration von PackML Komponenten in Modulare Anlagen

Orchestration of Hybrid Plants – Towards Integration of PackML Machines into Modular Process Plants

Dr.-Ing. **Sten Grüner, Mario Hoernicke**,
ABB Corporate Research Center, Ladenburg

Kurzfassung

Modularisierung der Produktionsanlagen ist eine vielversprechende Möglichkeit die kommenden Anforderungen der Flexibilität, Veränderbarkeit und Interoperabilität sowohl in der diskreten Fertigung als auch in der Prozessindustrie zu adressieren. Zurzeit existieren zwei Standards für die Orchestrierung der Produktionsmodule: Module Type Package (MTP) und PackML, welche jeweils in prozesstechnischen und diskreten Anlagen eingesetzt werden. Trotz gewisser Ähnlichkeit existieren konzeptionelle und technische Unterschiede zwischen der beiden Spezifikation, die eine Interoperabilität erschweren. Die negativen Folgen sind insbesondere in hybriden Anlagen mit gemischtem Equipment bemerkbar (z. B. in Pharmaindustrie oder Batteriefertigung). In dieser Arbeit stellen wir eine Lösung für dieses Problem vor, indem wir die PackML automatisch auf MTP mappen. Unser Ansatz wurde durch eine Erweiterung eines MTP-kompatiblen Process Orchestration Layers (POL) umgesetzt und an zwei unabhängigen Softwareimplementierungen von PackML OPC UA Spezifikation validiert.

Abstract

Modular plants are a promising option to address upcoming flexibility, changeability, and interoperability requirements both in discrete manufacturing and process industry. Currently, there are two standards defining orchestration interfaces of physical modules: Module Type Package (MTP) and PackML, used in process and manufacturing plants, respectively. Despite a high-level similarity, conceptual and technical differences between those two standards have a negative impact on interoperability. This is especially severe in hybrid production facilities which utilize both equipment standards at the same time (e.g., pharmaceuticals or battery production). In this work we address this issue by presenting an automatic mapping

from PackML to MTP. Our approach is implemented by extending an MTP compatible Process Orchestration Layer (POL) system with PackML interfaces and validating it with two independent software implementations of PackML OPC UA interfaces.

1. Introduction

Modularization of process plants is seen as a promising possibility to solve upcoming requirements from process industries, like flexibility and changeability, as well as interoperability between plant assets.

For that reason, several working groups within the German user association of automation technology in process industries (NAMUR) have been established to work together with ZVEI (German association of automation technology vendors) on creating requirements and concepts for the automation of modular plants. The expected result is a definition of process module interfaces, the Module Type Package (MTP) [1] modelled in AutomationML [2], which allows a seamless integration of process modules into a process orchestration layer [1].

Besides the initiatives in process industries, in discrete industries, machine automation follows a similar approach. Machine vendors use the PackML [3] specification to describe the interfaces of the machine in a standardized way and to integrate the machine into a line automation system, enable machine to machine communication and finally have a modular plant as well. The PackML runtime interface is specified as an OPC UA companion specification [4].

In industries where continuous process modules, as well as discrete machines or general discrete automation is needed in the same plants, e.g., food and beverages, pharmaceuticals, or battery production, both approaches are used. To have a common supervisory system for these hybrid plants, a side-by-side integration of PackML and MTP based modules is required. The integration is preferably done in the same engineering tool in order to provide an integrated engineering process. Additionally, the look and feel in engineering and during operations should be the same for both models.

In the paper we present an approach to integrate PackML based on an MTP-compatible POL prototype which is extend with the capabilities to engineer and communicate to PackML modules. The remainder of this work is structured as follows: In Section 2 state of the art and previous work are described. Section 3 defines the envisioned co-existence architecture between MTP and PackML and presents a mapping of PackML concepts to MTP. In Section 4 we describe engineering and runtime aspects of side-by-side MTP and PackML integration within one tool including an automatic mapping of PackML to MTP. The software-PackML

modules used for the validation of our approach are reviewed in Section 5. Finally, a summary and an outlook can be found in Section 6.

2. State of the Art and Previous Work

Currently there is little published information available about use cases of hybrid plant co-existence esp. with respect to MTP and PackML. Within the scope of a public funded research project BaSys 4.2 the aspect of co-existence of process automation and discrete manufacturing modular equipment is addressed on at least two dimensions. The first dimension is the application of currently discrete-dominated Industry 4.0 use cases within process industry, e.g., on using Asset Administration Shell for MTP models, is not in scope of this paper. Therefore, this work is on the second dimension of the co-existence of MTP and PackML – the orchestration of modular equipment within hybrid production plants.

In our previous work on this topic, we presented a structured competitive analysis of MTP, PackML and BaSys Control Component (CC) [6]. The conclusion was that despite similar business-case motivation, all three concepts differ, all three orchestration concepts differ in conceptual details and technical realization. A follow-up publication [7] presented a generic reference architecture for modular industrial automation systems including different possible co-existence strategies of MTP and PackML equipment. In this work we continue this work by picking one co-existence strategy from [7] and presenting a set of technical measures to ensure joint operations based on the comparison from [6].

3. Targeted Co-Existence Architecture of PackML and MTP

According to the reference architecture of modular industrial automation systems [7], multiple alternatives of PackML and MTP co-existence are possible ranging from two dedicated orchestration systems – communicating with each other – to cross-module communication between module instances, as discussed in the previous work. The targeted MTP-centric co-existence, we present in this work, is depicted in the blue part of Fig. 1.

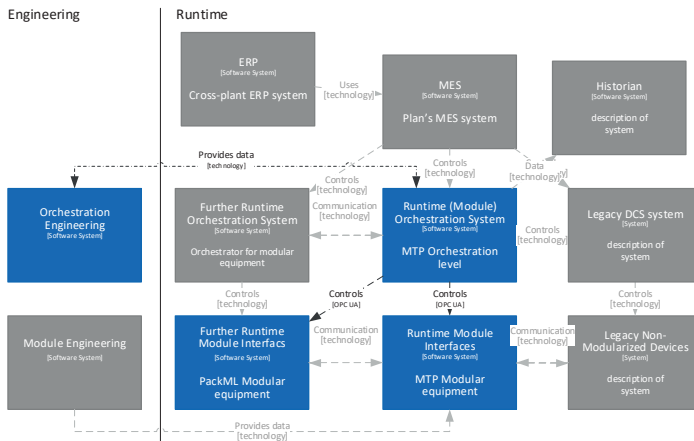


Fig. 1: MTP-centric Co-Existence Architecture (systems components in scope of this work are marked blue).

In this work we assume an MTP-centric co-existence, i.e., there is only one runtime (Module) Orchestration System which is an MTP-compatible POL. At runtime this POL is interfacing to both MTP-compatible PEAs and PackML-compatible machines using OPC UA communication interfaces. Furthermore, an MTP-compatible Orchestration Engineering system present in the plant which communicates to the POL. Due to this design decision, we need to ensure the compatibility of PackML concepts with those of MTP both during engineering and runtime by creating a suitable mapping which we describe in the following sections.

Mapping of PackML concepts to MTP

In this section a mapping of PackML concepts to respective MTP elements is presented. Focus is on the most important parts for service-based operation. Optional or administrative mappings are not presented with this contribution.

State Model

Both state models look similarly on the first glance as depicted in Fig. 2 where MTP and PackML state models can be found on the left-hand and right-hand sides, respectively.

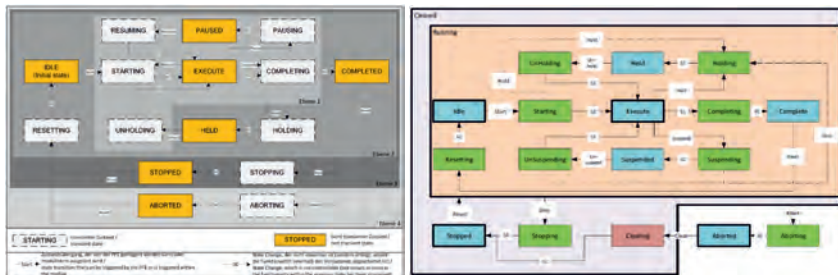


Fig. 2: State models of MTP [5] (left) and PackML [4] (right).

In general, both models distinguish between two types of states: transient (MTP) aka acting (PackML) and non-transient (MTP) aka wait (PackML) states. Transient states end with an “ing” suffix and are passed between non-transient states where machine is in a steady state. State transition can be triggered using a command or an internal “State Change (SC)” event trigger, e.g., for a self-terminating function.

Key differences within state model of PackML and MTP can be summarized as follows:

- Hold loop: The hold loop includes Holding, Held and Unholding states in both state models. The Holding state can be entered from every level 1 (“Ebene 1” in German) state within the MTP model, i.e., Execute, Resuming, Paused, Pausing, Unholding and Completing. In PackML the Holding state can be entered from Execute state only, additional extensions marked in the OPC UA companion specification allow to change to Holding from Execute, Suspending, Suspended, UnSuspending, Starting, and Unholding state.
- Pause/Suspend loop: no differences, Suspend loop is mapped to Pause loop.
- Stop loop: no differences.
- Abort loop and Clearing state: MTP has no equivalent to PackML Clearing state. Furthermore, PackML logic expects the machine to visit the Stopped state when coming from Aborted state via the Clearing state.
- Restart command: MTP has a possibility to restart a service out of the Execute state, this is not possible in PackML.
- Complete command: no differences, PackML OPC UA companion specification has an optional “toComplete” command to enter Completing state from Executing.

Based on the listed differences between the PackML and MTP state model, we would recommend following actions in case a PackML module should be used in conjunction with an MTP POL:

- Keep Clearing state unimplemented if possible, to stay close to MTP state layout.
- Skip Stopped state when coming from Aborted state or expect POL to execute a Clear + Reset command sequence when coming from Aborted state.
- Expect calls of Stop + Reset + Start sequence to “emulate” MTP’s restart command.

Operation Mode

MTP defines offline, manual, and automatic (external and internal) operation modes. PackML requires production, maintenance and manual modes to be implemented with a room for user-defined operation modes. In the prototype we use production mode within PackML machine as equivalent for MTP’s automatic external operation mode.

Services and Procedures

Conceptual comparison of MTP and PackML shows that a PackML machine can be considered as a single MTP service. Therefore, service parameters in MTP are comparable to unit parameters in PackML. For example, changes in the operation mode can be mapped to OPC UA method “SetUnitMode” within the unit representation. Similarly current unit mode can be read from the “UnitModeCurrent” variable within the StatusFolderType object.

Within MTP service model, each service has at least one procedure marked as default. Since there is no equivalent concept in PackML we propose to assume that MTP-style interface of the PackML module has exactly one default procedure which maps to the actual PackML state machine.

Alternative mappings of procedures to different “Products” from PackML domain which can be selected using a set of parameters passed to “SetProduct” methods within OPC UA representation is scope of future work.

Service Parameters

The last essential piece are service/procedure parameters which are passed to the service. The related MTP unit classes AnaServParam, BinServParam etc. can mapped to the parameters of PackMLDescriptorDataType within OPC UA information model which are passed to the Start method of the PackML state machine. For service parameters, we assume the synchronization mode, i.e., “Sync” Boolean parameter of *ServParam unit, to be active to synchronize service’s operation mode with the operation mode of the PEA, because PackML parameters do not have an own operation mode, thus have the same as the PackML state-machine.

4. Side-by-side Integration of PackML Modules within MTP POL

In this section we describe the side-by-side integration of PackML machines into an MTP-compatible POL. Achieving this goal required extending both the engineering and the runtime capabilities of the POL.

Engineering phase

Existing MTP-compatible POLs are built around the core engineering artifacts – MTP files describing various aspects of a module including its service interfaces and HMI. Obviously, since an MTP file is missing for existing PackML machines, we have taken a decision to generate an MTP-based representation of a PackML machine.

MTP file generation

A generic MTP-based representation of *any* PackML module is not feasible, due to the machine-specific differences between machines implementing PackML: the number of PackML machines within an OPC UA server, implemented optional states within the PackML state machine, available Pack Tags etc. Therefore, a PackML module-(type)-specific MTP representation needs to be generated out of the PackML unit instance representation.

We have chosen an approach to persist the OPC UA information of a module in a NodeSet format specified by OPC Foundation due to the lack of an agreed exchange format. Nodeset files are XML-based copies of OPC UA address spaces and can either be generated by engineering tools in an offline way or used to “snapshot” the content of a running OPC UA server. This fact contributes to the flexibility of the engineering process where the PackML-module description may be provided upfront, i.e., before the arrival of the actual product or taken from an already available physical module with an OPC UA communication interface. In the next step, an MTP-compatible representation of the PackML address space representation is created.

For demonstration purposes, we illustrate our approach on the example on an OPC UA PackML implementation provided by BaSys 4.2 Control Component (CC) [8] server with a suitable PackML profile. A description of this component can be found in Section 5.

First step is the mapping of available services. Each PackML machine (instance of PackMLBaseObject) in the OPC UA address space gets represented as an MTP service definition with a default procedure which represents the whole PackML machine. An example of such a mapping can be found in Fig. 3 where the expanded PackML machine on the left is mapped to the “PackMLBaseObject1” service within MTP (note, that we provide unique

names to MTP-services by adding a numbering suffix, since the example PackML server contains three PackML machine with same browse names).

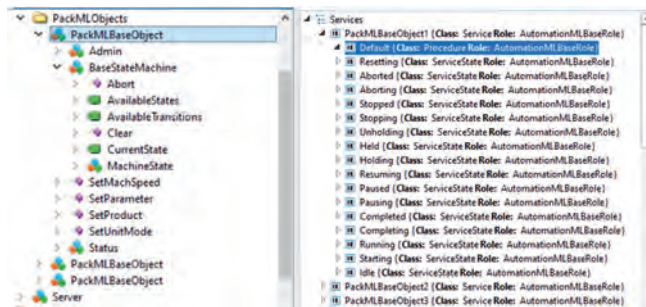


Fig. 3: Mapping of services with PackML OPC UA representation (left) and corresponding MTP (right).

Along with a default MTP procedure, available states within the PackML machine are mapped using a custom System Unit Class within MTPServiceSUC AutomationML library. This is needed, since every PackML machine instance may have a unique set of implemented states and pack tags.

Second step is creating service control instances for each service definition as shown in Fig. 4. In our example of three PackMLBaseObjects in the OPC UA server, we have three ServiceControl instances in the instance list. According to MTP specification ServiceControl instances are connected to the respective service definition, using the LinkedObject concept according to [1].

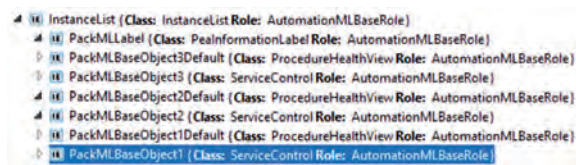


Fig. 4: ServiceControl instances for each PackML service.

The last step is to connect modeled service instances to its representation within OPC UA address space, i.e., OPC UA node IDs, of the PackML machine. This is done by modeling the PackML interface within a OPCUAObject of the generated MTP file. Here we use OPCUAObject class instances to store OPC UA node IDs of relevant nodes within the PackML machine: there include PackMLBaseObject and its sub-elements like methods (stored as OPCUAMethod instances) to invoke state machine transitions and contained hierarchical state-machines containing those. An example is shown in Fig. 5.

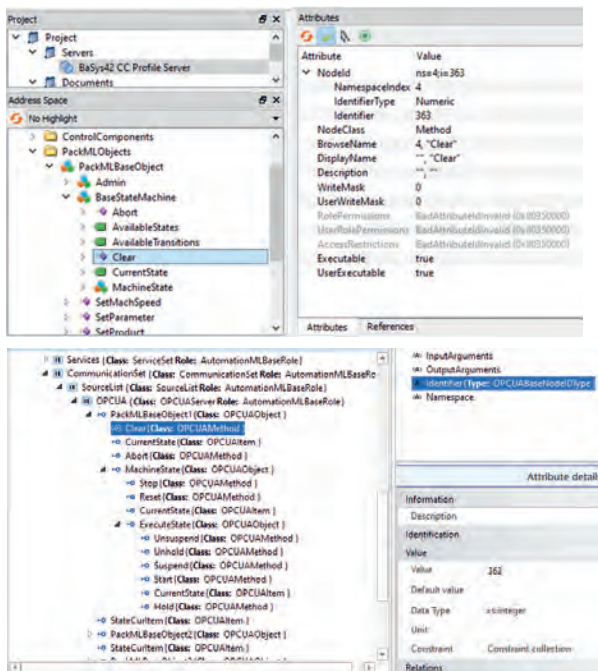


Fig. 5: Modeling of OPC UA node IDs within MTP.

The traceability between service control elements and OPC UA node representations is kept using the ID-Link concept, as described in [1]. Furthermore, ID-Links of OPC UA node representations of the PackMLBaseObjects are stored within the attributes of service control attribute like "CommandExt" to indicate the destination of OPC UA reads and writes, e.g., to determine the current state of the PackML state machine.

To sum up, the generated MTP file wraps a particular PackML machine representation and can be generated automatically out of a Nodeset file or an online OPC UA server. It contains all instance-specific information for a machine, like node IDs and implemented states.

PackML PEA Instantiation

With the creation of an MTP file wrapping the PackML state machine we can continue with the process of PEA instantiation out of this file.

The PackML-specific MTP file if first imported into the MTP-POL like any other MTP artifact. The MTP content is detected based on the used OPCUAObject and OPCUAMethod instances within the file and the POL sorts this module type into a specific category of modules type

(cf. tree on the left-hand side of Fig: 6 where “PackMLModuleTypePackge” is located after the import of PackML-specific MTP file).

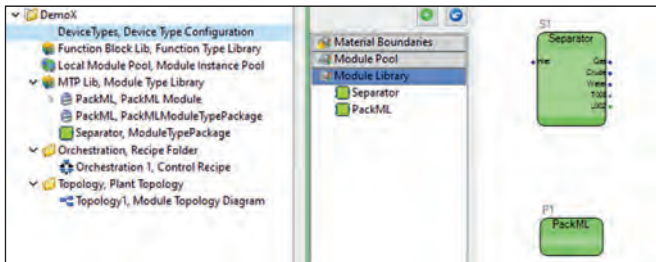


Fig: 6: PackML module instantiation within PackML-compatible MTP-POL prototype.

In the next step the module instance is created out of this module type using drag-and-drop operation into the right-hand part of the POL tool.

PackML has no HMI description – and thus no connection points, therefore no intuitive graphical connections are possible between the PackML instances within the engineering tool. Nevertheless, “stubs” can be added using the module HMI editor for material in and out from a PackML module to indicate machine's position within plant's material flow layout.

After the instantiation, some PEA-specific information needs to be completed including the IP address of the actual OPC UA endpoint provided by the PackML machine. Once this information is complete, we are clear for the runtime phase.

Runtime Orchestration

Runtime orchestration of MTP PEAs and PackML machines can happen manually by the operator using pre-generated faceplates or using recipes defined in the POL.

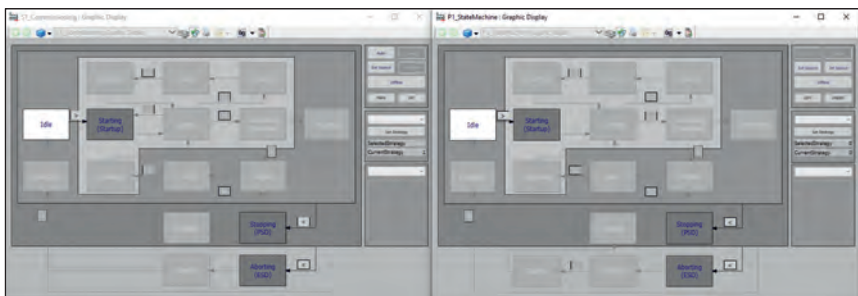


Fig: 7: Generated faceplates for manual operation of MTP (left) and PackML (right).

Generated faceplate for MTP and PackML are shown side-by-side in Fig. 7. Due to the similarity of the state models discussed before, a joint faceplate templates with a set of alternative elements are used which are shown or hidden based on the PEA type.

Behind the user-friendly interface wrt. the faceplate, some additional PackML-specific processing is needed within the POL during runtime. For example, one complexity related to the OPC UA Information Model representation of PackML state machines is the calculation of the current state of the module. The OPC UA model requires POL to continuously trace three "CurrentState" variables on each of three levels of the nested state machines within the OPC UA server and determine whether each level is currently activated or deactivated.

5. Validation

Our prototype was successfully evaluated on two software implementations of PackML OPC UA companion specification: an internal one and one obtained from BaSys¹ Control Component using PackML interface profiles. The bridge to BaSys CC via its PackML communication profile allows our MTP-compatible POL prototype to operate any unit which is compatible to BaSys automation architecture. The implementation of a test server including a PackML profile can be found via git repository² of RWTH Aachen University.

6. Summary and Outlook

In this paper we presented an approach for orchestration of hybrid plants – joint production facilities including equipment from both the process automation and discrete automation. For modular plants such a hybrid setting requires a side-by-side engineering and operation of MTP and PackML modules. For an MTP-centric orchestration, i.e., orchestration from an MTP-compatible POL, our approach includes generation of PackML-specific MTP files which can be used in a same engineering and operation workflow as standardized MTP files and derived PEAs. Our prototype of PackML-enabled POL was successfully validated with two independent software implementations of PackML OPC UA companion specification.

Despite successful initial validation, the work on the prototype is not finished and requires a number of extensions to fully cover the PackML specification including:

- PackTags within the PackML interfaces and their mapping to service parameters,
- Full support of service control properties, i.e., mapping procedure selection to custom operation modes of PackML.

¹ https://wiki.eclipse.org/BaSys/_/Documentation/_/API/_/ControlComponentProfiles

² <https://git.rwth-aachen.de/acplt/basys4.2/ccProfilesUA>

- Support of parameterized start method within PackML model.
- Adaptation of the presented compatibility measures to the updates of MTP state machine which expected in the upcoming final specification release.

Acknowledgements

We would like to thank Mr. Julian Grothoff from RWTH Aachen University for providing PackML-compliant modules for testing purposes. Authors were partially supported by German Federal Ministry of Education and Research in the scope of the BaSys 4.2 project (funding code 01IS19022).

References

- [1] VDI/VDE/Namur 2658: "Automation engineering of modular systems in the process industry." Part 1: "General Concept and Interfaces." 2019-10
- [2] IEC 62714 Edition 1.0: "Engineering data exchange format for use in industrial automation systems engineering - Automation Markup Language." 2014
- [3] ANSI/ISA-TR88.00.02-2015: "Machine and Unit State: An implementation example of ANSI/ISA-88.00.01," 2015
- [4] OPC 30050 RC1.01: "OPC UA for PackML," 2020-11
- [5] VDI/VDE/Namur 2658: "Automation engineering of modular systems in the process industry." Part 4 (Draft): "Modelling of module services." 2020-08
- [6] Grüner, S. et al.: "Cross -Industry State of the Art Analysis of Modular Automation." Automation 2020, Baden-Baden, 2020-06
- [7] Grüner, S. et al.: "A Reference Architecture for Modular Industrial Automation Systems" IEEE ETFA 2021, Vasteras, 2021-09
- [8] Grothoff, J. und Kleinert, T.: "Mapping of Standardized State Machines to Utilize Machine Learning Models in Process Control Environments", ESTEP 2020, Pisa, 2020-10

Selbstorganisation von Softwarekomponenten für produktive Aufgaben

Self-organization of software components for productive tasks

Dr.-Ing. **M. Riedl**, ifak e.V., Magdeburg;
Dr.-Ing. **H. Zipper**

1 Kurzfassung

Smart Factories gewinnen ihre Flexibilität und Anpassungsfähigkeit durch den Einsatz verteilter, kooperierender Systeme. Darauf aufbauend lassen sich Szenarien zur Selbstorganisation von Produktionsaufgaben umsetzen. Der Beitrag zeigt dafür einen Vorschlag, bei dem gleichzeitig durch eine formale Prüfung sichergestellt wird, dass die Arbeitsteilung auch erfolgreich umgesetzt werden kann. Das Konzept ist bis auf die Aufteilung der Arbeitsaufgabe in Teilprozesse generisch angelegt.

2 Abstract

Smart factories achieve their flexibility and adaptability through the use of distributed, cooperating systems. Based on this, scenarios for the self-organization of production tasks can be implemented. The article shows a proposal for this purpose, in which a formal test ensures that the division of tasks can also be successfully implemented. The concept is generic except for the division of the work task into sub-processes.

3 Motivation

Heutige Softwarestrukturen in Komponenten eines Produktionssystems wie Roboter, Maschinen oder Logistiksystemen sind nur bedingt in der Lage, Herausforderungen zur Selbstorganisation eines Produktionsprozesses zu bewältigen. Dies betrifft ebenso den Informationsfluss zwischen diesen Komponenten, der momentan überwiegend durch etablierte industrielle Kommunikationssysteme ermöglicht wird. Benötigt werden dafür jedoch autonome, dynamisch kooperierende Produktionskomponenten, die flexibel und zuverlässig miteinander interagieren. Dabei soll der Managementaufwand für den Anlagenplaner minimal sein und der spätere Anwender darf sich mit spezifischen Konfigurationen nicht befassen müssen.

Stand heute ist eine Anpassung des Produktionsprozesses aufgrund der hierarchischen Struktur der zentralen Fabrikleit- und Steuerungssysteme zeitaufwändig und teuer. Modularisierung und eine dynamische Machine-to-Machine (M2M)-Kommunikation bilden die Grundlage für Smart Factories und leistungsfähige und flexible Produktionssysteme von morgen. In den Steuerungssystemen werden die Softwaremodule Datenakquise, Regelung und Ansteuerung anhand der Anforderungen der zu steuernden Aufgabe mehr oder weniger als Einheit implementiert. Dabei ist die Hardware (Knoten), auf denen die Software läuft, von vornherein (in etwa) bekannt. Die Komposition der Module geschieht durch den Entwickler. Auch die Abstimmung der Parameter der einzelnen Module aufeinander muss durch den Entwickler erfolgen. Ein Trend in der industriellen Automatisierung geht dahin, die IoT-Technologie an die Vorverarbeitung von Daten anzupassen oder das Smart Manufacturing bei der Entwicklung kooperierender Systeme zu unterstützen [1][2]. Dabei sollen Methoden zur automatischen Aggregation und Disaggregation verteilter Produktionskomponenten die geforderte Flexibilität in der Interaktion ermöglichen und auch die Resilienz gegenüber Netzwerkausfällen erhöhen. Auch soll damit das Ausfallrisiko verringert werden, da das bisher angewandte komplexe und zentrale Management vereinfacht wird. Ebenso muss das Engineering verteilter Systeme so verbessert werden, dass dieses beispielsweise auf Merkmalen aufsetzt oder die Systemsimulation besser integriert.

Der Beitrag ist wie folgt strukturiert: Kapitel 2 betrachtet die technischen Möglichkeiten, die mit dem heutigen Stand der Technik umsetzbar sind und weist auf erfolgversprechende Forschungsarbeiten hin, die sowohl beim Entwurf des Ansatzes (Kapitel 3) als auch bei der Umsetzung einer Beispielanwendung (Kapitel 4) betrachtet wurden. Eine Zusammenfassung schließt den Beitrag ab.

4 Herausforderungen vs. Stand der Technik

Der anhaltende Trend zur Digitalisierung in der Industrie führt zu einer zunehmenden Akzeptanz von verteilten Steuerungssystemen. Derzeit ist die Standardisierung im Zusammenhang mit Industrie 4.0 ein großer Treiber für den Einsatz von verteilten Steuerungssystemen in der Industrie. Mit den Vorschlägen aus [3], [4] und [5] werden verteilte Steuerungssysteme für eine flexible und adaptive Produktion ausgelegt.

Die in industriellen Umgebungen verwendeten Kommunikationsprotokolle entwickeln sich von Protokollen, für die der Ingenieur die Bedeutung jedes Bits kennen muss, zu semantischen Protokollen, die Aufgaben und Anlagen auf der Grundlage ihrer Anforderungen, Eigenschaften und Fähigkeiten beschreiben. In [6] wird ein solches Protokoll unter Verwendung eines Ontologie-Informationsmodells gezeigt.

In [7], [8] wurde ein semantisches Protokoll für die fähigkeitsbasierte Steuerung von Robotern gezeigt. Es werden Ontologien verwendet, um die Daten wie Aufgaben, Greifer und 3D-Geometrien zu beschreiben. Als Kommunikationsmittel wird OPC-UA verwendet. Der Vorteil von OPC-UA ist das semantische Modell.

Der Standard IEC 61499 [9] erweitert den etablierten Standard IEC 61131 für den Einsatz in verteilten, ereignisbasierten Systemen. Neuere Forschungsaktivitäten wie [9], wenden OPC UA auf IEC 61499-Systeme an und realisieren damit Industrie 4.0-konforme Steuerungssysteme. Auf der anderen Seite werden serviceorientierte Architekturen erforscht. Ein Beispiel ist das Arrowhead [10] Framework.

Verteilte Steuerungssysteme (DCS) sind Systeme, die aus räumlich getrennten Steuerungen bestehen, die ihrerseits mit Aktoren, Sensoren usw. interagieren, die über ein Netzwerk verbunden sind. Obwohl sie intensiv erforscht werden, bleibt eine große Herausforderung in diesem Bereich eine stabile und gut funktionierende Steuerung unter Berücksichtigung von Beschränkungen der Kommunikation, zu denen vor allem unerwartete Verzögerungen und Paketausfälle zählen [11]. Aktuelle Forschungen zielen auf die Stabilisierung dieser Systeme z. B. zur Behandlung von Verzögerungen [12], [13], [14], während der Entwurf von DCS, die gleichzeitig Netz- und Reglerentwurf berücksichtigen, ein offenes Feld bleibt [15], [16].

5 Ansatz

Der gewählte Ansatz berücksichtigt die in Abschnitt 4 skizzierten aktuellen technologischen Möglichkeiten und greift interessante Forschungsideen auf. So wurde der Ansatz darauf ausgelegt, dass ein fähigkeitsbasierter Arbeitsablauf organisiert wird, der die Eigenschaften einer Ressource als Parameter der Fähigkeiten auffasst. Fähigkeiten sind eine abstrakte Beschreibung dessen, was eine Entität zu leisten vermag [17]. Fähigkeiten sind assoziiert mit Eigenschaften. Eine Eigenschaft beschreibt mit Minimal- und Maximalwerten den Wirkungsbereich einer Fertigkeit. Eine Zusammenarbeit von Produktionskomponenten erfolgt stets auf Basis von Ausschreibungen / Verhandlungen.

Die beteiligten Automatisierungskomponenten werden als Module betrachtet, die spezifische Aufgaben erfüllen können, z. B. ein Temperaturmodul besitzt die Fähigkeit zur Messung einer Temperatur oder ein Kran besitzt die Fähigkeit ein Werkstück zu bewegen. Die Module können ihre Fähigkeiten in Form von Metadaten beschreiben.

Diese Ziele führen zu einem lose gekoppelten System. Somit wird das Steuerungsprogramm der Produktionsanlage kein statisches SPS-Programm mehr sein, wie es heute der Fall ist. Stattdessen wird ein verteilter Steuerungsablauf adhoc erstellt, der auf den erforderlichen

Aufgaben und den verfügbaren Ressourcen wie Fertigungsanlagen und Kommunikationskapazitäten basiert. Zudem kann sich der Steuerungsablauf ändern, wenn Fehler auftreten oder Änderungen bei den oben genannten Ressourcen auftreten.

Der Ansatz zur Selbstorganisation wird in einer Applikationsschicht ausgeführt, die dem Steuerungsablauf noch vorgelagert ist. Die dabei wichtigen Komponenten sind das Mission Management (Verwaltung der Gesamtaufgabe) und der Proxy für das Modul.

Das Mission Management erhält Aufgaben aus dem ERP System. Diese Aufgaben sind komplex, wie z. B. "Produziere ein Produkt mit den angegebenen Eigenschaften". Im Mission Management wird diese übergeordnete Aufgabe in Teilaufgaben zerlegt. Die Teilaufgaben sind anschließend so atomisiert, dass eine Teilaufgabe einer Gruppe von Modulen zugewiesen werden kann, die diese Teilaufgabe ausführen und über Erfolg / Misserfolg berichten. Teilaufgaben repräsentieren ihre Anforderungen als Eigenschaften - Tupel von Typ, Minimalwert, Maximalwert. Für das oben erwähnte Temperaturmodul würde ein Tupel für die Fähigkeitsanforderung wie folgt sein: {„Bereitstellung eines Temperaturmesswertes“, min=200K, max = 300K}.

Der Proxy des Moduls ist eine Softwarekomponente, die auf dem Konzept der Verwaltungsschale basiert. Diese Softwarekomponente kann sich innerhalb der IoT-Komponente oder in einer (Edge-)Cloud befinden. Es gibt keine zentral organisierte Zuweisung von Teilaufgaben auf IoT-Komponenten. Stattdessen beziehen Proxies und damit die IoT-Komponenten in Abhängigkeit ihrer Auslastung, ihres Zustands und ihrer Programmierung selbständig Teilaufgaben im Pull-Verfahren vom Mission Management. Um eine Teilaufgabe auszuführen, prüft sie die Vorbedingungen für die Teilaufgabe und organisiert eine kooperative Strategie mit anderen Proxies. Die Strategie zur Durchführung einer Teilaufgabe kann jedoch nicht nur die Anforderungen dieser Teilaufgabe berücksichtigen, sondern muss auch andere Teilaufgaben und die verfügbaren Kommunikationsressourcen beachten.

Im Prinzip kann die Verhandlung als ein Anpassungs- und Optimierungsproblem aus den Anforderungen einer Teilaufgabe und den Fähigkeiten der verfügbaren Entitäten aufgefasst werden. Vielversprechende Ergebnisse zu solchen Problemen wurden in der Forschung vorgestellt [18]. Dort wurde das Problem als Satisfiability Modulo Theories (SMT)-Logik formuliert. Die Autoren zeigen, dass der Stand der Technik der SMT Solver für die Optimierung verwendet werden können und schnell Lösungen finden. Dieser Ansatz wurde erfolgreich übernommen.

6 Beispielanwendung

Die Funktionsweise zur Selbstorganisation solle anhand des folgenden Simulationsbeispiels verdeutlicht werden. In einem momentan in der Umsetzung befindlichen Demonstrator des BMBF-Projektes VERITAS soll eine Wand mit einem beliebigen Schriftzug aus Einzelsteinen kooperierend zusammengesetzt werden. Als Produktionskomponenten stehen Roboter und ein autonomes Transportfahrzeug (AGV) zur Verfügung. Als Verarbeitungsgegenstände dienen Legosteine unterschiedlicher Länge und Farbe, so dass daraus ein Schriftzug gebaut werden soll. Die Roboter arbeiten in einem gemeinsamen Arbeitsraum. Die Steine liegen in zwei Regalen: Ein Regal für weiße Steine und ein Regal für rote Steine. Die Regale können durch eine eingebaute Technik Steine auf ein davor anhaltendes AGV legen. Dieses AGV kann dann jeweils diesen Stein zu einem der Roboter fahren. Dort nimmt der Roboter den Stein herunter und die Mauer mit diesem Stein bauen. Folgende Arbeitsschritte werden ausgeführt:

1. Von einem übergeordneten System (ERP) wird eine neue Aufgabe in das System gegeben → Mauer mit Schriftzug bauen
2. Durch einen aufgabenspezifischen Algorithmus wird diese Aufgabe in mehrere Teilaufgaben zerlegt.

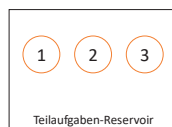


Bild 1: Vorrat an Teilaufgaben ohne Abhängigkeiten

3. Die Teilaufgaben können jedoch noch Abhängigkeiten enthalten. Im Beispiel muss aufgrund produkt- und produktionsspezifischer Gründe die Teilaufgabe 1 nach Teilaufgabe 2 und Teilaufgabe 3 erledigt werden, siehe Bild 2 (links), die Teilaufgaben 2 und 3 besitzen keine Abhängigkeiten und können vielleicht sogar parallel erledigt werden (hängt natürlich von Platz, Ressourcen usw. ab.)



Bild 2: Vorrat an Teilaufgaben

4. Alle Teilaufgaben werden im Aufgabenreservoir des Mission Management abgelegt. Wird eine Teilaufgabe erledigt, verschwindet sie aus dem Reservoir. Dadurch verschwinden sie auch als Abhängigkeit, siehe Bild 2 (links Mitte). Aus formaler Sicht bedeutet dies, dass zunächst alle Teilaufgaben bearbeitet werden, welche keine ausgehenden Kanten haben.
5. Einige Module im System können Strategien zur Umsetzung bestimmter Teilaufgaben besitzen. Es kann durchaus passieren, dass unterschiedliche Module ganz unterschiedliche Strategien zur Umsetzung ein und derselben Teilaufgabe haben. Im Beispiel besitzen die Roboter eine Strategie zum Setzen eines Steins.
6. Ein Modul kann von sich aus entscheiden, wenn es sich um die Bearbeitung einer Teilaufgabe kümmern möchte. Im Beispiel entscheiden sich die Roboter dies zu tun, sobald sie nichts anderes mehr zu tun haben.
7. Entscheidet sich ein Modul zur Bearbeitung einer neuen Teilaufgabe, fragt es beim Aufgabenreservoir nach. Ist noch eine Teilaufgabe vorhanden, wird sie aus dem Aufgabenreservoir herausgenommen (siehe Schritt 4). Die Abhängigkeiten zu dieser Teilaufgabe bleiben weiterhin im Aufgabenreservoir vermerkt. Dieses Modul ist nun das organisierende Modul für diese Teilaufgabe.
8. Das organisierende Modul entscheidet nun anhand seiner Strategie sowie der Anforderungen der Aufgabe, wie diese Aufgabe durchgeführt wird. Dabei kann es passieren, dass das organisierende Modul nicht alle Fähigkeiten zur Durchführung besitzt. Im Beispiel hat der Roboter zwar die Fähigkeiten zum Setzen der Steine, nicht jedoch zum Abholen aus dem Lager.
9. Für jede notwendige Fähigkeit fragt das organisierende Modul alle anderen Module im System: "Ich benötige jemand, der diese Fähigkeit mit jenen Anforderungen (Zeit von..bis, Ort, Form, ...) bereitstellen kann". "Wer kann Steine fahren, wer hat einen roten Stein, ..."
10. Jedes andere Modul im System entscheidet für sich, ob und wie es helfen kann. Kann es helfen, antwortet es mit einer Beschreibung ihrer Hilfe: "Ich kann helfen und biete eine Fähigkeit mit folgenden Parameter und Kosten (Energie, Kommunikation) zur Zeit XYZ an. "AGV: „Ich kann in 3 Minuten einen Stein fahren“, Regal: „Ich habe rote Steine der Längen 2 und 4“.
11. Das organisierende Modul sammelt alle Antworten ein. Aus allen Antworten und den Anforderungen der Teilaufgabe wird ein Optimierungsproblem aufgestellt. Dessen Lösung liefert die folgenden Informationen:

- a. Menge der Module, welche zur Umsetzung der Teilaufgabe benötigt werden. Diese Module werden als ausführende Module bezeichnet. Im Beispiel wären ausführende Module möglicherweise Roboter 1, Regal rot, AGV
 - b. Menge von Jobs, genannt Plan.
 - c. Jeder Job beschreibt, wann ein bestimmtes ausführendes Modul seine Fähigkeit einsetzt und mit welchen Parametern.
 - d. Jobs müssen in einer Reihenfolge ausgeführt werden. Dazu enthält jeder Job eine Liste von Jobs, die vorher stattfinden müssen. Im Beispiel: "Job1 = AGV fahre zum Regal 1, Job2=Regal 1 belade AGV..."
12. Das organisierende Modul versendet den Plan inklusive der Jobbeschreibungen an alle ausführenden Module. Das organisierende Modul kann ein ausführendes Modul sein, muss aber nicht.
 13. Sobald alle Vorbedingungen eines Jobs erfüllt sind, kann ein ausführendes Modul diesen Job ausführen. Sobald der Job erledigt ist, wird dies an alle anderen ausführenden Module gemeldet. Diese können daraufhin die Vorbedingungen ihrer gespeicherten Jobs aktualisieren. Außerdem wird das Erledigen eines Jobs an das organisierende Modul gemeldet. Dieses aktualisiert den Zustand des Plans.
 14. Sind alle Jobs eines Plans erledigt, meldet das organisierende Modul dies dem ERP.

Die Selbstorganisation wird als Optimierungsproblem mit Nebenbedingungen beschrieben. Dieses Optimierungsproblem wird in SMT aufgestellt und mit Z3 als Solver gelöst. Zwischen den Punkten 8 und 11 der obigen Liste wird das Optimierungsproblem in mehreren Phasen aufgestellt. Um eine Zuordnung zwischen Statements in SMT und den Informationen zur Selbstorganisation zu ermöglichen, erfolgt eine Kodierung in den Namen, wobei hier grob die 5 Phasen vorgestellt werden:

Phase 1: findet auf dem organisierenden Modul statt. Es werden die aus der Strategie resultierenden Informationen und Anforderungen dem Optimierungsproblem hinzugefügt.

Phase 2: findet auf allen gefragten Modulen statt. Hier werden für Fähigkeiten, welche zur Anfrage passt, Informationen zum Optimierungsproblem hinzugefügt.

Phase 3: Das organisierende Modul fügt zusätzliche SMT-Statements für jede angeforderte Fähigkeit in Abhängigkeit der empfangenen Antworten ein. Diese sollen sicherstellen, dass zu jeder in der Strategie geforderten Fähigkeit tatsächlich nur eine Ausgewählt wird. Weiterhin wird hier die Berechnung der Gesamtkosten sowie die Berücksichtigung der Startzeitpunkte der Fähigkeiten erledigt: Für jede angeforderte Fähigkeit werden daher weitere Informationen zum Optimierungsproblem hinzugefügt.

Phase 4: Nachdem alle im System vorhandenen Fähigkeiten bekannt sind, wird die Zielfunktion hinzugefügt, z. B. $\min(T_{\text{end}} + \text{Cost}_{\text{Teilaufgabe}_0} + \dots + \text{Cost}_{\text{Teilaufgabe}_n})$

Phase 5: Nach erfolgreicher Prüfung / Optimierung durch Z3 wird ein Modell extrahiert. Dieses Modell enthält Definitionen für die zuvor deklarierten Funktionen (technisch werden in SMT alle Variablen als Funktionsdeklaration abgelegt). Aus diesem Modell werden dann die Informationen extrahiert:

- Reihenfolge der Anwendung der Fähigkeiten
- Startzeitpunkte der einzelnen Fähigkeiten
- Welche konkreten Fähigkeiten genutzt werden
- Parameter der Fähigkeiten

Analog zu existierenden Automatisierungssystemen kann zu jeder Zeit der Status einer Teilaufgabe überwacht werden. Nachdem eine Teilaufgabe aus ihrem Pool genommen wird, wird ihre Bearbeitung weiterverfolgt. Gibt es bei der Bearbeitung ein Problem, darf die Aufgabe nicht verschwinden, sondern muss zurückgelegt werden. Dies ist je nach Art der Aufgabe, ihrem Bearbeitungsfortschritt und den ausführenden Komponenten mit mehr oder weniger komplizierten Fail-Over Mechanismen verbunden.

Folgende Designentscheidungen wurden bei der Umsetzung der Module in IoT-Geräte getroffen, siehe auch Bild 3:

- Jedes IoT Gerät hat eine Job Queue.
- Job Queue leer $\rightarrow$ IoT Gerät sucht sich eine neue Aufgabe.
- Der Einsatz einer Fähigkeit kann zu mehreren Jobs führen. Alle Jobs kommen der Reihenfolge nach in die Job-Queue



Bild 3: IoT-Gerät mit Job Queue

7 Zusammenfassung

Der Beitrag zeigt einen Ansatz zur Flexibilisierung der Produktionsprozesse auf Basis autarker Komponenten, die dynamisch zur Bearbeitung von Aufgaben zusammenarbeiten. Grundlage

dieser Zusammenarbeit ist das Aushandeln von passenden Aufgaben bzw. der autarken Suche nach Kooperationspartnern. Der gezeigte Ansatz ist generisch, einzig die Aufteilung der Gesamtaufgabe in Teilaufgaben muss spezifisch umgesetzt werden, da diese Aufteilung sich von Produkt zu Produkt unterscheidet. Die Funktionsfähigkeit des Ansatzes wurde in einem Simulationssystem bereits unter Beweis gestellt, reale Komponenten werden auf Basis der Robot Operating System Infrastruktur diese Arbeitsteilung umsetzen. Die Autoren planen, die gezeigten Konzepte auch in weiteren verteilten Steuerungskonzepten umzusetzen.

8 Danksagung

Die Autoren danken dem Bundesministerium für Bildung und Forschung für die Unterstützung und den Partnern des BMBF-Projekts VERITAS - Verteilte Emergente Regelung im IoT mittels Agil kollaborierender Systeme (Förderkennzeichen: 01IS18073C) für ihre Arbeit und Beiträge, die diesen Beitrag ermöglicht haben.

9 Literatur

- [1] D. Zuehlke, "SmartFactory-Towards a factory-of-things", Annual reviews in control, vol. 34(1), pp. 129–138, 2010.
- [2] S. Jeschke, C. Brecher, T. Meisen, D. Özdemir, and T. Eschert, "Industrial internet of things and cyber manufacturing systems", in Industrial Internet of Things, pp. 3–19, 2017.
- [3] Interaction Model for Industrie 4.0 Components, Federal Ministry for Economic Affairs and Energy (BMWi) Std., March 2016.
- [4] Details of the Asset Administration Shell - Part 1, Federal Ministry for Economic Affairs and Energy (BMWi) Std., May 2020.
- [5] D. Vögeli, S. Grabmaier, M. Jüttner, M. Weyrich, P. Göhner, and W. M. Rucker, "Intelligent and distributed solving of multiphysics problems coordinated by software agents", in International Conference on Agents and Artificial Intelligence, Madeira, Portugal, 2018.
- [6] A. Perzylo, J. Grothoff, L. Lucio, M. Weser, S. Malakuti, P. Venet, V. Aravatinos, and T. Deppe, "Capability-based semantic interoperability of manufacturing resources: A basys 4.0 perspective", IFAC-PapersOnLine, vol. 52, no. 13, pp. 1590–1596, 2019.
- [7] S. Profanter, A. Breitkreuz, M. Rickert, and A. Knoll, "A hardwareagnostic OPC UA skill model for robot manipulators and tools", in 2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA). IEEE, 2019, pp. 1061–1068.

- [8] P. Zimmermann, E. Axmann, B. Brandenbourger, K. Dorofeev, A. Mankowski, and P. Zanini, "Skill-based engineering and control on field-device-level with OPC UA", in 2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA). IEEE, 2019, pp. 1101–1108.
- [9] T. Terzimehic, M. Wenger, A. Zoitl, A. Bayha, K. Becker, T. Müller, and H. Schauerte, "Towards an industry 4.0 compliant control software architecture using IEC 61499 & OPC UA", in 2017 22nd IEEE International Conference on Emerging Technologies and Factory Automation (ETFA). IEEE, 2017, pp. 1–4.
- [10] P. Varga, F. Blomstedt, L. L. Ferreira, J. Eliasson, M. Johansson, J. Delsing, and I. M. de Soria, "Making system of systems interoperable– the core components of the Arrowhead framework", *Journal of Network and Computer Applications*, vol. 81, pp. 85–95, 2017.
- [11] X. Ge, F. Yang, and Q.-L. Han, "Distributed networked control systems: A brief overview", *Information Sciences*, vol. 380, pp. 117–131, 2017.
- [12] L. Ren, S. Xie, Y. Zhang, J. Peng, and L. Zhang, "Chattering analysis for discrete sliding mode control of distributed control systems", *Journal of Systems Engineering and Electronics*, vol. 27, no. 5, pp. 1096–1107, 2016.
- [13] G. Lympopoulos and P. Ioannou, "Adaptive control of networked distributed systems with unknown interconnections", in 2016 IEEE 55th Conference on Decision and Control (CDC). IEEE, 2016, pp. 3456–3461.
- [14] Y. Alipouri and B. Huang, "Distributed multiple step ahead prediction considering communication delays", *IET Control Theory & Applications*, vol. 13, no. 6, pp. 806–814, 2019.
- [15] J. P. Hespanha, P. Naghshtabrizi, and Y. Xu, "A survey of recent results in networked control systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 138–162, 2007.
- [16] D. Zhang, P. Shi, Q.-G. Wang, and L. Yu, "Analysis and synthesis of networked control systems: A survey of recent advances and challenges", *ISA transactions*, vol. 66, pp. 376–392, 2017.
- [17] K. Evers, J. R. Seyler, V. Aravantinos, L. Lucio, and A. Mehdi, "Roadmap to skill based systems engineering", in 2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA). IEEE, 2019, pp. 1093–1100.
- [18] S. F. Roselli, K. Bengtsson, and K. A. Kesson, "SMT solvers for job-shop scheduling problems: Models comparison and performance evaluation", in 2018 IEEE 14th International Conference on Automation Science and Engineering (CASE). IEEE, 2018, pp. 547–552.

Vertrauensbasierte Integration von Geräten mittels OPC UA Device Provisioning

F. Volkmann, N. Baur, S. Höme, Siemens AG, Nürnberg;
A. Palmin, M. Madsen, Siemens AG, Karlsruhe

1. Kurzfassung

Mit der Digitalisierung und der wachsenden Sensibilität für Sicherheit werden neue Konzepte für eine vertrauensbasierte Integration von Geräten benötigt. Verschiedene Standards bieten Lösungen, um eine erste Vertrauensbasis zwischen Geräteinstanzen und Laufzeitumgebungen herzustellen – also eine sichere Identifikation der Kommunikationspartner zu ermöglichen. U.A. sind hier BRSKI (Bootstrapping Remote Secure Key Infrastructure, RFC 8995) oder FDO (Fido Alliance Device Onboarding) zu nennen. Aber auch Kommunikationsstandards, die aus dem Automatisierungsumfeld bekannt sind, nehmen sich vermehrt dem Thema an. Das OPC UA Device Provisioning spezifiziert eine sichere Identifikation von Geräteinstanzen unter Nutzung der OPC UA Infrastruktur. Um dieses Ziel zu erreichen, müssen alle Akteure ihren Teil beitragen, wie im Folgenden ausgeführt wird. Akteure sind insbesondere Hersteller, Distributoren, Maschinenbauer und Endverbraucher. Zur Vereinfachung wird hier der Hersteller und Endverbraucher betrachtet. Da beim Endverbraucher Geräte und Software aus unterschiedlichsten Bereichen aufeinandertreffen, ist es fundamental, die Verwaltung solcher Geräteinstanzen, bzw. Security Informationen und Konfigurationen, einheitlich zu verwalten.

Dies kann nicht ausschließlich auf der Kommunikationsebene von Gerät und Umgebung erfolgen, da diese in der Praxis zu heterogen ist. Vielmehr muss es möglich sein die Verwaltung auf einer höheren Ebene zu vereinheitlichen und zusammenzuführen.

Dies wird erreicht, indem die OT-Ebene (hier OPC UA) an eine etablierte PKI (Public Key Infrastructure) in der IT-Ebene angebunden wird und somit die Verwaltung der Security an diese abgegeben wird.

Dieses Dokument basiert auf der noch nicht veröffentlichten Draft Nummer 50 der Spezifikation OPC UA 10000-21 Device Provisioning [1].

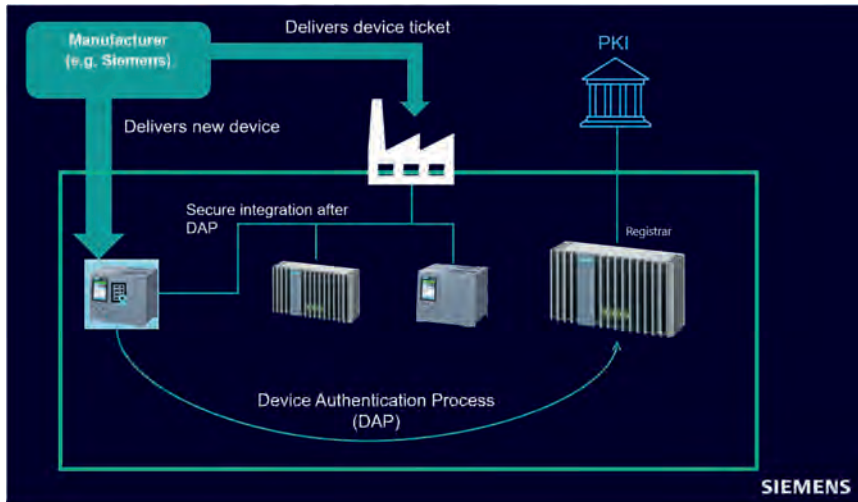


Bild 1: Integration einer Geräteinstanz mit Hilfe des Device Authentication Process (DAP)

2. OPC UA Part 21 Device Provisioning

Mit der neuen Spezifikation OPC UA Part 10000-21 Device Provisioning erweitert die OPC Foundation ihr Portfolio um die Möglichkeit, über OPC UA ein sicheres provisionieren von Geräten in eine OPC UA Infrastruktur. Dabei fokussiert der Standard auf die Geräte, ihren Lebenszyklus und die Möglichkeit, Geräte zu komplexeren Gebilden zusammenzufassen und diese wieder als Geräte aufzufassen (Composites).

Andere Device Provisioning Mechanismen, wie z.B. BRSKI oder FDO führen eine mutual Authentifizierung durch. Dies bedeutet, dass sich einerseits das Gerät gegenüber der Umgebung und andererseits die Umgebung gegenüber dem Gerät authentifiziert. Durch eine erfolgreiche gegenseitige Authentifizierung wird erreicht, dass das Gerät und die Umgebung dem Gerät vertraut. Die OPC UA-Geräte Provisionierung führt nur eine einseitige Authentifizierung durch, damit die Umgebung anschließend dem Gerät vertraut.

Die OPC Foundation erlaubt bzw. unterstützt explizit die Zusammenarbeit mit anderen Device Provisioning Mechanismen.

Das OPC UA Device Provisioning erlaubt bzw.- unterstützt Szenarien wie:

- Sichere Identifikation von Geräteinstanzen
- Sicherer Übergang der Geräteinstanzen von Sicherheitsdomänen
 - o Nachvollziehbarkeit des Vertriebsweges
- Wechseln der benutzten Autorität bei der Geräte Provisionierung
 - o Wahlweises Geräte Provisionierung über Hersteller-, Distributor- oder Maschinenhersteller Autorität.
- Automatisierte Installation und Softwarebereitstellung

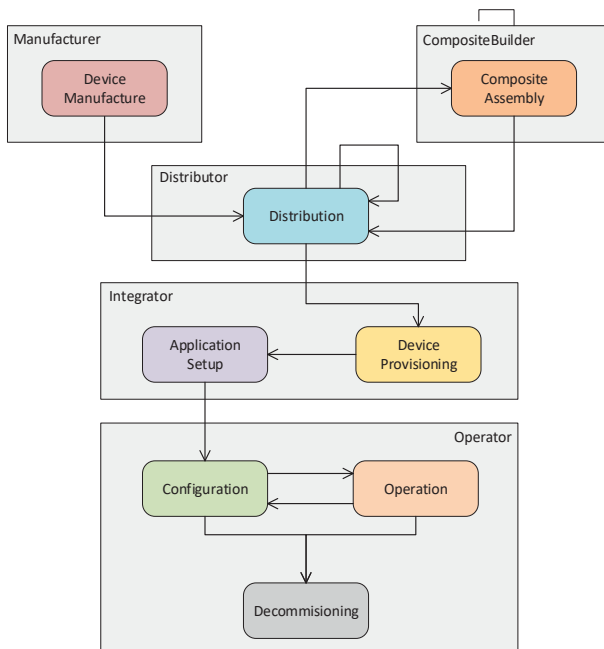


Bild 2: Akteure des DP gespiegelt am Geräte-Lebenszyklus

Zusammenfassend ist zu sagen, dass das Device Provisioning Verwaltungsaufgaben vereinfacht und automatisiert, so dass Fehlerquellen eliminiert werden und allgemein die Kosten der Verwaltung verringert werden.

Der Hersteller eines Gerätes, der Endkunde, sowie alle anderen Akteure (CompositeBuilder¹, Distributor²), die im Rahmen des Device Provisionings als Autorität agieren wollen, stellen eine passende Infrastruktur bereit. Die Infrastruktur, in der das Gerät provisioniert werden soll, beinhaltet unter anderem die Bereitstellung eines Registrars für den Zugriff auf eine Certification Authority (kurz CA) bzw. Registration Authority (RA).

2.1 IDevID Zertifikat und Ticket

Der Hersteller erstellt bei der Geräteherstellung ein IDevID Zertifikat, sowie ein Geräte Ticket. Der vorzugsweise hardwaregebunden im Gerät sicher hinterlegte privaten Schlüssel und das zugehörige IDevID-Zertifikat sowie die Zertifikate der zugehörigen Certification Authority (CA) und der zugehörigen Root CA werden mit dem Gerät ausgeliefert. Das zugehörige Geräteticket wird über einen anderen Weg an den jeweiligen Käufer übermittelt. Die Verbindung zwischen dem IDevID Zertifikat und dem Geräte Ticket wird über eine eindeutige ProduktInstanceURI hergestellt. Das IDevID-Zertifikat und das zugehörige Ticket sind durch die Signatur der zuständigen CA des Herstellers gegen unberechtigte Manipulationen geschützt (bzw. integritätsgeschützt). Die CA des Herstellers garantiert die Herkunft und Authentizität der beiden Komponenten.



Bild 3: ProductInstanceURI verbindet Ticket und IDevID Zertifikat

¹ Maschinenbauer, kombiniert mehrere Einzelgeräte zu einem komplexen Ganzen. Zum Beispiel Steuerung, Feldgeräte und Motoren zu einer Maschine.

² Wiederverkäufer, kauft Einzelgeräte oder Maschinen, um diese weiterzuverkaufen. Eine Veredelung findet nicht statt.

Andere Akteure können zusätzliche Tickets mit derselben ProductInstanceURI erstellen und über ihre eigene CA absichern. In diesem Fall obliegt es dem Kunden durch Auswahl des Tickets zu entscheiden, welcher CA Autorität er vertraut.

Ein Maschinenbauer (CompositeBuilder) kombiniert mehrere Geräte zu einem komplexeren Gebilde. Dieses wird im Device Provisioning dadurch dargestellt, dass er LDevID Zertifikate oder Tickets mit eigener CompositeInstanceURI erstellt. Diese enthalten zusätzlich alle ProductInstanceURIs der beinhaltenden Geräte. LDevID Zertifikate und Tickets werden durch die CA des Maschinenbauers abgesichert.

2.2 Prozess zur Geräte Provisionierung

Das Gerät als OPC UA Client wird das erste Mal in der Endkunden Infrastruktur aufgenommen, hierbei werden folgende Schritte durchgeführt:

- Das Gerät sucht einen Registrar
- Das Gerät sendet seine Identitäten (ProductInstanceURIs) an den Registrar
- Der Registrar wählt eine Identität aus, die zum weiteren provisionieren verwendet werden soll
 - o Zu der er ein gültiges Ticket hat
 - o Das Ticket von einer ihm ausgewählten Autorität ausgestellt ist. Zum Beispiel das Ticket eines Maschinenherstellers
- Der Registrar sendet die ausgewählte Identität an das Gerät zurück
- Bevor die Herstellung des Vertrauensverhältnisses zwischen Gerät und Registrar beginnt, muss der Softwarestatus des Gerätes über den Softwareupdate Manager überprüft werden.
- Der korrekte Softwarestatus wird vom Softwareupdate Manager an den Registrar übertragen. Ohne diesen lehnt der Registrar das weitere provisionieren ab.
- Um das Vertrauensverhältnis herzustellen, baut das Gerät einen neuen UA Secure Channel unter Verwendung des ausgewählten Zertifikats, nebst zugehörigem Private Key auf
- Das Gerät akzeptiert jedes gültige Zertifikat des Registrars – OPC UA Device Provisioning führt keine mutual authentication durch.
- Wird der UA Secure Channel erfolgreich aufgebaut, ist damit bewiesen, dass das Gerät über den private Key des Geräte Zertifikats verfügt. Damit hat das Gerät seine Authentizität bewiesen.
- Das Gerät fordert in diesem Kanal das DCA (Device Configuration Application) Zertifikat (StartSigningRequest) und die Trust List (GetTrustList) an.
- Der Kanal wird geschlossen und für weitere Kommunikation baut das Gerät neue UA Secure Channels unter Benutzung des DCA- oder weiterer Application Instance Zertifikate

an. Da das IDevID Zertifikat unendliche Gültigkeit hat, ist die Nutzung des Schlüsselmaterial des Zertifikats zu minimieren.

- Weitere Zertifikate (Application Instance Zertifikate) können über neue UA Secure Channel Verbindungen angefordert werden, die das DCA Zertifikat verwenden.
- Die Entscheidung ob Zertifikate ausgestellt werden, liegt bei der Registration Authority.
- Weitere Manager Instanzen können verwendet werden, zum Beispiel um weitere Überprüfungen vorzunehmen. Diese können entweder zwingend oder optional sein, je nach Konfiguration.

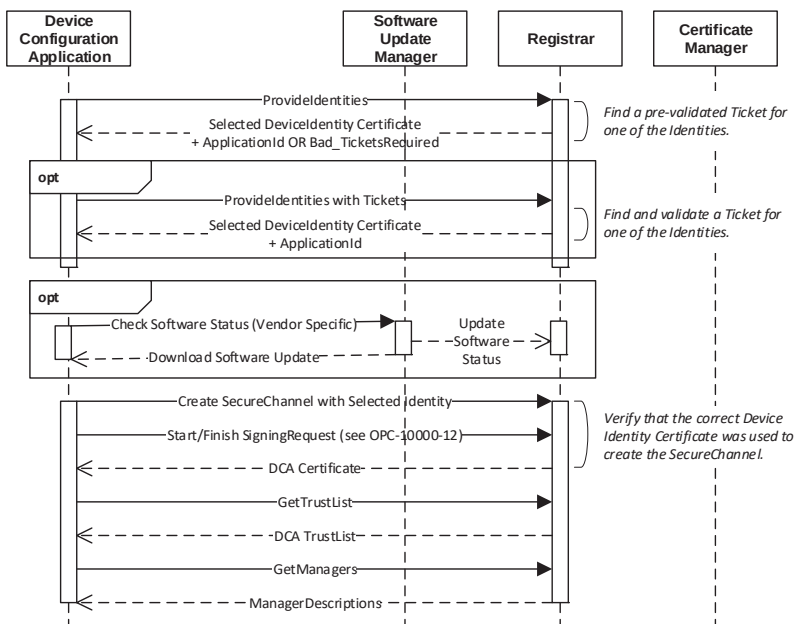


Bild 4: Provisionierungs Prozess nach Pull Model

2.3 Eigenschaften der OPC UA Geräte Provisionierung

ToFU (Trust on First Use)

Der Provisionierungs-Prozess des Gerätes kann nur ein einziges Mal durchgeführt werden. Einen weiteren Provisionierungs-Versuch würde das Gerät ablehnen – somit kann der Registrar erkennen, ob das Gerät bereits provisioniert wurde. Soll ein gebrauchtes Gerät erneut provisioniert werden, so muss es erst in einen definierten Zustand zurückgesetzt werden.

Herstellen der Vertrauensbeziehung durch Etablierung eines UA Secure Channels

Bei der ersten Verbindung, die beim Provisionierungsprozess zwischen dem Gerät und dem Registrar etabliert wird, handelt es sich um eine nicht sichere Verbindung. Sie dient nur zur Auswahl des Zertifikats, mit dem die Vertrauensbeziehung hergestellt werden soll.

Nach Auswahl des Zertifikats, stellt das Gerät einen UA Secure Channel zum Registrar mit diesem Zertifikat und dem dazugehörigen private Key her. Kommt diese Verbindung zustande, hat das Gerät bewiesen, dass es über den private Key verfügt und somit das im Zertifikat beschriebene Gerät ist.

Client – Pull, Server – Push

OPC UA Clients bauen Verbindungen zu OPC UA Servern auf und OPC UA Server warten auf solche Verbindungsanforderungen. Der Registrar kann beliebige OPC UA Applikationen provisionieren – daher muss er sich in dem einen Fall als OPC UA Client und im anderen Fall als OPC UA Server verhalten:

	Registrar: OPC UA Server	Registrar: OPC UA Client
Gerät: OPC UA Client	Pull	--
Gerät: OPC UA Server	--	Push

Bild 5: Client – Pull, Server – Push

- Pull – Modell
 - o Hier ist das Gerät als Client der aktive Part, der die Verbindung zum Registrar aufbaut.
- Push – Modell
 - o Hier ist das Geräte als Server der passive Part, der darauf wartet vom Registrar angesprochen zu werden.

Auf der funktionalen Ebene unterscheiden sich Push und Pull Modell nicht, daher wird in diesem Dokument nur das Pull Modell beschrieben.

3. Aufgaben und Rollen

Die einzelnen Akteure benötigen eine spezielle Infrastruktur, um die Geräte Provisionierung zu realisieren.

3.1 Hersteller

Der Hersteller besitzt eine vertrauenswürdige CA (Manufacturer CA), mit deren Hilfe bei der Produktion eines neuen Gerätes ein IDevID Zertifikat nach dem IEEE 802.1AR Standard, sowie ein Ticket nach OPC UA Part 21 erstellt wird.

Das Zertifikat sowie das Ticket, enthalten eine eindeutige ProductInstanceURI, z.B. nach dem Standard DIN 91406. Sollte die ProductInstanceURI als URL dargestellt werden, so darf diese auf keine reelle Ressource verweisen, da dies auf schädlichen Informationen verweisen könnte.

Beispiele für mögliche ProductInstanceURI:

```
urn:some-company.com:5ff40f78-9210-494f-8206-c2c082f0609c
http://some-company.com/snr-16273849
urn:some-company.com:model-xyz:snr-16273849
```

Das IDevID Zertifikat wird auf das Gerät selbst gebrannt und zudem wird auf die Verpackung die passende ProductInstanceURI aufgedruckt. Das IDevID Zertifikat enthält zusätzliche beschreibende Informationen zu dem Gerät.

Field	Description accord. to IEEE 802.1AR
Version	3
Serial Number	[positive integer of up to 20 octets] unique in scope of Issuing CA
Signature Algorithm	RSA-2048/SHA-256, ECDSA P-256/SHA-256 or ECDSA P-384/SHA-384
Issuer	Subject Distinguished Name of Issuing CA
Validity	indefinitely into the future (should use the GeneralizedTime value 99991231235959Z)

Subject	Distinguished Name serialNumber = [Product Serial Number] (unique within issuer's domain of significance)
Subject Public Key Info	[Algorithm to Public Key], [Public Key]
Authority Key Identifier	[Key Identifier of Issuing CA]
Subject Key Identifier	not recommended / shall not be used
Key Usage (critical)	digitalSignature (shall) keyEncipherment – optional marked as critical
Subject Alt Name	(optional) if used should contain: [HardwareModuleName] as specified in RFC 4108
Basic Constraints	„This extension MAY appear as a critical or non-critical extension in end entity certificates.“ RFC 5280

Bild 6: Auszug IDeVID Zertifikat

Das Ticket wird von der vertrauenswürdigen CA signiert.

Field	Description
DeviceIdentityTicketType	Subtype of the BaseTicketType DataType defined in 4.5.6.
Uri = ProductInstanceUri	The ProductInstanceUri for the Device.
manufacturerName	The name of the Manufacturer for the Device.
modelName	The model name assigned by the Manufacturer. Not specified if no model name was assigned or known.
modelVersion	The model version assigned by the Manufacturer. Not specified if no model version was assigned or known.
hardwareRevision	The hardware revision assigned by the Manufacturer. Not specified if no hardware revision was assigned or known.
serialNumber	The serial number assigned by the Manufacturer. Not specified if no serial number was assigned or known.
manufactureDate	When the Device was manufactured. DateTime.MinValue if the date is not known.
authorities	A list of CAs need to validate DeviceIdentity Certificates on the Device that were installed when the Ticket was created.

Bild 7: Auszug Geräte Ticket

3.2 Distributor

Ein Distributor kann ebenfalls als Autorität agieren und eigene Tickets für das Gerät - mit derselben ProductInstanceURI - erzeugen. Im Provisionierungs Prozess kann dann der Akteur, z.B. der Endkunde, entscheiden anhand welchen Tickets er die Provisionierung durchführt. Die Auswahl des Tickets spiegelt das ausgewählte Vertrauensverhältnisses wider.

3.3. Maschinenhersteller

Ein Maschinenhersteller, welcher als Autorität agiert, hat dieselben Möglichkeiten wie der Distributor ein neues Ticket für ein Gerät auszustellen. Zusätzlich kann er mehrere Geräte zu einem komplexen Gebilde zusammenführen.

Er kann hierzu neue Zertifikate oder Tickets erzeugen. Diese enthalten als Identifikation eine CompositeInstanceURI, sowie alle beteiligten ProductInstanceURIs. Analog zu Geräte Zertifikaten oder Geräte Tickets können diese ebenfalls weitere Informationen wie Maschinenhersteller etc. enthalten.

3.4 Integrator / Operator

Der Integrator/Operator nimmt das Gerät oder die Maschine in Betrieb. Hierzu soll die Herkunft des Geräts/Maschine sichergestellt werden (Provisionierung). Er stellt hierzu einen Registrar als Kommunikationspartner zur Verfügung.

Der Registrar stellt die Verbindung zwischen dem Gerät/Maschine und der Sicherheitsinfrastruktur her.

4. Mögliche Integration in PKI (PoC)

Der Registrar stellt die Verbindung zur Public Key Infrastruktur (PKI) her, diese wird i.d.R. durch eine RA/CA realisiert – optimalerweise kombiniert mit einem Geräteinventar.

Die Interaktion mit dieser erfolgt zwischen dem StartSigningRequest und dem Ausliefern des DCA Zertifikates. In dem PoC wurde von dem Registrar mit der PKI über das Certificate Management Protocol (CMP, RFC 4210) und dem Lightweight CMP Profile [2] kommuniziert. Bei den verwendeten Certification Authorities handelte es sich um Standardprodukte.

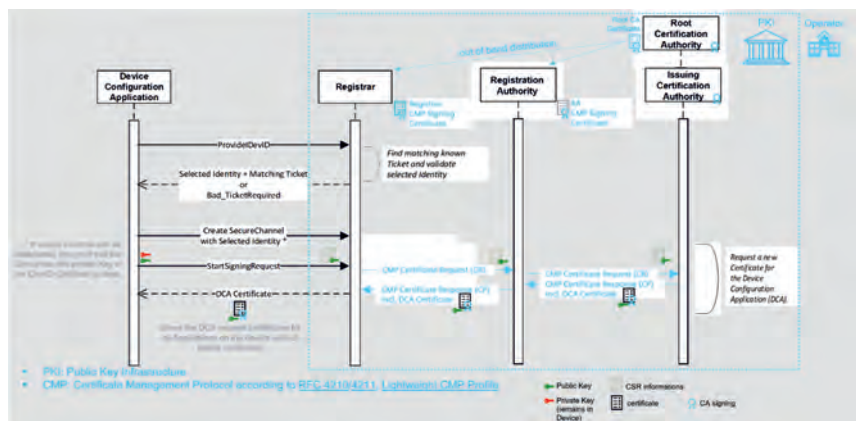


Bild 8: PoC - Kommunikation zu PKI

Während die Root CA als Vertrauensanker integritätsgeschützt und sicher abgekapselt von der produktiven Umgebung ist, können Zertifikate von einer Issuing CA ausgestellt werden. Eine

anlagenspezifische Registration Authority (RA) übernimmt die Kommunikation zwischen dem Registrar und den in der Anlage vorhandenen CAs. Somit übernimmt die RA neben der Authentifizierung und Autorisierung der Zertifikatsanträge auch kontextspezifisches Routing der Zertifikatsanträge zu den CAs. Als Anlagenkomponente ist der Registrar aus Sicht der RA eine Local Registration Authority (LRA), weshalb Vertrauensbeziehungen vor der Provisionierung des Gerätes konfiguriert sind.

Da die DCA selbst das Kommunikationsprotokoll der Anlage nicht bedient, fungiert der Registrar als ein Gateway. Der Inhalt des (OPC UA) StartSigningRequests der DCA wird in eine CMPNachricht konvertiert. Bei der Implementierung wurde der open-source basierte "generic CMP Client" <https://github.com/siemens/genccmpclient> verwendet.

Während somit eine Vielzahl von Anwendungsfällen direkt abdeckbar ist (Zertifikatsanfrage, -erneuerung und -revokation), beschränkt sich die Nutzung im Kontext des PoCs auf die Zertifikatsanfrage.

Da der Registrar bereits als LRA in die CMP-basierte PKI integriert ist, somit schon anlagen spezifische Zertifikate für die Kommunikation nutzt, ist kein "Initial Request" (IR) mehr notwendig. Die RA nimmt den vom Registrar aus dem OPC UA StartSigningRequest erstellten CR entgegen und legt nach der erfolgreich durchgeführten Prüfung der Vertrauensbeziehung ein Gerät in ihrem Inventar an, um Administratoren transparent den automatischen Prozess nachverfolgen lassen zu können. Basierend auf dem vom Registrar angeforderten Zertifikatsprofil (in diesem Fall ein LDevIDGeneric-Zertifikat), wird die Nachricht bzw. der Zertifikatsantrag in Form eines Certificate Request (CR) an die entsprechende CA weitergeleitet. Diese stellt das Zertifikat entsprechend den Angaben im Zertifikatsantrag und dem entsprechenden Zertifikatprofil aus und schickt an die RA eine Rückmeldung in Form eines Certificate Response (CP), welcher das ausgestellte Zertifikat enthält, zurück an die RA. Diese wiederum leitet die Antwort nach internen Konformitätsprüfungen an den Registrar weiter.

5. Zusammenfassung

OPC UA Part 21 Device Provisioning:

- definiert eine Schnittstelle für die sichere Geräte Integration auf Basis von OPC UA
- kann mit anderen Device Provisioning Verfahren kombiniert werden
- kann als Zugang zu einer PKI genutzt werden (PoC)
- bietet Möglichkeiten zur automatisierten Installation und Softwarebereitstellung

Literaturverzeichnis

- [1] OPC Foundation, „OPC UA Part 10000-21: Device Provisioning (Draft 50),“ 2022.
- [2] H. Brockhaus, D. von Oheimb und S. Fries, „Lightweight Certificate Management Protocol (CMP) Profile,“ 2022.

FAIRness in der Automatisierungstechnik am Beispiel Modellierung und Simulation

M. Stüber, Prof. Dr.-Ing. **G. Frey**,
Universität des Saarlandes, Saarbrücken

Kurzfassung

Modellierung und Simulation sind ein wichtiges Werkzeug in Ingenieurwissenschaften und Industrie. Die Erstellung von Modellen ist oft sehr aufwendig und damit teuer. Deshalb ist eine Wiederverwendung einmal erstellter Modelle anzustreben. Im Kontext von verteilten Entwicklungsprozessen und bei der Integration von Modellen aus verschiedenen Quellen, beispielsweise bei Industrie 4.0, ist zudem die Interoperabilität von Modellen zu sichern. Im Bereich der Forschung sind durch die Diskussion um die Nachvollziehbarkeit von Forschungsergebnissen Konzepte zur FAIRen Veröffentlichung von Daten entwickelt worden (FAIR = Findable–Accessible–Interoperable–Reusable). Zu der oben für Modelle bereits geforderten Interoperabilität und Wiederverwendbarkeit treten hier noch Auffindbarkeit und Zugänglichkeit. Aufgrund der Tatsache, dass Modelle letztlich abstrakte Daten sind, untersucht dieser Beitrag wie FAIRness deren Nutzung verbessern kann. Hierzu werden exemplarisch Modelle nach dem offenen Standard FMI (Functional Mock-up Interface) über eine Ontologie mit Metadaten angereichert und durch Semantic Web Technologien als Services zur Verfügung gestellt. Der Ansatz erlaubt auch die Nutzung der Services durch intelligente Softwareagenten.

Stichwörter: FAIR-Prinzipien; Simulation; Modellaustausch; FMI/FMU; Semantic Web

1. Einleitung

Die fünfzehn FAIR-Prinzipien [13] sollen dazu beitragen, die Wiederverwendbarkeit und damit den Nutzen von Forschungsdaten zu maximieren. Sie adressieren organisatorische und technische Aspekte sowie die Frage, welche Metadaten mit den Daten veröffentlicht werden sollten. Organisatorische Aspekte betreffen die langfristige Verantwortungsübernahme für einen veröffentlichten Datensatz, wie beispielsweise die Nutzung persistenter IDs und die andauernde Verfügbarkeit von Metadaten selbst dann, wenn die Daten nicht mehr zur Verfügung stehen.

Die FAIR-Prinzipien sind technologieoffen formuliert und explizit sowohl auf Menschen als auch auf Software-Agenten ausgerichtet.

Wichtig ist bei den FAIR-Prinzipien, dass zugänglich nicht „ohne Einschränkungen und kostenfrei“ bedeutet: vielmehr geht es darum sicherzustellen, dass potenzielle Nutzende von der Existenz und dem Inhalt digitaler Objekte oder Sammlungen Kenntnis erlangen. Der Inhalt kann aus verschiedenen Gründen vor allgemeinem Zugriff geschützt werden; etwa auch, weil der Inhalt nur gegen Bezahlung oder unter bestimmten Bedingungen Dritten zur Verfügung gestellt werden soll [8]. Das bedeutet, dass die FAIR-Prinzipien genauso relevant für Forschung wie für gewinnorientierte Unternehmen sein können.

Forschungsdaten FAIR zu machen, wird allgemein als erstrebenswert angesehen—beispielsweise haben sich große Forschungsnetzwerke dazu verpflichtet, Forschende dabei zu unterstützen, ihre Daten in FAIRer Art und Weise zu veröffentlichen [14]. In Deutschland hat sich die NFDI4ing-Initiative¹ gegründet, um Anforderungen, Methoden und Services für das FAIRe Forschungsdatenmanagement in den Ingenieurwissenschaften zu sammeln, zu entwickeln, zu standardisieren und zu kommunizieren. NFDI4ing gliedert ihre Arbeit anhand von sieben Archetypen, die typische Arbeitsweisen abbilden, sowie ergänzenden Anwendungsfeldern und Basisservices.

Vor allem Archetyp „Betty“² bildet dabei die Problemfelder ab, in die auch diese Arbeit einzuordnen ist: Programmierung und Validierung von Forschungssoftware sowie Kontext für und Reproduzierbarkeit von mit dieser Software erstellten Ergebnissen. Die Arbeitsgruppe „Betty“ verfolgt unter anderem eine integrierte Toolchain, Best-Practice-Sammlungen, Containerisierung, standardisierte Metadaten und einen Software-/Datenkatalog als Lösungsansätze für die genannten Problemstellungen. Modellierung und Simulation werden dabei als Spezialfälle von Forschungssoftware gesehen.

¹ <https://nfdi4ing.de/>

² <https://nfdi4ing.de/archetypes/betty/>

Modelle und Simulationen des dynamischen Verhaltens von Systemen sind ein essenzielles Werkzeug, um die Komplexität der Systeme in Entwicklung und Betrieb beherrschen zu können. In der Regel werden Modelle lokal unter Verwendung von spezieller Software entwickelt und durch Simulation angewendet. Es gibt aber Szenarien, in welchen es vorteilhaft ist, Simulationen auf einem verteilten Satz von Rechnern auszuführen, der abhängig von der Nachfrage erweitert oder verkleinert werden kann. Beispielsweise kann dies sinnvoll sein, wenn hunderte Simulationen als Ganzes ausgewertet müssen (Parameterfitting, Optimierungen) oder die Simulationen Teil eines überlagerten verteilten Prozesses sind; etwa für den Betrieb von Smart Grids oder die kundenspezifische Zusammenstellung von Anlagen aus Komponenten durch Mitarbeitende im Vertrieb, beim Kunden vor Ort.

Der Begriff „Cloud Computing“ beschreibt wünschenswerte Charakteristiken und typische Ausprägungen für den Zugriff auf IT-Ressourcen über das Internet [6]. Aus Nutzendensicht ist essenziell, dass Software oder Ressourcen (Rechner, CPUs, Speicher, ...) *als Service* über das Internet und in der Regel in unlimitierter Menge/Leistung verfügbar sind. Die Nutzung wird gemessen und entsprechend in Rechnung gestellt.

Modellierung und Simulation (M&S) als Service bereitzustellen, erfordert interdisziplinäre Arbeit an den Schnittstellen von Ingenieurwissenschaften, Informatik und Software-Engineering. Kernhypothese von Modellen und Simulation als Service (MSaaS) ist, dass Zugänglichkeit und Wiederverwendung verbessert werden können, indem Modelle und Simulationsfähigkeiten über das Internet für ein breiteres Publikum erreichbar werden. Außerdem kann die Funktionalität durch Komposition verschiedener Services und die Leistung durch den Betrieb in der Cloud verbessert werden (siehe [11] für eine detailliertere Darstellung und weitere Referenzen).

Die einfache Erreichbarkeit von Modellen und Simulationsumgebungen über das Internet ist notwendig, aber nicht ausreichend, um die erwünschten Vorteile vollumfänglich zu realisieren. Dafür sollten Modelle und Simulationsfähigkeiten die FAIR-Prinzipien erfüllen.

2. Forschungsstand

Zum Thema Modellierung und Simulation als Service gibt es umfangreiche Arbeiten. Als Einstieg empfehlen sich die Artikel [1] und [9]; der Stand der Forschung hinsichtlich verwendeter Softwarearchitekturen, notwendiger Funktionalitäten für den Aufbau großskaliger MSaaS-Anwendungen und der Nutzung von Cloud Computing-Konzepten wird in den Übersichtsartikeln [10], [3] und [5] dargestellt.

Fokussiert man sich auf die Arbeiten in der Modelica-Community, können diese in drei Kategorien eingeteilt werden [12]: Arbeiten zur Bereitstellung von Webapplikationen für ExpertInnen; Arbeiten zur verteilten Co-Simulation; sowie Arbeiten, die sich mit der Frage auseinandersetzen, wie die Designprinzipien des Web möglichst sinnvoll für M&S-Werkzeuge eingesetzt werden können (in diese Kategorie fällt die vorliegende Arbeit).

Cloud-basierte Software-Services werden typischerweise als Komposition unterschiedlicher Basisservices realisiert, die jeweils für nur eine Aufgabe zuständig sind (Service-orientierte Softwarearchitektur, Mikroservices). Aus technischer Sicht sind Services üblicherweise als HTTP-APIs realisiert, deren Interface sich am Architekturstil des Web (Representational State Transfer (REST)) orientiert.

Aus praktischer Sicht kann REST wie folgt sehr grob zusammengefasst werden: ein Service veröffentlicht Ressourcen, die ein Konzept der Anwendungsdomäne darstellen und durch eine URL identifiziert und lokalisiert werden. Als Reaktion auf die Anwendung eines HTTP-Verbes (GET, POST, ...) wird eine *Repräsentation* der Ressource übermittelt. Für Systeme, die den Architekturstil REST umsetzen, ist wesentlich, dass die Interaktion zwischen NutzerIn und Service durch die nutzerseitige Auswahl von in den Repräsentationen enthaltenen Optionen (Links, Eingabemasken) erfolgt und *nicht*, in dem ein Benutzerhandbuch gelesen werden muss, um valide Serviceanfragen konstruieren zu können. Dieses Prinzip wird „Hypermedia As The Engine of Application State (HATEOAS)“ genannt und von Menschen zur Verwendung von Webseiten selbstverständlich benutzt. Allerdings ist es noch die Ausnahme, dass Software-Agenten APIs auf dieselbe Art und Weise nutzen.

APIs, die den Architekturstil REST vollumfänglich umsetzen, werden als „Hypermedia-APIs“ bezeichnet. Unsere erste Hypothese ist, dass Hypermedia-APIs ein wesentlicher Schritt sind, um die FAIRness von Modelle und Simulationen zu erhöhen.

Die FAIR-Prinzipien fordern die Verwendung einer zugänglichen, frei verwendbaren, universell einsetzbaren und *formalen* Sprache für die Repräsentation von Wissen. Es sind also ein Datenmodell und entsprechende Serialisierungen nötig, die die Kodierung von Links und Wissen nativ unterstützen. Dieses Datenmodell und die notwendigen Serialisierungen wurden im Kontext des Semantic Web entwickelt. Das Semantic Web ist eine Erweiterung des Web, deren Kernidee es ist, die Bedeutung von Entitäten explizit zu machen, um Inhalte im Web für generische Softwareagenten zugänglich zu machen [2].

Die Bedeutung von Inhalten wird im Semantic Web unter Verwendung des „Resource Description Framework (RDF)“-Datenmodells kodiert, welches Daten als gerichteten Graphen darstellt: Subjekt-Knoten sind mit Objekt-Knoten durch ein Prädikat verbunden. Der Graph entsteht durch die Sammlung von Subjekt—Prädikat—Objekt-*Tripeln*. Subjekte und Prädikate sind URLs (somit sind Links nativ darstellbar) und auch generelle Wissensbeziehungen wie etwa „Jedes valide Modell, das in der Sprache Modelica implementiert wurde, kann in der Softwareumgebung OpenModelica simuliert werden“ können abgebildet werden. Daher ist es naheliegend (und unsere zweite Hypothese), die Konzepte und Technologien des Semantic Web für die Realisierung FAIRer digitaler Objekte im Allgemeinen und FAIRer Modelle und Simulationen im Speziellen zu verwenden.

Arbeiten zur Kombination von Ideen und Technologien des Semantic Web mit Modellierung und Simulation fokussieren hauptsächlich die Verwendung von Ontologien für die Verbesserung des Model-Based Systems Engineering (MBSE). Themen sind die bessere Integration von Wissen und Informationen unterschiedlicher Beteiligter; die automatische Betrachtung der Interoperabilität von Modellen; und die Generierung von Simulationsmodellen aus Systembeschreibungen auf Basis von Ontologien. Nur wenige AutorInnen haben die Verwendung von Hypermedia-APIs für die Bereitstellung von M&S-Funktionalität untersucht. Software wurde nicht offengelegt; eine Betrachtung der FAIRness ist uns nicht bekannt.

Aus unserer Sicht limitiert die fehlende FAIRness von Modellen und Simulationen den positiven Nutzen und die Wirkung, die diese entfalten könnten. Deshalb stellt diese Arbeit folgende Forschungsfrage: Kann die FAIRness von Modellen, Simulationsfähigkeiten und Ergebnissen verbessert werden, in dem diese durch eine Hypermedia-API zur Verfügung gestellt werden, welche RDF-Repräsentationen für die Ressourcen nutzt?

3. Technische Umsetzung

Die Beantwortung der Forschungsfrage erfordert die Implementierung von zwei Kernfunktionalitäten. Erstens muss eine API implementiert werden, die Modelle, Simulationsfähigkeiten und -Ergebnisse über das Internet verfügbar macht und dabei den Architekturstil des Web einschließlich des HATEOAS-Prinzipes realisiert. Zweitens müssen die RDF-Repräsentationen der API-Ressourcen ausgehend vom unterstützten Modellformat soweit wie möglich automatisch generiert werden.

Unterstützt werden Modelle, die dem offenen Standard Functional Mock-up Interface (FMI) [7] entsprechend abgespeichert wurden. Diese werden als Functional Mock-up Units (FMUs) bezeichnet. FMI ist weit verbreitet und wird von mehr als 150 Tools zumindest teilweise unterstützt³. Es gibt zwei Varianten von FMUs: FMUs für den Modellaustausch enthalten nur die Modellgleichungen und benötigen einen externen Solver für die Simulation. FMUs für Co-Simulation enthalten dagegen einen Solver und können deswegen sowohl als alleinstehende, ausführbare Form eines Modells; als auch im Rahmen einer Co-Simulation mit anderen FMUs eingesetzt werden. In dieser Arbeit wird Co-Simulation nicht betrachtet.

Um relevante M&S-Konzepte wie FMUs im Semantic Web (und damit im Datenmodell RDF) beschreiben zu können, bedarf es Ontologien. Ontologien sind in sich konsistente, formale Spezifikationen von Konzepten und deren Beziehungen in einer Anwendungsdomäne [4].

³ <https://fmi-standard.org/tools/>

Für die vorliegende Arbeit wurden drei Ontologien entwickelt: die FMI-Ontologie erlaubt die Beschreibung von FMUs in RDF. Die Systems, Models, Simulations (SMS)-Ontologie ermöglicht, (Teile von) Systemen zu (Teilen von) Modellen in Bezug zu setzen. Die SMS-FMI-Ontologie formuliert schließlich die Beziehung zwischen den Konzepten der beiden erstgenannten Ontologien. Außerdem wurde ein Parser entwickelt, der die RDF-Beschreibung einer FMU soweit möglich automatisiert erstellt. Mit den Ontologien und dem Parser können die Repräsentationen der durch die Hypermedia-API angebotenen Ressourcen erstellt werden.

Beim Entwurf der Hypermedia-API muss die bereitzustellende Funktionalität (Hinzufügen, Parametrieren, Simulieren von Modellen) in Ressourcen und darauf anwendbare HTTP-Verben übersetzt werden. Zudem müssen die Repräsentationen so ausgestaltet werden, dass sie neben den eigentlichen Daten auch Triples enthalten, die zur FAIRness beitragen und solche, die das HATEOAS-Prinzip ermöglichen.

In Tabelle 1 sind die wesentlichen Elemente des Serviceinterfaces zusammengefasst. Grundlegende Ressource des Serviceinterfaces ist ein Modell. Modelle können instanziiert werden, indem die Modellparameter definiert werden. Mit den resultierenden Modellinstanzen können Simulationen spezifiziert werden, indem Simulationseinstellungen, Start- und Randbedingungen/Eingangsdaten festgelegt werden. Die eigentliche Simulation wird im Hintergrund gestartet, wenn durch einen POST-Request eine Simulationsressource angelegt wird. Sobald das Ergebnis bekannt ist, wird es in der Repräsentation der Simulation verlinkt und kann über diesen Link abgerufen werden.

Tabelle 1: Serviceinterface der Hypermedia-API (Auszug)

HTTP-Methode	Ressource	Beschreibung
POST	/models	Modell hinzufügen
GET	/models/<id>	Modellrepräsentation erhalten
POST	/models/<id>/instances	Modell durch Parametrierung instanziiieren
GET	/models/<id>/instances/<id>	Repräsentation einer Modellinstanz erhalten
POST	/models/<id>/instances/<id>/experiments	Eine Simulation durch Festlegen aller Einstellungen starten
GET	/models/<id>/instances/<id>/experiments/result	Simulationsergebnis abrufen

Die Implementierung dieser Funktionalität ist in vier Teile gegliedert (Bild 1). Die API-Komponente stellt das Serviceinterface zur Verfügung, leitet Anfragen an die eigentliche Implementierung der Funktionalität weiter und erstellt die Repräsentationen der Ressourcen. Die Worker-Komponenten erledigen alle Aufgaben, die spezifisch für ein bestimmtes Modellformat sind, etwa die Generierung der RDF-Repräsentation von FMUs oder die eigentliche Simulation. Verbunden sind API und Worker-Komponenten durch eine Warteschlange, die aus zwei Teilen (Aufgabentransfer zu den Worker-Komponenten; Rücktransfer der Ergebnisse) besteht. Es ist vorgesehen, dass die einzelnen Komponenten in sogenannten Containern verwendet werden. Diese bündeln die Anwendung und alle Abhängigkeiten und bilden den grundlegenden Baustein für den Betrieb auf Softwareplattformen wie Kubernetes, welche die horizontale Skalierung in Abhängigkeit der Last regeln.

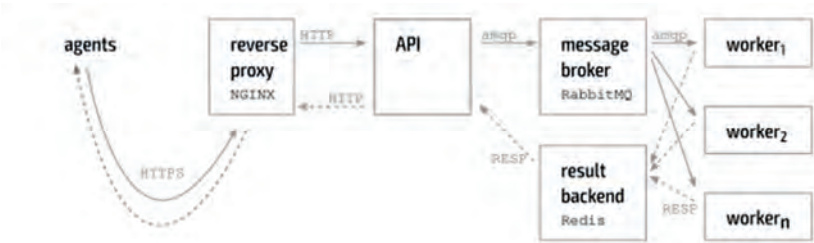


Bild 1: Softwarearchitektur der Hypermedia-API

Sämtliche entwickelten Software-Komponenten einschließlich der Ontologien wurden unter einer offenen Lizenz zur Verfügung gestellt (siehe Tabelle 2; Details können den Software-Repositorien entnommen werden).

Tabelle 2: Softwarekomponenten und -Repositorien

FMI-Ontologie, Parser	https://github.com/UdSAES/fmi2rdf
SMS-Ontologie, SMS-FMI-Ontologie	https://github.com/UdSAES/sms-ontology
API-Implementierung	https://github.com/UdSAES/simaas-api
Worker-Implementierung	https://github.com/UdSAES/simaas-worker

4. Ergebnisse

Die wesentlichen, durch die implementierte Software neu erlangten Fähigkeiten, sind erstens die Simulation von als FMU vorliegenden Modellen in der Cloud und zweitens die Repräsentation von Ressourcen im Datenmodell RDF.

4.1. Anwendungen und Softwareeigenschaften

Die Vorteile der Simulation in der Cloud wurden an zwei Beispielen gezeigt (siehe [12] für Details): eine Ensembleprognose der Leistungserzeugung einer PV-Anlage repräsentiert die vielfache Simulation derselben Modellinstanz mit unterschiedlichen Randbedingungen. Die Bestimmung optimaler Parameterwerte für eine elektrische Schaltung mit Hilfe eines genetischen Algorithmus repräsentiert die Simulation hunderter unterschiedlicher Modellinstanzen mit denselben Randbedingungen. Für beide Anwendungen müssen viele Simulationen als Ganzes ausgewertet werden, die aber voneinander unabhängig sind. Deswegen profitieren beide Anwendungen von der parallelen Ausführung in einem horizontal skalierbaren Pool von CPUs in der Cloud. Außerdem wird die Einbettung in überlagerte (verteilte) Prozesse stark vereinfacht, da Nutzende keine Software installieren müssen. Durch die Trennung von API und Worker-Komponenten durch eine Warteschlange wird erstens sichergestellt, dass auch viele gleichzeitig eintreffende Anfragen robust abgearbeitet werden und zweitens ermöglicht, für API und Worker-Komponenten unterschiedliche Programmiersprachen zu verwenden und so den unterschiedlichen Anforderungen gerecht zu werden.

Die Wahl von RDF für die Repräsentation von Ressourcen legt die Bedeutung aller Entitäten eindeutig fest: alle Bestandteile eines Triples sind in Ontologien definiert und können dort nachgeschlagen werden. Zusätzlich zu den eigentlichen Daten kodieren weitere Triples Metadaten wie beispielsweise Provenienz-Informationen oder beschreiben, welche Aktionen ausgehend von dieser Ressource ausgewählt werden können. Somit entsteht insgesamt eine sich selbst beschreibende API, die ohne eine externe Dokumentation auch von Softwareagenten genutzt werden kann. Abschließend ermöglicht die Verwendung von Ontologien auch das Ableiten von nicht explizit in den Repräsentationen enthaltenen Fakten durch logisches Schlussfolgern.

4.2. FAIRness

Die FAIRness der durch eine Instanz der Hypermedia-API bereitgestellten M&S-Fähigkeiten wurde im Vergleich zu einzelnen FMUs evaluiert; die Ergebnisse sind in Bild 2 visualisiert.

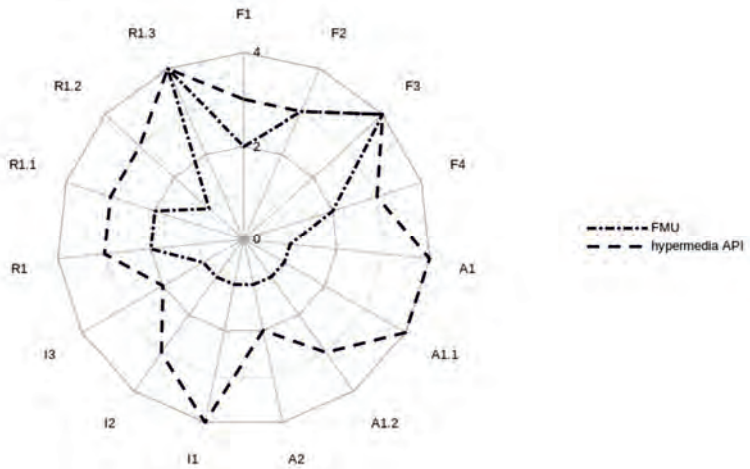


Bild 2: Unterstützung der FAIR-Prinzipien durch Hypermedia-API und FMUs

Für jedes der 15 FAIR-Prinzipien wurden beide Varianten als ‚nicht unterstützt‘ (numerischer Wert 1), ‚unterstützt‘ (2), ‚teilweise umgesetzt‘ (3) oder ‚vollständig umgesetzt‘ (4) kategorisiert. Nicht unterstützt bedeutet, dass die gewählte Architektur der Umsetzung des Prinzips widerspricht; in den Kategorien 2–4 kann das Prinzip mit der gewählten Architektur vollständig umgesetzt werden.

Es gibt keinen Unterschied in der FAIRness hinsichtlich den Prinzipien F2 (Beschreibung der Daten durch Metadaten), F3 (explizite Verlinkung von Metadaten zu Daten) und R1.3 (Metadaten entsprechen Community Standards). Alle anderen Prinzipien sind durch die Hypermedia-API mindestens unterstützt und in der Regel teilweise umgesetzt, während einzelne FMUs sieben der 15 FAIR-Prinzipien nicht unterstützen. Durch umfangreichere Metadaten in den Repräsentationen, insbesondere hinsichtlich Lizenzen und Provenienz, sowie Verlinkungen zu anderen relevanten Daten, könnte die FAIRness weiter gesteigert werden.

5. Ausblick

Zusätzlich zur wesentlich verbesserten FAIRness der M&S-Ressourcen ermöglicht das gewählte Konzept die Implementierung weiterer wünschenswerter Funktionalitäten. Die Verwendung des RDF-Datenmodells und die Erweiterung der Repräsentationen von Ressourcen um auf den ersten Blick nicht-essentielle Informationen kann bessere Nachvollziehbarkeit und damit höheres Vertrauen in die Ergebnisse ermöglichen. Auch das Rückverfolgen von Änderungen während des MBSE-Prozesses wird dadurch erleichtert, dass einmal erstellte Ressourcen nicht geändert werden können. Die Cloud-native Softwarearchitektur vereinfacht den Betrieb auf modernen Cloudinfrastrukturen inklusive der Möglichkeit einer horizontalen Skalierung. Weitere Modellformate könnten ohne Änderung der API unterstützt werden, indem neue Worker-Implementierungen ergänzt werden.

Außerdem bildet die Hypermedia-API einen essenziellen Grundbaustein für komplexere Anwendungsfälle wie die Nutzung durch generische (nicht für diese API programmierte) Softwareagenten sowie die Modellerstellung aus Teilkomponenten basierend auf einer Systembeschreibung in RDF.

6. Zusammenfassung

Dieser Bericht gibt einen groben Überblick über einen möglichen Weg, um die fehlende FAIRness von Modellen, Simulationsfähigkeiten und Simulationsergebnissen zu adressieren. Das Thema FAIRness ist in Bezug auf Forschungsdatenmanagement Gegenstand vielfältiger Arbeiten; aus unserer Sicht sollten Modelle und Simulationen dabei unbedingt mitgedacht werden.

Modelle und Simulationen wurden über eine Hypermedia-API bereitgestellt und in Beispielanwendungen verwendet. Die implementierte Software demonstriert Funktionalität und wünschenswerte Eigenschaften wie horizontale Skalierbarkeit in der Cloud. Die FAIRness konnte im Vergleich zu einer Sammlung von FMUs wesentlich gesteigert werden. Alle Softwarekomponenten wurden unter einer offenen Lizenz veröffentlicht.

7. Literaturangaben

- [1] Cayirci, E. (2013): Modeling and simulation as a cloud service: A survey. 2013 Winter Simulations Conference (WSC). IEEE.
- [2] Glimm, B., Stuckenschmidt, H. (2016): 15 Years of Semantic Web: An Incomplete Survey. *KI – Künstl. Intell.* Vol. 30, Nr. 2, S. 117–130.
- [3] Hannay, J.E., Berg, T. van den, Gallant, S., Gupton, K. (2020): Modeling and Simulation as a Service infrastructure capabilities for discovery, composition and execution of simulation services. *J. Def. Model. Simul.*
- [4] Hofmann, M., Palii, J., Mihelcic, G. (2011): Epistemic and normative aspects of ontologies in modelling and simulation. *J. Simul.* Vol. 5, Nr. 3, S. 135–146.
- [5] Kratzke, N., Siegfried, R. (2020): Towards cloud-native simulations – lessons learned from the front-line of cloud computing. *J. Def. Model. Simul.*
- [6] Mell, P., Grance, T. (2011): The NIST Definition of Cloud Computing. National Institute of Standards and Technology.
- [7] Modelica Association (2020): Functional Mock-up Interface for Model Exchange and Co-Simulation Version 2.0.2. Modelica Association. Linköping.
- [8] Mons, B. u. a. (2017): Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Inf. Serv. Use.* Vol. 37, Nr. 1, S. 49–56.
- [9] MSG-131, S.T. (2015): Modelling and Simulation as a Service: New Concepts and Service-Oriented Architectures. North Atlantic Treaty Organization NATO.
- [10] Shahin, M., Babar, M.A., Chauhan, M.A. (2020): Architectural Design Space for Modeling and Simulation as a Service: A Review. *J. Syst. Softw.* Vol. 170, S. 110752.
- [11] Stüber, M., Exel, L., Frey, G. (2018): Using Modelling and Simulation as a Service (MSaaS) for Facilitating Flexibility-based Optimal Operation of Distribution Grids. *Proceedings of the 15th International Conference on Informatics in Control, Automation and Robotics - Volume 2: ICINCO.* SciTePress. S. 613–620.
- [12] Stüber, M., Frey, G. (2021): A Cloud-native Implementation of the Simulation as a Service-Concept Based on FMI. *Proceedings of 14th Modelica Conference 2021*, Linköping, Sweden, September 20-24, 2021. Linköping University Electronic Press.
- [13] Wilkinson, M.D. u. a. (2016): The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data.* Vol. 3, Nr. 1.
- [14] (2020): Sorbonne declaration on research data rights, <https://sorbonnedatadeclaration.eu/data-Sorbonne-declaration.pdf>.

Strukturgraph der Produktion

Ein Konzept für Prozessdatenintegration und -analyse mit Anwendungsbeispielen

Structure Graph of Production:

A concept for process data integration and analysis with application example

W. Li M.Sc., **M. Winter** M.Sc., Prof. Dr. rer. nat. **C. Roos**,
Prof. Dr.-Ing. **T. Kleinert**, RWTH Aachen University, Aachen

Kurzfassung

Aufgrund der ständig steigenden Komplexität von Produktionsanlagen und -prozessen ist eine Ursachen-Wirkungs-Analyse häufig schwierig. In diesem Beitrag stellen wir einen Modellierungsansatz vor, mit dem die Produktionsanlagen und -prozesse, Feldgeräte und Produkte sowie deren Eigenschaften in einem Modell, dem „Strukturgraph der Produktion (SGP)“, strukturiert abgebildet werden. Dabei werden Einflussfaktoren und die Beziehungen zwischen ihnen als Knoten und Kanten im SGP modelliert. Darüber hinaus stellt SGP eine Methode zur Verfügung, um die in der Datenbank gespeicherten Daten im SGP-Modell zu assoziieren, so dass die Daten leichter zugänglich sind und eine direkte Interpretation durch eine solche Verknüpfung möglich ist. Sind die Einflussfaktoren und ihre Beziehungen einmal ermittelt, werden diese im SGP-Modell visualisiert. Um solche Einflussfaktoren und ihre Beziehungen für zukünftige Analysen nutzbar zu machen, werden diese bestimmten Prozessanlagen, Prozessschritten und Zwischen- oder Endprodukten eindeutig zugeordnet. In diesem Beitrag werden die Modellelemente mit Hilfe von UML-Klassendiagrammen formal definiert und eine Benutzeroberfläche zur Handhabung des SGP vorgestellt. Der Modellierungsansatz wird exemplarisch an einem Beispiel aus einer Laboranlage demonstriert.

Abstract (optional)

Due to the increasing complexity of production plants and processes, a cause-effect analysis is often difficult. In this paper, we present a modeling approach with which the production plants and processes, field devices and products as well as their properties are mapped in a structured way in a model, the "Structure Graph of Production (SGP)". Influencing factors and the relationships between them are modeled as nodes and edges in the SGP. In addition, SGP

provides a method to associate the data stored in the database in the SGP model, making the data easier accessible and allowing direct interpretation through such association. Once influencing factors and their relationships are identified, they will be visualized in the SGP model. In order to make identified such factors and their relationships useful for future analysis, they will be uniquely associated with specific process equipment, process steps, and intermediate or final products. In this paper, the model elements are formally defined using UML class diagrams and a user interface is provided for handling the SGP. The modeling approach is demonstrated using an example from a laboratory plant.

1. Introduction

With the development of Industry 4.0 technologies, numerous manufacturers are establishing production digitization as their corporate strategy to maintain and increase competitiveness. More and more data on production facilities and processes is collected and stored to answer recurring questions e.g., causes of machine downtimes or product quality defects.

For such cause-effect analysis, all potential influencing factors have to be gathered, some of which are, however, often difficult to identify. Therefore, efficient methods for determining the influencing factors are necessary. In addition, the compilation of available historical data is of great importance. For data processing, meta-information is needed i.e., on data availability and on where and how to access it. Once the influencing relationships are determined, they should be reasonably explained and visually represented. In order to make influencing factors and their relationships accessible and usable for future analysis, they must be clearly associated to certain process elements, process steps, and intermediate or finished products.

The keyword "data" has a central role in these scenarios, where manufacturers need a sensible and practical concept for managing data and information. In this paper, a novel and generalized modelling approach based on the "Structure Graph of Production" (SGP) is presented. To address the need, the following central questions will be answered, which are derived from the interdisciplinary areas of activity:

- How to collect the information on plant structure, production process, material, and data flow as well as the corresponding parameters and properties within an integrated approach?
- How is the required information modelled and presented to achieve the goal of data integration and causal analysis?
- How will the required data be accessed and made usable for a specific data analysis project using the SGP?

- How will relevant influencing relationships be identified and represented in the model and visualized in the SGP?
- How can changes and modifications be integrated?

Section 2 addresses the related work. Section 3 introduces the concept of SGP modeling. Section 4 describes potential methods for identifying influencing relationships between model elements. Section 5 demonstrates the modelling approach on a demonstrative example.

2. Related work

Causal questions can be answered through combining data analysis and process model [18]. Traditional cause-effect analysis, namely Ishikawa-diagram is a management method, which provides a collection of qualitative influencing factors in several categories including measurements, methods or processes, machines and materials [11]. A quantitative causal analysis is driven by mathematical methods and data of the influencing factors [17], typically available from engineering databases and Plant Information Management System (PIMS) [2]. Plant engineering documentation e.g., piping and instrumentation diagram (P&ID) or procedures, provides information on deployed equipment, production processes, measuring systems and their parameters (properties) [22]. Operating data of a production, including data of machines and field devices, processes, materials and jobs, is stored in PIMS [16]. Historical data in PIMS can be used for determining causes of product defects or machine downtimes. [2]. Maintenance and optimization measures taken in the past are documented in dedicated systems (typically ERP), which can also be used to analyze influencing factors [13].

Properties of elements in a production system can be accessed through measuring systems and their values are stored in data variables. Relationships between data variables in a production system can be analyzed by causal graph, which delineates causal mechanisms of the system [15]. In a causal graph, data variables are modelled as vertices (nodes) and relationships between variables are described using (directed) edges (i.e., arrows) indicating that a change in X possibly causes a change in Y [17]. However, variables in such a pure causal graph are usually not categorized or annotated, which is necessary for automated data handling and interpretation, especially for a growing number of variables and relationships.

The *phase model of production (PMP)* provides system analysis for process engineering production processes [19], which consists of a formal description language [5] with partitioned sub-processes of in- and output ports that are connected via material or energy flows [12]. Additionally, a link between the processes and the technical equipment as well as the information flow can be represented. This supplementary information allows represent knowledge on the processes, which establishes a networked information base that can support

a causal analysis [21]. A transition from the product net to the causal net can be made evident with the help of a Petri net and consideration of time [20]. However, a direct synthesis between both nets is yet to be established [9].

The authors introduced the idea of "Structure Graph of Production" in a discussion paper submitted to MATHMOD 2022 [14]. Based on this, the concept for structuring information of production plant, process and product with the purpose of building a net of influencing relationships is introduced in detail, which provides the general basis for an integrated and automated analysis and visualization of influencing relationships in process data.

3. Concept of the structure Graph of Production

3.1 Overview of the Structure Graph of Production

The quality of produced goods is affected by a multitude of different process parameters, control variables, machine conditions and conditions of the preliminary products. Thus, an analysis of these factors is necessary to identify potential influencing variables and to formally verify them.

A production system in the current consideration is composed of production processes, machines, raw materials, produced (intermediate) products, actuators, and measuring devices, which are considered as central elements of the SGP:

- A production process consists of several process steps, in which products are processed and their states are changed [20]. A process step is characterized by process parameters, which are target variables affected by machines in order to achieve the required product quality state.
- A producing machine consists of one or more machine parts. Each machine part is assigned to one or more process steps and affects a processed product with its corresponding parameters to achieve the process specifications. A variation or deviation in the specification of the parameters causes variation in the product quality. An actuator is also considered a machine part that purposefully influences changes in certain properties of a product. The actuator is usually guided by open- or closed-loop control.
- A product has several properties that are determined by all upstream process steps and the used machines. A raw material is considered as an input product at the beginning of a considered production process.
- A measuring device has one or more measuring modules that measure certain properties of a process, machine or product. It is assumed that this device has no direct feedback effect on the measured object.

Modelling approaches for dynamic systems in areas of e.g., economics, biology and engineering are widely discussed [15], where various possible factors need to be taken into account when building such system models. In order to analyze the causes of situations such as product quality defects and equipment wear, the above-mentioned elements of a production system have to be included, which are considered objects for SGP modelling. However, in contrast to traditional dynamic system modelling, the aspects of metadata and automatic data integration are especially in scope of SGP modeling. The idea of nodes and edges used for causal graph modelling will apply

3.2 Modelling language for Structure Graph of Production

SGP can be modelled and visualized e.g., by using the Graphviz DOT Language [7]. The metamodel of Graphviz [4] is used to define model elements of SGP. For the SGP, this metamodel is extended with a further Graph Element *subgraph* (Fig. 1). A graph consists of nodes, edges and subgraphs [8], which can have a list of attributes e.g., label, color, style or shape. A subgraph has a graph-like structure, which groups graph elements and further subgraphs, indicating certain semantic information about the grouped graph elements. Subgraphs have a special form of cluster subgraphs, which enables associate graph elements to the cluster. However, *cluster subgraphs* are only a layout convention [8]. Therefore, it is not part of the metamodel.

Due to the consistent use of terms in the presented concept, the definitions described in the following sections are introduced. A production system is composed by various *production elements* including process steps, machines, products and field devices. The term *Product* includes raw materials, intermediate products and finished products. An *influencing factor* is a *property* of a production element, whose values are stored in a *data variable* in a database.

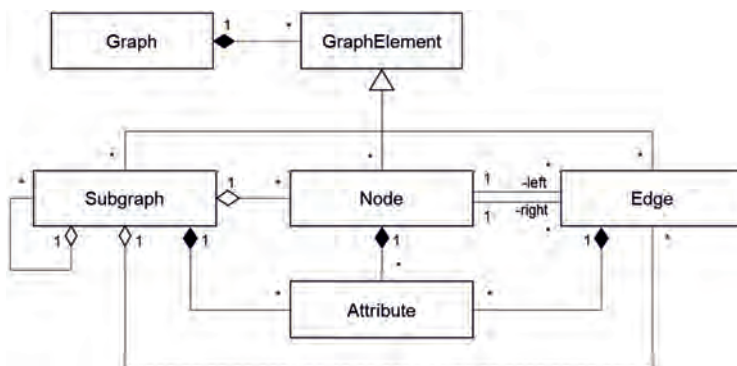


Fig. 1: Graphviz DOT metamodel extended by Subgraph

3.3 Model elements of Structure Graph of Production

The SGP class model (Fig. 2) is the basis for the intended representation of a production system by abstracting and structuring the properties of the considered system elements. The PMP [20] models production processes and products in SGP. The system elements are described by properties (nodes) and are modelled as property clusters. The term *property cluster* refers to cluster subgraphs of properties and is, therefore, of type *Subgraph*. The *relationship* between properties, across clusters or within a cluster, are components of the SGP or cluster, respectively. The modelling of nodes, relationships, PMP, and property clusters and nodes are described in more detail below.

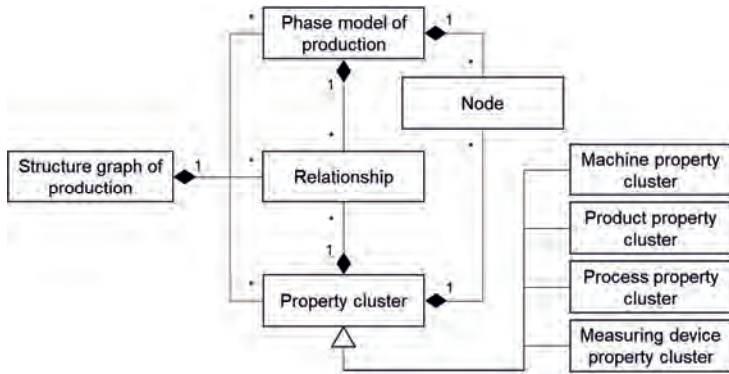


Fig. 2: General SGP class model

3.4 Modelling of nodes and relationships

Nodes and relationships are the most basic model elements of the SGP. A *node* (Fig. 3) has a uniquely identifiable ID that can be represented by a textual label. The class node has two subclasses: structural node and property node. *Structural nodes* can be product or process nodes, representing process steps or products in the PMP. This type of node has no further attributes. *Property nodes* describe certain aspects of a machine, product, process and measuring device. Each property node has two further attributes: (1) *associated cluster* to which the node belongs to; (2) *associated data variable* with basic metadata (e.g., name, data type and location) which are referencing the data stored in a database.

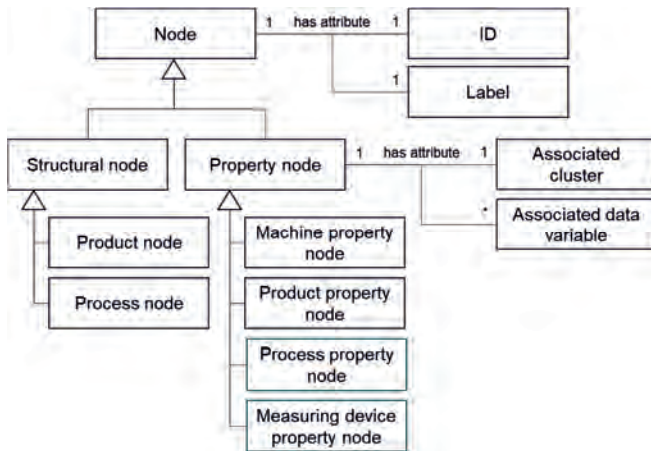


Fig. 3: Node class model

A *Relationship* (Fig. 4) is a subclass of an edge, which is represented by a directed line pointing from a source node to a target node. Each relationship is uniquely identifiable in a SGP by its ID or by a node tuple (source node, target node). A Relationship has two subclasses: structural or influencing relationship. A *structural relationship* represents the direction of the material flow in PMP or the data flow between the property nodes of a measuring device cluster and a machine, product or process cluster, respectively. An *influencing relationship* (causal or correlational) indicates that the source node may have an influence on the target node with a certain probability and intensity, which can be determined by using methods of either data analysis, expert knowledge or experimental observation (see Section 4). When an influencing relationship is valid, its attribute "isValidated" is set TRUE.

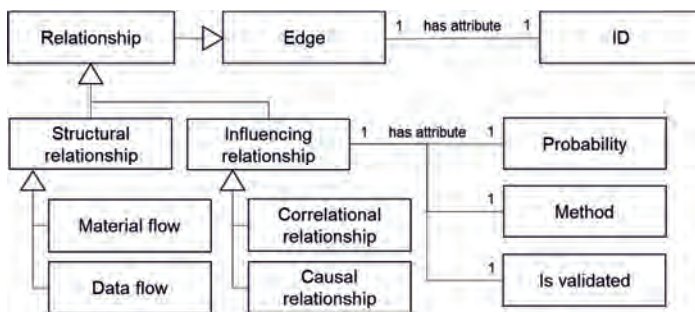


Fig. 4: Relationship class model

3.5 Modelling of Phase Model of Production

The PMP models production processes with process steps and products alternately. A PMP model consists of structural nodes and relationships (Fig. 4) that form the fundamental structure of the SGP. Process steps are represented by process nodes and product stages by product nodes. A structural relationship connects always one product node to a process node (or vice versa) and suggests the material flow direction.

The PMP model can be considered as the skeleton of the SGP model. It is used to structure and link all property clusters in SGP, which is connected to corresponding nodes in PMP. Since process and product clusters contain the properties of processes and products, they are assigned to the described process and product nodes. The same is applied to machine and measuring device clusters. In each process step, certain machines and possible measuring devices are deployed to execute process tasks, affecting or measuring states of process and product, respectively. Therefore, their property clusters shall be located in the same layer of the corresponding process or product nodes. In Section 3.6, the modeling of the property cluster and property node is described in detail.

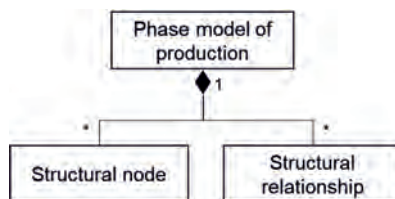


Fig. 5: Phase Model of Production (PMP) class model

3.6 Modelling of property cluster and property nodes

A *property cluster* (Fig. 6) consists of property nodes, influencing relationships and further property clusters. For instance, a machine can include several machine parts and can be described by a machine property cluster. Therefore, the property cluster of the whole machine can be seen as the parent cluster, and those of the machine parts are child clusters.

Corresponding to the property cluster subclasses, four subclasses of property nodes (Fig. 6) are derived from the parent class. A *property node* (except that of measuring devices) is actually or potentially measurable, which is given by the attribute “Measurability” (Boolean). In addition, each *machine property node* takes on two more attributes: “Controllability” (Boolean) and “Control Mode” (Automatic, Manual or None). As soon as controllability is given, a distinction is made between an automatic and a manual Control Mode. Otherwise, this property is occupied by “None”. A *product property node* describes certain product state. The product

states that correspond to or influence certain quality characteristics of the final product are considered quality-related, otherwise, non-quality related states. A *measuring device property node* delineate the measurement capability of a measuring device. Therefore, the measuring principle and measuring point are to be given for the nodes.

3.7 Visualization of Structure Graph of Production

SGP can be visualized e.g., using Graphviz. Visualization conventions for distinction of different nodes are proposed. In Fig. 7, a SGP is illustrated in a general manner. The PMP as well as the property nodes of the system elements are represented by different shapes, colors and frames. The PMP is represented by blue nodes and edges, from "Raw materials" to "Process 1" and "Intermediate product" to "Process 2" and "Final product". The direction of the edge corresponds to the direction of the material flow.

Oval nodes, with frame (quality related) or without frame (non-quality related), in green (measurable) or grey (non-measurable), represent product and process properties. Property clusters in rounded rectangles (surrounding such property nodes) are assigned to the corresponding blue product and process nodes.

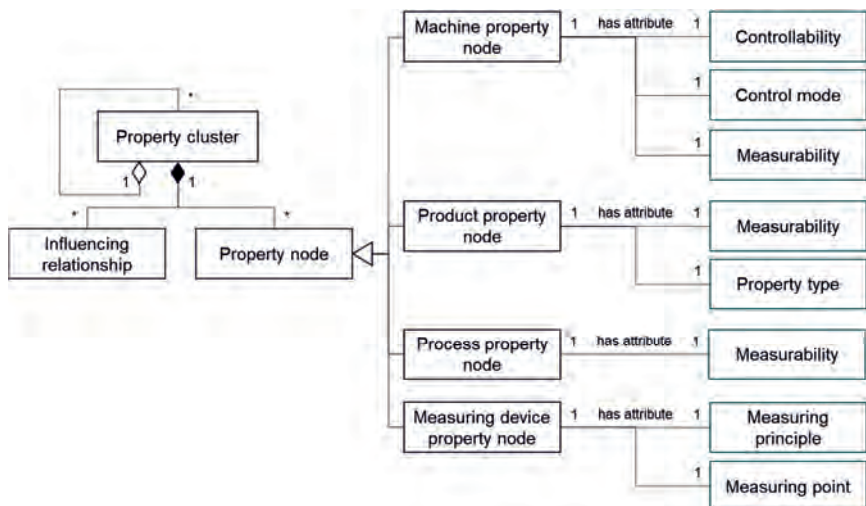


Fig. 6: Property cluster and property node class model

Rectangle nodes represent measuring device properties as well as non-controllable machine properties, while arrow nodes represent controllable machine properties, with frame (automatic) or without frame (manual). The color coding is the same as for product and process

property. Property clusters in normal rectangle (enclosing such property nodes) are located on the level of corresponding blue product and process nodes.

A property can potentially be influenced and it can also have an influence on other properties. These influences are modelled as black directed edges between these nodes. Dashed edges show the data flow, which indicates that properties are measured by measuring devices.

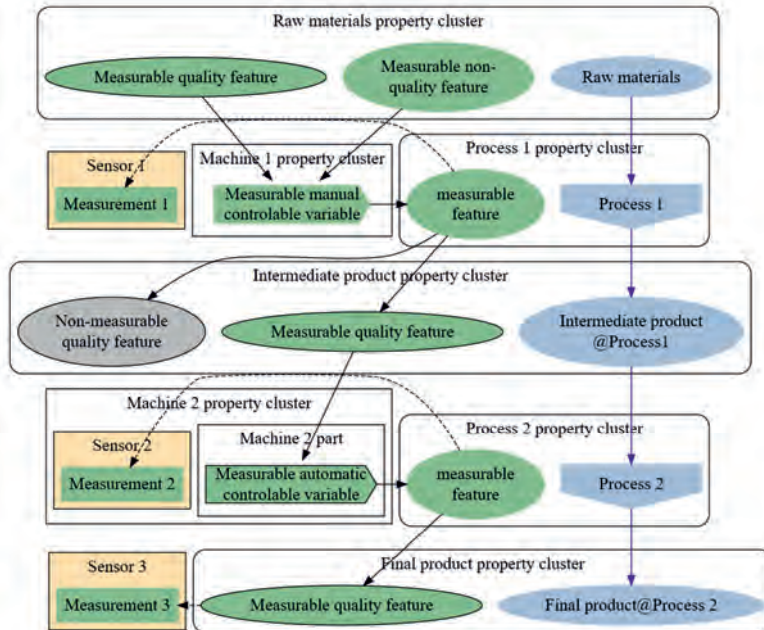


Fig. 7: General example of a structure graph of production.

4. Methods to Identify the Influencing Relationships of the SGP

4.1 Necessity of identifying influencing factors

Numerous methods are available for data analysis. These are applied in preparation towards the concept of data integration and analysis as well as determining the relationship dependencies between the individual machine or product attributes regarding the influencing relationships. A dependency may be quantifiable which, however, may vary over time. Thus, a weighting of the dependencies in the working states must take place. Therefore, the weights of the particular edges from a source node to the target node need to be identified. Furthermore, the edges are uniquely labeled. If a sufficient number of edges on relevant

properties are identified and weighted, the SGP can be used as basis to analyze cause and effect behavior of the modelled production process.

As a first step, expert knowledge can be used to determine a structure graph with various nodes and their corresponding relationships. These relations are directed in order to build a causality model. Initially, edges that skip intermediate nodes should be avoided. Thus, the causality can be followed up along the nodes in a sequential order. The bare representation of edges does not provide influencing factors. For quantitative causality analysis, the influencing relationships are required with individual weighting of the edges which can have different forms of mathematical functions.

4.2 Correlation analysis to identify influencing factors

Correlation analysis is one approach to determine quantitative magnitude of influencing relationships in SGP. It is important to note that, correlation does not indicate causation [1]. The correlation analysis in the current work is used for searching possible influencing relationships (and factors). For instance, if values of the characteristics X and Y are arranged in such a way that for increasing one of the characteristics also the other tends to have larger values, it is reasonable to suspect a relationship between them. A measure of the strength of this relationship is the well-known empirical correlation coefficient, also known as the Bravais-Pearson correlation coefficient [6]. This method allows the identification of the relationships between two or more properties in the SGP. The requirements for performing the Bravais-Pearson correlation are fulfilled with the normal distribution and at least an interval scaling of two distinct characteristics X and Y . Furthermore, the examined relationship between these characteristics must be linear [10].

The correlation analysis aims at determining dependencies between two or more characteristics on the basis of measured values [3]. Therefore, the scattering of each characteristic X and Y (σ_X, σ_Y) is to be determined as well as the covariance between X and Y (σ_{XY}) to calculate the correlation coefficient:

$$\varrho_{XY} = \sigma_{XY} / (\sigma_X \sigma_Y)$$

The correlation coefficient is a value for linear dependency between X and Y . This is applied, when all points (X, Y) are lying on a straight line with probability 1, namely:

$$\varrho_{XY}^2 = 1.$$

If X and Y are independent random variables, then:

$$\varrho_{XY} = 0$$

applies [3].

This approach can be used to make inferences about the direction and strength as well as to identify influencing relationships in SGP in general. The direction can be positive or negative. The correlation coefficient is between -1 and 1, depending on a negative or a positive correlation between the values of the characteristics X and Y.

5. Use case and example of a SGP

In the following, the modelling approach of SGP is explained with a simple demonstrative example of a heating system of a laboratory plant. Firstly, information about the plant structure and the process is needed e.g., P&ID (Fig. 8), loop list, process parameters, expert knowledge and other related information. As shown in Fig. 8, Pump N13 feeds the outflow from an upstream tank B1 into a heat exchanger and feeds it further to the downstream processes through valve Y16. A temperature sensor (T15) and a flow meter (F17) measure the process and product properties, respectively. The pumping power, heating power and valve opening are controlled automatically.

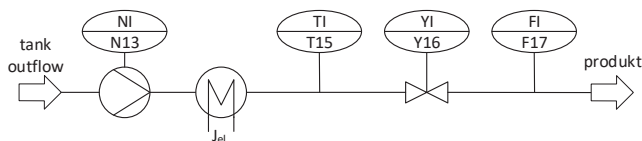


Fig. 8: The partial P&ID of a laboratory plant

With the given information, the process is modelled as PMP and the respective properties of transporting and heating process, water, pipelines and sensors are modelled by following the rules defined in Section 3.3 - 3.6. The displayed property nodes are all measurable, whereby corresponding meta-information of measurement data is stored in the attributes of the nodes. For instance, the flow rate values measured by the flow sensor F17 have data type float and are stored in the historian database "Pumpwerk.DB" with the path '//F17/PV/PV.CV'. Thus, measurement data can be extracted and analyzed for identifying correlational and causal relationships between the properties. The influencing factors of warm water temperature are identified by correlations analysis. The identified relationships are then added in the SGP model containing the corresponding correlation coefficient, which can be queried by need. As shown in Fig. 9, The SGP model is visualized in a tool using the defined convention described in Section 3.7. With the help of the tool, an analyst can, for example, filter nodes, edges and clusters to view the SGP in desired way. It is obvious that the temperature of warm water is directly influenced by the heater power. Results of correlation analysis also suggest

that the temperature of outflow water from tank B1, the heater setpoint as well as the flow rate measured by F17 are correlated with the heater power due to the automatic control of the heating power. Through simple experiment design, causalities (red edges) are derived from the observed correlations. It is also obvious that the pump power and the valve opening are influencing the flow rate in the transport process.

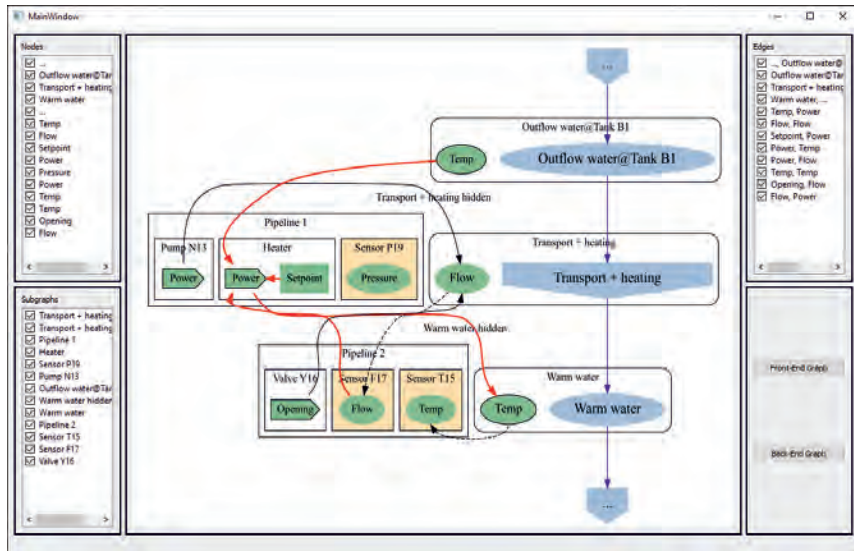


Fig. 9: SGP model of the heating process shown in Figure 8

6. Conclusion

The authors introduced the basic concept and formal description of the Structure Graph of Production (SGP) to systematically model elements of a production system, relationships between their properties, and their semantic link to measurement data. By grouping properties in clusters and associating data variables with the properties, SGP enables semantic interpretation of data and analysis results. In addition, the authors realized the SGP as a python package and provide a graphical user interface, which enables create, modify and visualize SGP models. The SGP package also provides an interface to other exiting python packages, which enables graph analysis. Furthermore, the SPG considers aspects of dynamic system modelling, data processing and causal analysis in its modeling systematic, which lays a basis for automated data integration, interpretation and visualization to describe causal mechanisms

within the production system. A review of other methods for automated data integration and causal analysis was out of the scope of the presented work.

In future work, the authors will evaluate the SGP and conduct further reviews of methods in this area. The SGP will be further developed towards automated data integration and causal analysis. Moreover, aspects of changeability, scalability and user-friendly visualization will also be considered.

Acknowledgements: The modelling approach of SGP was prepared for application to an industrial glass container production line in cooperation with International Partners in Glass Research e.V. (IPGR). The authors wish to thank IPGR for supporting this study. Special thanks go to Dr. Domink Orzol for numerous productive discussions and suggestions.

- [1] Aldrich, J. (1995). Correlations Genuine and Spurious in Pearson and Yule. *Statistical Science*. 10 (4): 364–376. doi:10.1214/ss/1177009870.
- [2] Berge, J. (2005). *Software for automation: Architecture, integration, and security*. ISA, North Carolina, U.S.
- [3] Bronstein, I. N., & Semendjajew, K. A. (2012). *Taschenbuch der Mathematik* (8., voll. überarb. Aufl.), p. 852. Hades. Frankfurt am Main: Wissenschaftlicher Verlag Harri Deutsch GmbH.
- [4] Demirezen, Z., Sun, Y., Gray, J., Jouault, F. (2009). Supporting Tool Reuse with Model Transformation. In: 18th International Conference on Software Engineering and Data Engineering (SEDE-2009), Las Vegas, Nevada.
- [5] Ebert, J., Schulze, M. (2003). Semantics of the Phase Model of Production: Formalization. In *Proceedings of MATHMOD 2003*, Vienna.
- [6] Fahrmeir, L., Heumann, C., Künstler, R., Pigeot, I., & Tutz, G. (2016). *Statistik: Der Weg zur Datenanalyse*, p. 126. Berlin, Heidelberg: Springer Berlin Heidelberg
- [7] Graphviz. <https://graphviz.org/>. Accessed 15 Dez. 2020.
- [8] Graphviz - Subgraphs and Clusters. <https://graphviz.org/doc/info/lang.html>. Access 12 Jul. 2021
- [9] Gruner, S., Weber, P., & Eppele, U. (2014). A model for discrete product flows in manufacturing plants. In *Proceedings of the 2014 IEEE Emerging Technology and Factory Automation (ETFA)* (pp. 1–8).
- [10] Hartung, J., Elpelt, B., & Klösener, K.-H. (2012). *Statistik: Lehr- und Handbuch der angewandten Statistik* (15., überarb. Aufl.). München: Oldenbourg: Oldenbourg.
- [11] Ishikawa, K. (1976). *Guide to Quality Control*. Asian Productivity Organization, Tokyo, Japan.

- [12] Lauber, J. (1994). Phase model of production for process description and process analysis. In Sympos. 94 'Automat. of Mixed Processes: Dyn. Hybrid Syst.' Brüssel, Belgien
- [13] Li, Z., Wang, Y. & Wang, K.S. (2017). Intelligent predictive maintenance for fault diagnosis and prognosis in machine centers. In Industry 4.0 scenario. Adv. Manuf. 5, 377–387.
- [14] Li, W., Winter, M., Roos, C., Kleinert, T. (2022). Submitted to MATHMOD 2022. 10th Vienna Conference on Mathematical Modelling, Vienna, Austria, July 27-29, 2022.
- [15] Lloret-Climent, M., Nescolarde-Selva, J., Alonso-Stenberg, K., Selva-Barthelemy, M. (2019). Causal Analysis of the Spanish Industrial Sector Through Smarta. IEEE Access, 7, p.33556–33564.
- [16] Mertens, P. (2013) *Integrierte Informationsverarbeitung 1: Operative Systeme in der Industrie*. Gabler Verlag, Wiesbaden, Germany
- [17] Pearl, J. (2009). *Causality: Models, reasoning, and inference*. Cambridge, U.K, New York: Cambridge University Press.
- [18] Pearl, J. (2018). *The book of why: the new science of cause and effect*. Chap. 10. Basic Books, New York, U.S.
- [19] Polke, M. (1992). Phasenmodell der Produktion als nicht-mathematische Präsentationsform für Automatisierungs-aufgaben. In Entwurf komplexer Automatisierungs-systeme, Braunschweig, Germany
- [20] Polke, M., Eppler, U., & Ahrens, W. (Eds.) (1994). *Process control engineering (1st ed.)*, p. 34 – 39, VCH, Weinheim
- [21] Schulze, M. Ebert, J. (2002). Prozessmodell-basierte Präsentation von Produktionsfehler-Beschreibungen. In Modellierung 2002, Modellierung in der Praxis - Modellierung für die Praxis, Tutzing, Germany
- [22] Tauchnitz, T. (2017) Schnittstellen ermöglichen Daten-integration in der Prozessindustrie. In: *Handbuch Industrie 4.0* Bd.2. Springer Vieweg, Berlin, Heidelberg.

Konzeption eines vereinheitlichenden Energieinformationsmodells zur Integration unterschiedlicher Energieprofile

Conception of a unifying energy information model for the integration of different energy profiles

Leif-Thore Reiche M.Sc., Prof. Dr.-Ing. **Alexander Fay**,
Helmut-Schmidt-Universität / Universität der Bundeswehr, Hamburg;
Maxim Runge M.Eng., Prof. Dr.-Ing. **Karl-Heinz Niemann**,
Hochschule Hannover, Hannover

Kurzfassung

Zur Übertragung von Energieinformationen werden in den industriellen Kommunikationssystemen verschiedene Energieprofile eingesetzt, die sich allerdings hinsichtlich ihrer Inhalte und Semantik unterscheiden. Dies führt zu unterschiedlichen Beschreibungen für identische Informationen. Diese Unterschiede erschweren es den Anwendern, die Energieprofile auf einfache Art in die vorhandenen Energie-Management-Systeme zu integrieren. Dieser Beitrag zeigt mittels eines Vergleichs zum einen Unterschiede und Gemeinsamkeiten von ausgewählten Energieprofilen auf. Zum anderen wird das Konzept für ein vereinheitlichendes Energieinformationsmodell vorgestellt, mit dem es möglich ist, die Energieinformationen verschiedener Energieprofile interpretieren und standardisiert nutzen zu können.

Abstract

For the transmission of energy information, various energy profiles are used in industrial communication systems, but these differ in terms of content and semantic. The differences make it difficult for users to integrate the energy profiles into existing energy management systems. This paper shows differences and similarities of the energy profiles by a comparison. Furthermore, the concept for a unified energy information model is presented that allows to interpret and process the energy information of different energy profiles in a standardized manner.

1 Einleitung

Vor dem Hintergrund des voranschreitenden Klimawandels steht insbesondere der industrielle Sektor vor der Herausforderung Maßnahmen zu ergreifen, mit denen die Energiewende reali-

siert werden kann. In den kommenden Jahren spielt daher für viele Unternehmen die Energieeffizienzerhöhung der eigenen Produktionsstätten eine wichtige Rolle. Mit der Einführung eines Energiemanagementsystems nach ISO 50001 [1] können diese Bestrebungen maßgeblich unterstützt werden, denn ein derartiges Managementsystem ermöglicht es, energieeffizienzrelevante Ziele zu definieren und diese zu verfolgen [12]. Damit diese Ziele in der realen Welt der Produktion auch umgesetzt werden können, bedarf es der Einführung eines technischen Energiemanagementsystems (kurz tEnMS). Mit Hilfe eines tEnMS können Energieinformationen in der Feldebene erfasst, gespeichert und anschließend über das industrielle Kommunikationssystem an höher liegende Ebenen des Automatisierungssystems (z. B. Betriebsleitebene mit MES) zur Darstellung und Überwachung übermittelt werden. Außerdem erlaubt das tEnMS, dass Steuerbefehle aus der Betriebsleitebene zur aktiven Beeinflussung der elektrischen Lasten in die Feldebene übermittelt werden können und der Betrieb in der Feldebene so energieeffizienter gestaltet werden kann.[17]

Um Energieinformationen zwischen der Feld- und der Steuerungsebene austauschen zu können, werden im Rahmen der industriellen Kommunikation sogenannte Energieprofile eingesetzt. Diese Profile ermöglichen es, dass die Energieinformationen unter Nutzung des jeweils verwendeten Protokolls in einem standardisierten Format übermittelt werden können [18]. Aufgrund des Einsatzes von unterschiedlichen Bussystemen (z. B. PROFINET [19] oder CIP [20]) in der Feldebene werden dementsprechend auch verschiedene Energieprofile mit unterschiedlich ausgeprägten Eigenschaften und Semantiken eingesetzt.

Damit die Energieinformationen von der Steuerungsebene im Anschluss weiter über die Prozessleitebene bis zu Betriebsleitebene kommuniziert werden können, müssen diese auf andere Protokolle (z. B. OPC UA [21]) übersetzt werden (Mapping). Dieser Übersetzungsprozess birgt jedoch das Risiko, dass energiespezifische Informationen später nicht richtig interpretiert werden können oder sogar verloren gehen, wenn die Informationen auf dem Zielprotokoll nicht richtig übersetzt und abgebildet werden können. Mit Hilfe von sogenannten OPC UA Companion Spezifikationen (Erweiterung der bestehenden Energieprofile zur Abbildung der Informationen auf OPC UA) wird aktuell versucht dieses Problem zu beheben. Allerdings werden für die verschiedenen OPC UA Companion Spezifikationen unterschiedliche Semantiken verwendet. Anwender stehen somit oft vor der Herausforderung einen, hohen Engineering-Aufwand zu leisten, um die Energieinformationen von unterschiedlichen Energieprofilen in ihr übergeordnetes Kommunikationssystem einheitlich integrieren zu können, damit sich die Informationen auf der Betriebsleitebene standardisiert auswerten lassen. Dieses Problem findet sich besonders in Produktionsanlagen, die heterogene Bussysteme einsetzen, da die Energieinformationen der einzelnen Anlagenteile durch verschiedene Energieprofile beschrieben werden

können und dementsprechend mehr Zeit für die Zusammenführung der Energieinformationen und Integration in das übergeordnete Kommunikationssystem benötigt wird.

Damit die Engineering-Aufwände bei der Verwendung verschiedener Energieprofile zukünftig reduziert und die Aufwände für die Erstellung weiterer Companion Spezifikationen vermieden werden, wird eine Lösung benötigt, mit der sich Energieinformationen aus heterogenen Bussystemen einheitlich in ein übergeordnetes Kommunikationssystem abbilden lassen.

Dieser Beitrag verfolgt daher zwei Ziele: Zum einen soll aufgezeigt werden, welche Energieinformationen im Rahmen der industriellen Kommunikation mittels der Energieprofile übertragen und wie diese semantisch beschrieben werden. Zum anderen wird das Konzept eines vereinheitlichenden Energieinformationsmodells vorgestellt, das bei Anwendung die Möglichkeit bietet, heterogene Energieinformationen verschiedener Energieprofile zu erfassen und einheitlich zu beschreiben. Dadurch ist es möglich den aktuell bestehenden Engineering-Aufwand für die Aggregation der Energieinformationen aus verschiedenen Energieprofilen in ein übergeordnetes Kommunikationssystem (z. B. OPC UA) zu reduzieren. Zukünftig lassen sich somit auch die Energieinformationen von weiteren Energieprofilen vereinfacht integrieren, sodass nicht für jedes der Energieprofile eine eigene OPC UA Companion Spezifikation erstellt werden muss.

Der Beitrag ist wie folgt aufgebaut: Im folgenden Kapitel 2 werden der Aufbau, die Eigenschaften und die bereitgestellten Funktionen von verschiedenen Energieprofilen erläutert. Daran anschließend werden in Kapitel 3 ausgewählte Energieprofile miteinander verglichen. Anhand dieses Vergleichs können Gemeinsamkeiten und Unterschiede für die Eigenschaften und Funktionsumfänge aufgezeigt werden und so Aussagen darüber getroffen werden, welche Energieinformationen vereinheitlicht in ein übergeordnetes Kommunikationssystem übertragen werden sollten. Kapitel 4 erläutert, welche Vorteile die Nutzung eines semantisch beschriebenen, vereinheitlichenden Energieinformationsmodells bietet, das Energieinformationen über verschiedenste Kommunikationsschnittstellen (z. B. basierend auf OPC UA, MQTT) bereitstellen kann.

2 Stand der Technik

Dieses Kapitel dient dazu einen Einblick in die integrierte Energieinformationsübertragung zu geben und dabei verwendete Energieprofile darzustellen.

2.1 Integrierte Übertragung von Energieinformationen

In [22] werden von den Autoren zwei Möglichkeiten aufgezeigt, wie ein tEnMS realisiert werden kann. Die erste Möglichkeit besteht darin, parallel zum bestehenden Automatisierungssystem

ein weiteres System zu installieren, das ausschließlich für die Übertragung der Energieinformationen genutzt wird. Der Nachteil dieser Lösung besteht darin, dass zusätzliche Kosten für die Hardware, Software, Installation, Wartung und Instandhaltung dieser Zusatzsysteme anfallen. Die zweite Möglichkeit nutzt die Ressourcen des bereits vorhandenen Automatisierungssystems, indem die Energieinformationen vom Feldgerät über das bestehende Kommunikationssystem übertragen werden. Abbildung 1 zeigt diese integrierte Energieinformationsübertragung.

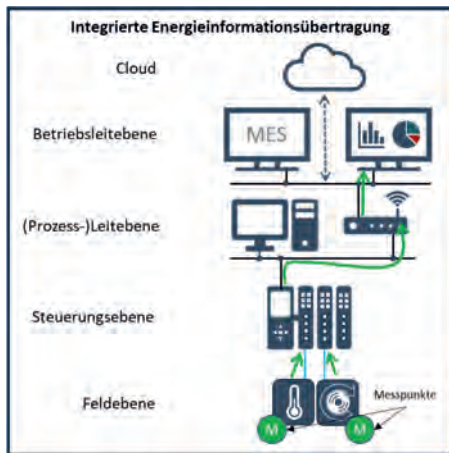


Bild 1: Integrierte Energieinformationsübertragung in Anlehnung an [22]

Im Allgemeinen können die Energieinformationen zwischen der Feld- und der Betriebsleitebene bidirektional übertragen werden. Dieser bidirektionale Informationsfluss besteht zum einen aus Energieinformationen, die für ein Energie-Monitoring aus der Feld- über die Steuerungs- bis in die Betriebsleitebene (MES) versendet werden, in der die Informationen dann über eine Energiemanagement-Applikation aufbereitet und ausgewertet werden können [23]. Zum anderen können Energieinformationen (z. B. Standby-Befehle) im Rahmen eines Lastmanagements aus der Energiemanagement-Applikation auf der Betriebsleitebene über die Steuerungsebene bis in die Feldebene versandt werden, um den Betrieb der Aktuatoren (z. B. Frequenzumrichter) im Sinne der Energieeffizienz zu beeinflussen. Dies ist immer dann sinnvoll, wenn Pausenzeiten oder Maschinenstillstände auftreten und die volle Betriebsbereitschaft der Aktuatoren nicht benötigt wird. [2]

Damit Energieinformationen im industriellen Kommunikationsnetzwerk über die verschiedenen Echtzeitprotokolle übermittelt werden können, werden sogenannte Energieprofile eingesetzt.

Diese Profile erweitern die bestehenden Protokolle, um die Funktionalität die Energieinformationen über das jeweilige Protokoll in einer standardisierten Semantik zu übermitteln. [18]

2.2 Energieprofile

Um ein besseres Verständnis für die Energieprofile im industriellen Kommunikationsnetz zu erlangen, werden nachfolgend verschiedene Energieprofile erläutert.

2.2.1 PROFenergy

Das Energieprofil PROFenergy [3] nutzt die azyklischen Dienste von PROFINET. PROFenergy spezifiziert zwei Energiemanagement-Funktionalitäten, die in der Spezifikation als „PE Entity class“ (PE Standby Management, PE Measurement Functionality) bezeichnet werden. Einerseits können mit PROFenergy Energieinformationen (z. B. Energiezählwerte in kWh) von einzelnen Aktuatoren (z. B. Frequenzumrichter) aus der Feldebene aufgezeichnet und bis in die Energiemanagement-Applikation übertragen werden. Andererseits können Lastmanagementfunktionen genutzt werden, indem einzelne Aktuatoren während Maschinenstillständen oder längeren Pausenzeiten über PROFenergy in Energiespar- oder Schlafmodi versetzt werden.

Um eine einfache Abbildung der in PROFenergy spezifizierten Energieinformationen auf das Kommunikationsprotokoll OPC UA zu ermöglichen, wurde PROFenergy um eine Erweiterungsspezifikation (sog. Companion Specification) [4] für OPC UA ergänzt. Mittels dieser zusätzlichen Spezifikation können die Energieinformationen von PROFenergy einheitlich auf OPC UA abgebildet werden.

2.2.2 SERCOS Energy

Sollen Energieinformationen mit dem Automatisierungsbuss Sercos (Sercos III) übermittelt werden, kann dies mit dem Sercos-spezifischen Energieprofil Sercos Energy [5] erfolgen. Sercos Energy bietet ebenfalls die Möglichkeit Energieinformationen bidirektional für ein Monitoring und Lastmanagement zu übermitteln, allerdings werden bei Sercos Energy die Energieinformationen mit einer eigenen Sercos-spezifischen Semantik beschrieben. Für Sercos III wurde ebenfalls eine OPC UA Companion Spezifikation veröffentlicht, mit der auch die Sercos Energy-Funktionalitäten abgedeckt werden [6].

Darüber hinaus wird in [7] angegeben, dass Feldgeräte, die Sercos Energy unterstützen, zukünftig auch in Teillastzustände versetzt werden können, um die Energieeffizienz der Produktionsanlage während geringerer Auslastungen zu erhöhen.

2.2.3 CIP Energy

Neben PROFenergy und Sercos Energy ist CIP Energy ein weiteres Energieprofil, das aus dem US-amerikanischen Markt verbreitete Common Industrial Protokoll (CIP) [8] der Standardentwicklungsorganisation ODVA ergänzt. Zu den CIP Protokollen gehört beispielsweise

auch EthernetIP. Mit CIP Energy können ebenfalls sowohl Lastmanagementfunktionen als auch Informationen für das Energie-Monitoring zwischen der Feld- und Betriebsleitebene übermittelt werden. Für CIP wurde die Entwicklung einer allgemeinen OPC UA Companion Spezifikation angekündigt [9].

2.2.4 Nicht-Energieprofil-basierte Energieinformationsübertragung

Auch Feldgeräte, die keine Energieprofil-basierte Energieinformationsübertragung unterstützen, können Energieinformationen bereitstellen. Die Bereitstellung der Energieinformationen kann zum einen durch Ethernet-basierte Kommunikationssysteme (z. B. Modbus TCP) [10] und zum anderen durch nicht-Ethernet-basierte Kommunikationssysteme (z. B. IO-Link) realisiert werden [11]. Die Semantik der Energieinformationen unterliegt dabei den proprietären Spezifikationen der Feldgeräte.

2.2.5 Bereitstellung der Energieinformationen bis zur Energiemanagement-Applikation

Die Energieinformationen der vorgestellten Energieprofil-basierten oder nicht-Energieprofil-basierten Feldgeräte werden auf Edge-Devices oder Speicherprogrammierbaren Steuerungen (SPS), falls erforderlich, im Datenformat angepasst und skaliert. Damit die Energieinformationen im tEnMS bereitgestellt und verarbeitet werden können, müssen diese dem übergeordneten Kommunikationssystem (z. B. basierend auf OPC UA oder MQTT) zur Verfügung gestellt werden. Feldgeräte, die ebenfalls zur Bereitstellung von Energieinformationen geeignet sind, können z. B. über OPC UA- oder MQTT-Schnittstellen eine direkte Kommunikation mit der Energiemanagement-Applikation aufbauen [13; 14]. Während MQTT besonders auf eine Datenübertragung mit geringer Netzwerkbandbreite und unbeständiger Netzwerkkonnektivität (z. B. bei Funkverbindungen mit geringer Netzwerkqualität) ausgerichtet ist, fokussiert OPC UA eine standardisierte Beschreibung der Semantik mittels sogenannter OPC UA Companion Spezifikationen [15].

3 Heterogene Beschreibung der Energieinformationen

Der Stand der Technik zeigt, dass es für verschiedene industrielle Netzwerkprotokolle jeweils eigene Energieprofile gibt, die alle für den Zweck der Energieinformationsübertragung entwickelt wurden. Da die benannten Netzwerkprotokolle jeweils eigene Protokollstrukturen und eigene Semantiken für die Beschreibung der übertragenen Informationen aufweisen, sind die Semantiken der Energieinformationen der Energieprofile dementsprechend vielfältig. Um Gemeinsamkeiten und Unterschiede von Energieprofilen im Hinblick auf die übertragenen Energieinformationen aufzeigen zu können, wurden die Energieprofile PROFlenergy, Sercos Energy und CIP Energy miteinander verglichen. Für den Vergleich wurden die Eigenschaften

der Energieprofile sowohl für Energie-Monitoring-Funktionalitäten als auch für Lastmanagement-Funktionalitäten in tabellarischer Form gegenübergestellt.

3.1 Vergleich von Energie-Monitoring-Funktionalitäten bestehender Energieprofile

Beim Vergleich der Energie-Monitoring-Funktionalitäten der Energieprofile PROFenergy, Serco Energy und CIP Energy lassen sich folgende Unterschiede und Gemeinsamkeiten aufzeigen (siehe Tabelle 1):

Tabelle 1: Vergleich der Energie-Monitoring-Funktionalitäten der Energieprofile

Energie-Monitoring	Energieprofile		
Eigenschaften der Energieprofile	PROFenergy	Sercos Energy	CIP Energy
Unterstützung elektrischer Messgrößen			
Unterstützung nicht-elektrischer Messgrößen	X	X	
Anzahl elektrischer Messgrößen	40	2	39
Anzahl nicht-elektrischer Messgrößen	0	0	12
Identifikation der Messgrößen über	spezifische MeasurementID	spezifische Parameternummer	spezifische Attribut-ID
Bezeichnung eines Spezifikationsmerkmals am Beispiel Energiezähler	Active Energy Import/ Export	Consumed / Supplied Energy Odometer	Real Energy Consumed / Generated Odometer
Datentypen der Messgrößen	Float32, Float64, signed integer	signed integer	Real, Struct (aus unsigned/ signed integer)
Abbildung der Messgenauigkeit	✓ Normbasiert (EN 50470-3, IEC 61557-12) oder relative Abweichung bezogen auf Messbereichsendwert/ aktuellen Messwert	✗ Keine Angaben zu Messgenauigkeiten in der Energieprofilspezifikation	✓ Relative Abweichung bezogen auf Messbereichsendwert/ aktuellen Messwert

Erste Unterschiede und Gemeinsamkeiten zeigen sich bei der Betrachtung der Energieprofile beim Umgang mit Messgrößen. Zwar können alle betrachteten Energieprofile elektrische Messgrößen (z. B. Strom in A, Spannung in V) abbilden, das Monitoring von nichtelektrischen

Messgrößen (z. B. Wärmemenge in J , Volumenstrom in $\frac{l}{min}$) ist zum Vergleichszeitpunkt allerdings nur mit CIP Energy möglich. Zusätzlich lassen sich weitere Unterschiede bei der Betrachtung der Anzahl an verfügbaren Messgrößen feststellen. So hat Sercos Energy mit zwei elektrischen Messgrößen das kleinste Angebot an Messgrößen. Wird das Energieprofil PRO-Flenergy eingesetzt, kann der Anwender zwischen 40 elektrischen Messgrößen wählen. CIP Energy hat zwar mit 39 elektrischen Messgrößen eine etwas kleinere Anzahl, kann dafür zusätzlich aber 12 verschiedene nicht-elektrische Messgrößen darstellen.

Die Messgrößen werden bei allen Profilen über spezifische Nummern aufgerufen. Bei PRO-Flenergy erfolgt der Aufruf einer Messgröße über eine *MeasurementID*. CIP Energy verwendet für den Aufruf *Attribute-IDs* und Sercos Energy nutzt für die Aufrufe Parameternummern.

Weitere Unterschiede sind bei den Bezeichnungen der Spezifikationsmerkmale erkennbar. Am Beispiel von Energiezählern für verbrauchte und erzeugte Energie ist ersichtlich, dass alle Profile voneinander abweichende Bedeutungen und dementsprechend verschiedene Semantiken für die gleichen Spezifikationsmerkmale verwenden. CIP Energy bezeichnet Energiezähler als *Real Energy Consumed / Generated Odometer*. Sercos Energy nutzt hierfür mit *Consumed / Supplied Energy Odometer* eine ähnliche, aber dennoch abweichende Bezeichnung für Energiezähler. Bei PROFlenergy hingegen werden die Energiezähler mit *Active Energy Import / Export* bezeichnet.

Auch bei den eingesetzten Datentypen für Messgrößen unterscheiden sich die drei Energieprofile. PROFlenergy verwendet mit *Float32*, *Float64* und *signed integer* gleich mehrere Datentypen, um sowohl kleinste Messwerte als auch große Messwerte ohne eine Überlaufproblematik des Datentyps darstellen zu können. CIP Energy setzt als Datentyp *REAL* und *Struct* (Kombination aus *unsigned* und *signed integer*) ein. Sercos Energy beschreibt die Messwerte mit dem Datentyp *signed integer*.

Größere Unterschiede sind auch bei der Abbildung von Messgenauigkeiten erkennbar. PRO-Flenergy setzt bei der Genauigkeit von Messwerten auf die Normen EN 50470-3 und IEC 62557-12. Zusätzlich kann bei PROFlenergy eine Messgenauigkeit über die relative Abweichung bezogen auf den Messbereichsendwert bzw. aktuellen Messwert dargestellt werden. CIP Energy verweist zwar nicht auf normbasierte Messgenauigkeiten, jedoch kann eine Messgenauigkeit wie bei PROFlenergy über die relative Abweichung bezogen auf den Messbereichsendwert bzw. aktuellen Messwert dargestellt werden. In der Spezifikation zu Sercos Energy werden keine Angaben zu Messgenauigkeiten gemacht.

3.2 Vergleich von Lastmanagement-Funktionalitäten bestehender Energieprofile

Die Lastmanagement-Funktionalitäten der betrachteten Energieprofile weisen ebenfalls Gemeinsamkeiten und Unterschiede auf (siehe Tabelle 2)

Tabelle 2: Vergleich der Lastmanagement-Funktionalitäten der Energieprofile

Lastmanagement-Funktionalitäten	Energieprofil		
	PROFenergy	Sercos Energy	CIP Energy
Eigenschaften der Energieprofile			
Bereitstellung der Funktionalität - Standby-Management -			
Bezeichnung der Spezifikationsmerkmale (Beispielbezeichnung: Zeitspanne bis Reaktivierung)	time_to_operate	Time To Energy Mode Operational	Resume Time
Bereitstellung der Funktionalität - Wake On LAN -		X	
Wechselmöglichkeit zwischen Energiesparmodi und Wake on LAN	X	-	

Alle betrachteten Energieprofile bieten die Möglichkeit über ein Standby-Management einzelne Geräte der Feldebene mit gezielten Anweisungen in einen Standby-Zustand zu versetzen. Somit können die Geräte beispielsweise in Pausen- oder Maschinenstillstandszeiten energieeffizienter betrieben werden, da der Energiebedarf durch den Standby-Zustand reduziert ist. Auch bei den Lastmanagement-Funktionalitäten sind für die Beschreibung der gleichen Merkmale Unterschiede bei den Bezeichnungen der Spezifikationsmerkmale erkennbar. Als Beispiel dient die „Zeitspanne bis Reaktivierung eines Gerätes“. Um ein besseres Verständnis für diesen Zeitraum zu haben, ist in Abbildung 2 die Energiebedarfskurve eines Gerätes der Feldebene (z. B. Frequenzumrichter) mit einer Lastmanagementmöglichkeit (Gerät wird zeitlich begrenzt in einen Energiesparmodus versetzt) dargestellt.

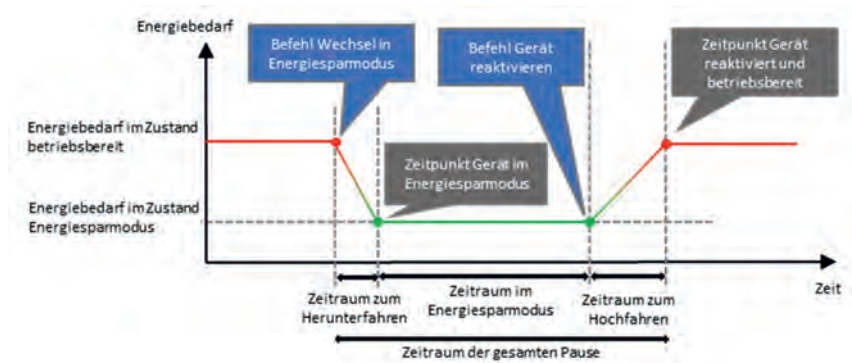


Bild 2: Energiebedarfskurve eines Feldgerätes mit Lastmanagementmöglichkeit

Die „Zeitspanne zur Reaktivierung des Gerätes“ oder auch „Zeitraum zum Hochfahren“ wird bei allen drei Energieprofilen unterschiedlich bezeichnet. PROFInergy nennt diesen Zeitraum *time_to_operate*. Bei Sercos Energy hingegen wird dieser Zeitraum als *Time To Energy Mode Operational* bezeichnet. CIP Energy wiederum benennt diesen Zeitraum *Resume Time*. Alle Energieprofile nutzen dementsprechend für das gleiche Merkmal jeweils eine eigene Semantik.

Darüber hinaus sind weitere Unterschiede für die Funktionalität -Wake on LAN- (WOL) ersichtlich. Mit WOL ist es möglich die Geräte der Feldebene aus einem Schlafzustand wieder zu wecken, sodass diese wieder in einen betriebsbereiten Zustand versetzt werden können. Geräte, die sich in einem Schlafzustand befinden, können keine Daten versenden und haben nur einen sehr geringen Energiebedarf. Sercos Energy stellt als einziges Energieprofil diese Funktionalität nicht bereit. PROFInergy und CIP Energy bieten hingegen die Möglichkeit WOL zu verwenden. Jedoch können Geräte, die PROFInergy unterstützen, nur direkt aus dem betriebsbereiten Zustand in einen Schlafmodus versetzt werden. Der Wechsel von einem Energiesparmodus in den Schlafmodus ist nicht möglich. CIP Energy bietet zusätzlich die Möglichkeit ein Gerät aus einem Energiesparmodus in einen Schlafzustand zu versetzen.

3.3 Zwischenfazit

Sowohl bei den Energie-Monitoring- als auch bei den Lastmanagement-Funktionalitäten ist ersichtlich, dass die verglichenen Energieprofile Unterschiede hinsichtlich der Funktionen und Semantik aufweisen.

Solange die Energieinformationen in einem tEnMS mit nur einem Energieprofil übertragen werden, sind die identifizierten Unterschiede für das tEnMS unbedeutend, da alle Energieinformationen durch das verwendete Profil einheitlich beschrieben werden.

Sollen allerdings mehr als nur ein spezifisches Energieprofil in ein tEnMS integriert werden, entsteht das Problem, dass die Energieinformationen durch die Nutzung von mehreren Profilen nicht mehr einheitlich beschrieben werden können. Nur durch einen erhöhten Engineering-Aufwand können die Energieinformationen der einzelnen Energieprofile vereinheitlicht werden.

In [16] wurde bereits aufgezeigt, dass der Einsatz eines vereinheitlichenden Energieinformationsmodells den Engineering-Aufwand reduzieren kann. Mit Hilfe dieses Informationsmodells können Energieinformationen von unterschiedlichen Quellen der Feldebene vereinheitlicht werden, ohne dass ein zusätzliches Mapping der Energieinformationen erforderlich ist. Wo ein derartiges Energieinformationsmodell zu verorten ist und welche Funktionen dieses Informationsmodell erfüllen müsste, wird im nächsten Kapitel aufgezeigt.

4 Einheitliches Energieinformationsmodell

Auf der Grundlage des vorangehenden Vergleichs der Energieprofile kann ein universelles Energieinformationsmodell (kurz UEIM) konzeptioniert werden. Abbildung 3 zeigt, wie eine standardisierte Bereitstellung von Energieinformationen über ein UEIM in der Automatisierungshierarchie durchgeführt werden kann.

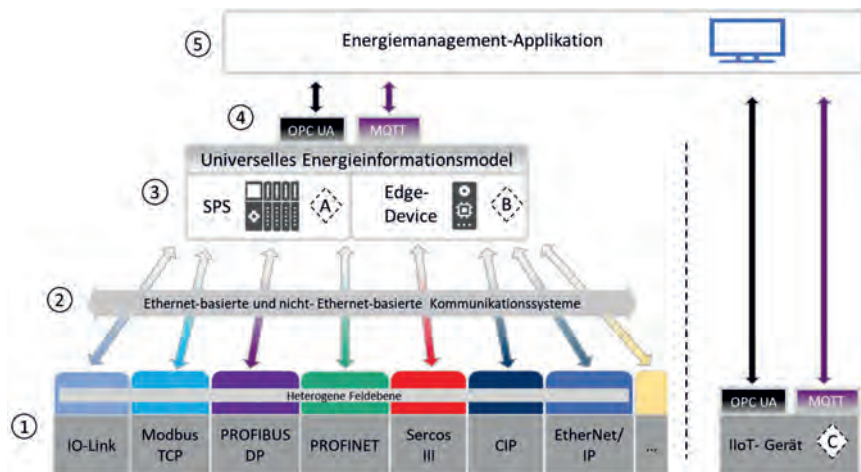


Bild 3: Einordnung des universellen Energieinformationsmodells im Kontext der Automatisierungshierarchie

Aus der heterogenen Feldebene (1) können die Energieinformationen über unterschiedliche Ethernet- oder nicht-Ethernet-basierte Kommunikationssysteme (2) übertragen werden. Die

Semantik der Energieinformationen ist nicht einheitlich und möglicherweise an verschiedene Energieprofile geknüpft. Über das jeweilige Kommunikationssystem (2) der Feldgeräte können Energieinformationen mit Edge-Devices oder Speicherprogrammierbaren Steuerungen (SPS) (3) bidirektional ausgetauscht werden. Bei der Nutzung einer SPS (Fall A) oder eines Edge Devices (Fall B) werden die Energieinformationen, wie in Kapitel 2.2.5 beschrieben, aufbereitet und in der standardisierten Semantik des UEIM (4) über Kommunikationsschnittstellen, wie beispielsweise OPC UA- oder MQTT, mit der Energiemanagement-Applikation (5) ausgetauscht. Bei Nutzung eines IIoT-Gerätes (Fall C) kann eine direkte, z. B. OPC UA- oder MQTT-basierte, Bereitstellung der Energieinformationen gegenüber der Energiemanagement-Applikation (5) in der standardisierten Semantik des UEIM erfolgen. Mit Hilfe des UEIM sollen die Geräte (A), (B), (C) mit einer semantisch standardisierten Kommunikationsschnittstelle ausgestattet werden. Durch diese Schnittstelle kann der Engineering-Aufwand für die Integration eines tEnMS in bestehende Automatisierungssysteme reduziert und die Energieinformationsübertragung mit mehreren Energieprofilen vereinfacht werden.

5 Zusammenfassung und Ausblick

Durch einen Vergleich von ausgewählten Energieprofilen konnte in diesem Beitrag aufgezeigt werden, dass Unterschiede, aber auch einige Gemeinsamkeiten bei der semantischen Beschreibung von Energieinformationen in den Energieprofilen vorliegen. Darüber hinaus wurde das Konzept eines universellen Energieinformationsmodells vorgestellt, mit dem Energieinformationen semantisch eindeutig beschrieben werden können, sodass es irrelevant ist, mit welchem Energieprofil die Energieinformationen übertragen werden und wie diese im Kommunikationsnetzwerk weiter gereicht werden sollen.

Zukünftig müssen die Energieinformationen von verschiedenen Energieprofilen nicht mehr einzeln auf das Protokoll eines übergeordneten Kommunikationssystems übersetzt (gemappt) werden, denn mit Hilfe des Energieinformationsmodells können Energieinformationen einfacher zusammengeführt und in ein übergeordnetes Kommunikationssystem integriert werden. Somit lässt sich nicht nur der Engineering- Aufwand reduzieren, sondern auch Fehler beim Mapping der Energieinformationen vermeiden. Das Energieinformationsmodell wird in der zukünftigen Arbeit ausgearbeitet und realisiert.

Die Inhalte dieses Beitrags wurden im Rahmen des IGF Projekts 21329N (IoT_EnRG), das vom Bundesministerium für Wirtschaft und Klimaschutz (BMWK) auf Grundlage eines Beschlusses des Deutschen Bundestages gefördert wird, erarbeitet.

Literatur

- [1] ISO 50001:2018-08. Energy management systems -Requirements with guidance for use, 2018.
- [2] K. Kostková, L. Omelina, P. Kyčina, P. Jamrich: *An introduction to load management*. Electric Power Systems Research, Vol. 95, S. 184–191, 2013. <https://www.sciencedirect.com/science/article/abs/pii/S037877961200288X> [Abruf: 2022-04-07]
- [3] PROFIBUS and PROFINET International: *Common Application Profile PROFIenergy: Technical Specification for PROFINET*. V1.3. <https://de.profibus.com/downloads/profienergy> [Abruf: 2022-04-07]
- [4] PROFIBUS and PROFINET International: *OPC UA for Energy Management*. RC 1.00. <https://www.profibus.com/download/opc-ua-for-energy-management-companion-specification> [Abruf: 2022-04-07]
- [5] Sercos International e.V.: *Sercos III 1.3.2 Subprofile Sercos Energy*. 1.3.2-1.1, 2018.
- [6] Sercos International e.V.: *OPC UA for Sercos Devices*. <https://opcfoundation.org/developer-tools/specifications-opc-ua-information-models/opc-unified-architecture-for-sercos-devices/> [Abruf: 2022-04-07]
- [7] J. Schlechtendahl: *Whitepaper sercos Energy*, 2012. <https://www.sercos.de/news-events/newsdetail/sercos-energy-spezifikation-des-energieprofils-fuer-sercos-iii-verfuegbar/> [Abruf: 2021-04-22]
- [8] ODVA, Inc.: The CIP networks library Volume 1: Common Industrial Protocol.
- [9] ODVA: *OPC UA Companion Specification to be developed for CIP*. <https://www.odva.org/news/opc-ua-companion-specification-to-be-developed-for-cip/> [Abruf: 2022-01-26]
- [10] M. Shamshiri, C.K. Gan, K.A. Baharin, M. Arif Mohd Azman: *IoT-based electricity energy monitoring system at Universiti Teknikal Malaysia Melaka*. Bulletin of Electrical Engineering and Informatics, Vol. 8 (2), S. 683–689, 2019. <https://beei.org/index.php/EEI/article/view/1281> [Abruf: 2022-04-07]
- [11] Siemens AG: *IO-Link - continuous communication through to the last metre*. Nürnberg, 2020. <https://new.siemens.com/global/en/products/automation/industrial-communication/io-link.html> [Abruf: 2022-04-07]

- [12] F. Marimon, M. Casadesús: *Reasons to Adopt ISO 50001 Energy Management System*. Sustainability, Vol. 9 (10), 2017. <https://www.mdpi.com/2071-1050/9/10/1740> [Abruf: 2022-04-07]
- [13] G. Leßmann, S. Schriegel: *MQTT, OPC UA und PROFINET für IIoT*, 2021. <https://opendata.uni-halle.de/handle/1981185920/41506> [Abruf: 2022-04-07]
- [14] A. Veichtlbauer, M. Ortmayr, T. Heistracher: *OPC UA integration for field devices*. In: *IEEE 10th International Conference on Industrial Informatics*, S. 419–424, 2012. <https://ieeexplore.ieee.org/abstract/document/8104808> [Abruf: 2022-04-07]
- [15] S. Profanter, A. Tekat, K. Dorofeev, M. Rickert, A. Knoll: *OPC UA versus ROS, DDS, and MQTT: Performance Evaluation of Industry 4.0 Protocols*. In: *2019 IEEE International Conference on Industrial Technology (ICIT): Melbourne Convention and Exhibition Centre, Melbourne, Australia, 13-15 February 2019: proceedings*. Piscataway, NJ: IEEE, S. 955–962, 2019. <https://ieeexplore.ieee.org/document/8755050> [Abruf: 2022-04-07]
- [16] L.-T. Reiche, M. Runge, K.-H. Niemann, A. Fay: *Communication of energy data in automation systems*. In: *Communication in automation: 17th IEEE International Workshop on Factory Communication Systems 2021: WFCs 2021: 9-11 June 2021, Linz, Austria*. Piscataway, NJ: IEEE, S. 45–48, 2021. <https://serwiss.bib.hs-hannover.de/frontdoor/index/index/year/2021/docId/2089> [Abruf: 2022-04-07]
- [17] A. Würger: *Automatisierte Generierung von Energiemanagementfunktionen auf der Basis des Profinet-Energieprofils*. Hamburg, 2020. <https://openhsu.ub.hsu-hh.de/handle/10.24405/9095> [Abruf: 2022-04-07]
- [18] PROFIBUS Nutzerorganisation e.V.: *Informationsmodelle: PI-Strategie für Industrie 4.0*, 2022. <https://www.profibus.com/download/information-models> [Abruf: 2022-04-07]
- [19] PROFIBUS and PROFINET International: *PROFINET Systembeschreibung: Technologie und Anwendung*, 2014. <https://www.profibus.com/index.php?eID=dump-File&t=f&f=51713&token=5e6746cb84a7421d187681a0d9bd545388cb2a5e> [Abruf: 2022-04-07]
- [20] ODVA, Inc.: *Common Industrial Protocol*. 3.31. Aufl., 2021.
- [21] DIN CLC/TR 62541-1. *OPC Unified Architecture - Teil 1: Übersicht und Konzepte (IEC/TR 62541-1:2010)*; Deutsche Fassung CLC/TR 62541-1:2010.

- [22] A. Würger, K.-H. Niemann: Integriertes Energiemanagement in Industrie 4.0-Anlagen - Potentiale, Trends und Ausblick. In: Hochschule Hannover (Hrsg.): Anwendungsorientierte Forschung für die digitale Transformation von KMU: Schriften des Forschungsclusters Industrie 4.0: Hochschule Hannover, S. 66–78, 2020. <https://serwiss.bib.hs-hannover.de/frontdoor/index/index/docId/1738> [Abruf: 2022-04-07]
- [23] E. O'Driscoll, G.E. O'Donnell: *Industrial power and energy metering – a state-of-the-art review*. Journal of Cleaner Production, Vol. 41, S. 53–64, 2013. <https://www.sciencedirect.com/science/article/pii/S0959652612005173>

Nachhaltigkeit in der Wasserstoffwirtschaft – Eine Methode zur Bewertung von Wasserstoff-Wertschöpfungsketten

J.-P. Beck, A.Fay, Helmut-Schmidt-Universität, Hamburg

Kurzfassung

Wasserstoff ist ein zentraler Teil der Dekarbonisierungsstrategie und damit auch wichtiger Bestandteil für eine erfolgreiche Energiewende. Daher ist es notwendig, die komplette Wasserstoff-Wertschöpfungskette zu betrachten und nicht nur Teile davon. Dieser Beitrag stellt ein Schema vor, um verschiedene Wasserstoff-Wertschöpfungsketten ableiten zu können und mit Hilfe einer Methode diese Wertschöpfungsketten anhand verschiedener Kriterien zu bewerten. In diesem Beitrag werden exemplarisch drei Wasserstoff-Wertschöpfungsketten anhand ihrer Investitionskosten und Betriebskosten miteinander verglichen.

Abstract

Hydrogen is a central part of the decarbonization strategy and thus also an important component for a successful energy transition. Therefore, it is necessary to consider the complete hydrogen value chain and not only parts of it. This paper presents a scheme to derive different hydrogen value chains and to use a method to evaluate these value chains based on different criteria. In this paper, three hydrogen value chains are compared based on their investment costs and operating costs.

1. Motivation

Die Energiewende mit ihren Anstrengungen beim Ausbau der erneuerbaren Energien und im Bereich der Energieeffizienz bildet die Grundlage für eine saubere, sichere und bezahlbare Energieversorgung. Die Nutzung von gasförmigen und flüssigen Energieträgern wird auch langfristig in allen Industrieländern - damit auch in Deutschland - relevant sein. Eine zentrale Rolle für die Weiterentwicklung und Vollendung der Energiewende wird dem Wasserstoff zugesprochen.

Wasserstoff ist ein vielseitig einsetzbarer Energieträger, ein Energiespeicher, ein wesentliches Element zur Sektorkopplung sowie bereits seit langem ein Grundstoff für verschiedene chemische und industrielle Prozesse. Mit Hilfe von Wasserstoff lassen sich langfristig bestimmte industrielle CO₂-Quellen dekarbonisieren. [1]

Die Nationale Wasserstoffstrategie [1] sieht Wasserstoff damit als zentralen Bestandteil der Dekarbonisierungsstrategie, und daher muss die gesamte Wertschöpfung von Wasserstoff in den Blick genommen werden. Dazu gehören Technologien der Erzeugung, der Speicherung, Infrastruktur und der Endanwendung einschließlich Logistik. [1]

Dabei ist zu beachten, dass eine Wasserstoffwirtschaft nur schrittweise aufgebaut werden kann und damit auch eine Umstellung von Erdgas auf Wasserstoff nur schrittweise umgesetzt werden kann. Dies bedingt aber auch, dass neue Strukturen geschaffen werden müssen und mittelfristig die Strukturen von Wasserstoff und Erdgas parallel existieren. Sich entwickelnde Wasserstoffteilnetze müssen mit vielen Unsicherheiten geplant werden, müssen aber auch langfristig Bedeutung haben und sind damit besonders zu Beginn überdimensioniert und unwirtschaftlich. Außerdem wird nicht überall ein zusätzliches Netz für Wasserstoff in unserer bestehenden Infrastruktur kurzfristig umzusetzen sein und es wird Übergangslösungen für den Transport, aber auch die Erzeugung und Speicherung geben müssen. Die Komplexität dieser Aufgabe der Gestaltung und Dimensionierung von Wasserstoff-Wertschöpfungsketten ist sehr hoch. Für den Bereich der Mobilität als Endanwendung gibt es bereits eine Vielzahl an Forschungsarbeiten, die sich mit kostenoptimalen Wertschöpfungsketten beschäftigt haben. Wie eingangs erwähnt, wird Wasserstoff aber in viel mehr Bereichen eine Rolle spielen, und die Anforderungen an die Wertschöpfungskette sind damit zum Teil abweichend.

2. Stand der Forschung

Yang und Ogden [2] haben eine Methode zum Vergleich von unterschiedlichen Transportmöglichkeiten von gasförmigem und flüssigem Wasserstoff entwickelt. Sie haben dabei Pipelines, Tube-Trailer und Flüssigtankwagen verglichen und haben aufgezeigt, dass es keine ideale Lösung für das gesamte System gibt. Viele nachfolgende Arbeiten bauen auf diesem Ansatz auf oder vergleichen eigene Ergebnisse mit den dort erzielten Ergebnissen. Mintz et al. [3] haben ein Tool auf Excel-Basis entwickelt, um die Kosten einer Wasserstoffversorgung anhand verschiedener Input-Parameter zu berechnen. Die Ansätze [2] und [3] sehen die Mobilität als Endanwendung und haben die Wasserstoffproduktion nur als Input angenommen und nicht mit abgebildet. Reuß et al. [4] haben eine Methode zur Berechnung eines räumlich aufgelösten Autobahn-Routing-Modells für Deutschland für den Transport von Wasserstoff zur Versorgung von 9683 Tankstellen entwickelt. Sie kommen zu dem Schluss, dass der gasförmige Transport von Wasserstoff mit Hilfe von Tube-Trailern für Entfernungen kleiner als 130 km wirtschaftlicher ist als der Transport in flüssigem Zustand. In einem anderen Beitrag analysieren Reuß et al. [5] flüssige organische Wasserstoffträger (LOHC) und zeigen ihren Einfluss auf eine zukünftige Wasserstoff-Mobilität mit der

Besonderheit, dass sie auch langfristige Speicherung berücksichtigen. Sie bauen dabei auf bestehenden Ansätzen in der Literatur auf. Der ökologische Vergleich von Erzeugungstechnologien für den Einsatz von Wasserstoff im Verkehr wird in der Arbeit von [6] durchgeführt. In der Arbeit von [7] wird ein Energiesystemmodell mit einem geographischen Informationssystem gekoppelt. Ziel ist es, eine Bewertung für den Aufbau einer Wasserstoffinfrastruktur zu geben. Weitere Arbeiten sind die von Almansoori et al. [8], Agnolucci et al. [9] sowie Samssatli et. al [10]. Diese Arbeiten berechnen mögliche Wasserstoffinfrastrukturen für das Vereinigte Königreich. Hier werden verschiedene Optimierungsansätze zur Problemlösung genutzt.

3. Wasserstoff-Wertschöpfung

In Bild 1 sind zwei vereinfachte Wasserstoff-Wertschöpfungsketten gezeigt. In beiden wird Wasserstoff mit Hilfe einer Elektrolyse im Ausland hergestellt und dann im oberen Fall per Schiff nach Deutschland transportiert. Der Transport vom Hafen zur Endanwendung geschieht per LKW. Da es sich hier um eine vereinfachte Darstellung handelt, ist beispielsweise die Konditionierung nicht dargestellt. Im unteren Fall wird der Wasserstoff mittels einer Transmission-Pipeline nach Deutschland transportiert, hierzulande gespeichert und anschließend mit Hilfe einer Distribution-Pipeline zur Endanwendung transportiert. Neben den unterschiedlichen Transportoptionen Schiff, LKW, Pipeline würde sich in diesem Fall auch der Aggregatzustand des Wasserstoffs unterscheiden. In flüssiger Form lässt sich Wasserstoff in deutlich höheren Mengen pro Container transportieren. In industriellen Anwendungen wird Wasserstoff in gasförmigen Zustand benötigt, daher muss der Wasserstoff wieder in den gasförmigen Zustand überführt werden.

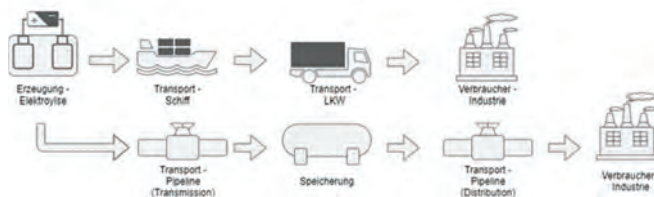


Bild 1: Beispielhafte Wasserstoff-Wertschöpfungsketten in vereinfachter Form

Erzeugung

Wasserstoff kann mit Hilfe von verschiedenen Verfahren hergestellt werden. Die Verfahren lassen sich dabei in thermo-chemische, elektro-chemische und biologische Verfahren

unterteilen. Außerdem gibt es noch die sog. thermo-chemischen Zyklen und die Erzeugung des Wasserstoffs als Nebenprodukt der chemischen Industrie.

Zu den thermo-chemischen Verfahren zählen Reforming, Pyrolyse und die Vergasung. Für die thermo-chemischen Verfahren kommen meist Erdgas, Kohle oder Biomasse als Primärenergieträger zum Einsatz. Diese Verfahren zählen daher - mit Ausnahme von Biomasse - nicht zu den grünen Wasserstoffherstellungsverfahren.

Die elektro-chemischen Verfahren sind die Elektrolyseverfahren mit der alkalischen Elektrolyse, der Proton-Austausch-Membran Elektrolyse und der Hochtemperatur-Elektrolyse. Diese Verfahren können grünen Wasserstoff herstellen und sind daher auch langfristig für die Wasserstoffproduktion präferiert. Diese Verfahren benötigen aber Strom als Reaktionsenergie und stellen somit nur grünen Wasserstoff her, wenn auch der Strom aus erneuerbaren Energien erzeugt wird.

Zu den thermo-chemischen Zyklen zählen bspw. der Schwefel-Iod-Kreislauf oder der Kupfer-Chlor-Kreislauf.

Zu den biologischen Verfahren zählen bspw. die Fermentation und die Biophotolyse.

Außerdem entsteht Wasserstoff in verschiedenen Produktionsverfahren als Nebenprodukt, bspw. bei der Chloralkali-Elektrolyse oder der Ethylen- oder Acetylenproduktion.

Distribution

Wasserstoff kann entweder mittels physikalischer oder stofflicher Speicherung per Schiff, Schiene oder LKW transportiert werden oder mit Hilfe einer Pipeline. Bei den physikalischen Speichern handelt es sich um Druckgasspeicher oder Flüssiggasspeicher. Diese können sowohl stationär oder auch als transportierbare Speicher realisiert werden. Bei den stofflichen Speichern wird der Wasserstoff in Feststoffen oder Flüssigkeiten gespeichert. Für die Speicherung bzw. den Transport ist in der Regel eine Konditionierung notwendig. Dabei handelt es sich meist um eine Kompression, Verflüssigung oder Regasifizierung.

Endanwendung

Die Anwendungsmöglichkeiten von Wasserstoff finden sich in der Energiewirtschaft, der Industrie, dem Verkehr und anderen Sektoren. Wasserstoff ist in der Lage, viele auf fossilen Energieträgern basierenden Prozesse zu dekarbonisieren. Die Endanwendung stellt unterschiedliche Anforderungen, bspw. in Bezug auf die Reinheit des Wasserstoffs oder die benötigte Menge oder die Gleichmäßigkeit der Versorgung, an die Wasserstoff-Wertschöpfungskette.

4. Vorstellung der Methode

Wie bereits eingangs in der Motivation beschrieben, sieht die Nationale Wasserstoffstrategie [1] den Bedarf der Betrachtung der gesamten Wasserstoff-Wertschöpfungskette. Der Stand der Forschung zeigt, dass es bereits einige Arbeiten zu diesem Thema gibt. Sehr wenige beschäftigen sich aber mit ganzen Wertschöpfungsketten, und niemand hat bisher eine Methode für die Bewertung von beliebigen Wasserstoff-Wertschöpfungsketten entwickelt. Falls die analysierten Wertschöpfungsketten nicht den eigenen Anforderungen entsprechen bzw. nicht alle gewünschten Technologien und Prozesse berücksichtigt werden, muss der Anwender immer noch selbst viel Arbeit in die Analyse investieren.

Um dieses Problem zu adressieren, wird von den Autoren eine Methode zur Bewertung von unterschiedlichen Wasserstoff-Wertschöpfungsketten entwickelt. Der Anwender kann mit Hilfe der Methode die für seine Anforderungen optimale Wertschöpfungskette im Hinblick auf die Investitionskosten, Betriebskosten, den resultierenden Wasserstoffpreis, die Effizienz der Wertschöpfungskette, sowie die CO₂-Emissionen bestimmen.

Da sich innerhalb der Wasserstoff-Wertschöpfungskette die Prozesse in vielfältiger Kombination anordnen lassen, entsteht eine Vielzahl von möglichen verschiedenen Wasserstoff-Wertschöpfungsketten. Zur Wahrung der Übersichtlichkeit ist daher eine grafische Darstellung sinnvoll.

In der Literatur finden sich bereits einige Abbildungen und Schemata, um möglichst viele Wasserstoff-Wertschöpfungsketten abbilden zu können, beispielsweise in der Arbeit von Bolat und Thiel [11]. Die Abbildungen sind jedoch meist schwer zu lesen. In diesem Beitrag wird daher ein Schema vorgestellt, welches auf den ersten Blick deutlich reduziert wirkt, aber mit dem sich eine Vielzahl von Wasserstoff-Wertschöpfungsketten abbilden lassen. Der Fokus dieses Schemas liegt auf gängigen Erzeugungstechnologien – wobei jede weitere Technologie erweiterbar ist – und der Nutzung von Wasserstoff in chemischen und industriellen Prozessen. Mit einigen Anpassungen wäre es aber ebenso für Wasserstoff-Mobilitätsanwendungen einsetzbar. In diesem Fall würde sich das Schema besonders im Bereich IV: Endanwendung unterscheiden und müsste dahingehend erweitert werden.



Bild 2: Schema zur Ableitung einer Vielzahl von Wasserstoff-Wertschöpfungsketten

Das Schema in Bild 2 wurde entwickelt, um einfach eine Vielzahl von Wasserstoff-Wertschöpfungsketten daraus ableiten zu können und diese einheitlich abzubilden. Das Schema besteht aus 4 Bereichen. Im Bereich I: Erzeugung ist die Erzeugungstechnologie auswählbar und der Zustand „E“ des Wasserstoffs direkt nach der Erzeugung ist ersichtlich. In diesem Zustand kann der Wasserstoff optional gespeichert oder aufbereitet werden. Im nächsten Bereich II: Transmission wird abgebildet, wie der Wasserstoff über lange Strecken – auch Länderübergreifend – transportiert wird. Zunächst wird der Wasserstoff im Zustand „E-T“ vom Standort der Erzeugung zum Standort des Beginns der Transmission per LKW, Zug oder Pipeline transportiert. Am Standort des Beginns der Transmission liegt der Wasserstoff im Zustand „T“ vor und kann optional gespeichert oder aufbereitet werden. Anschließend wird er im Zustand „T-D“ per Schiff, Zug oder Pipeline über weite Strecken transportiert. Der anschließende Bereich III: Distribution beschreibt die Verteilung des Wasserstoffs auf regionaler Ebene. Im Bereich III: Distribution liegt der Wasserstoff dann am Standort des Beginns der Distribution im Zustand „D“ vor. Hier kann der Wasserstoff wieder optional gespeichert oder erst jetzt aufbereitet werden, bevor er weiter verteilt wird. Im Zustand „D-End“ wird der Wasserstoff per LKW, Zug oder Pipeline zum Standort der Endanwendung transportiert. Im Bereich IV: Endanwendung liegt der Wasserstoff dann im Zustand „End“ vor und kann optional gespeichert oder aufbereitet werden, bevor er schlussendlich im Prozess der Endanwendung verbraucht wird. Das Schema bietet die Möglichkeit, durch die Auswahl

der optionalen Punkte zur Speicherung und Aufbereitung sich an individuelle Wasserstoff-Wertschöpfungsketten anzupassen. Während im Zustand E der Wasserstoff fast ausschließlich in gasförmiger Form vorliegt, kann er in den Zuständen E-T, T, T-D, D-End und End ebenfalls in flüssiger Form vorliegen. Aus dem Schema ist durch farbliche Hervorhebung der Aggregatzustand erkennbar. Falls der Wasserstoff also beispielsweise in Zustand E im gasförmigen Zustand vorliegt und in Zustand E-T im flüssigen Zustand ist klar, dass dazwischen eine zusätzliche Konditionierung stattgefunden haben muss, in diesem Fall eine Verflüssigung. Da die Konditionierung theoretisch zwischen allen jeweils benachbarten Zuständen möglich wäre, wurde im Schema darauf verzichtet, die Konditionierung überall optional mit aufzugreifen. Durch die farbliche Hervorhebung lässt sich aber direkt erkennen, dass eine Konditionierung erforderlich ist. Die Zustände E, T, D und End lassen sich festen Standorten zuordnen, während es sich bei den Zuständen E-T, T-D und D-End um Transportzustände handelt. Mit dem Schema lassen sich aber auch regionale/lokale Wasserstoff-Wertschöpfungsketten abbilden. In dem Fall fällt der Bereich II: Transmissionen weg und der Zustand E entspricht dem Zustand D. Um die korrekte Anwendung des Schemas zu verdeutlichen, wird in Bild 3 das Schema genutzt, um die Wertschöpfungskette aus Bild 1 (oben) abzubilden – wobei die jetzt gezeigte Wertschöpfungskette etwas komplexer ist.



Bild 3: Abbildung der Wasserstoff-Wertschöpfungskette aus Bild 1 oben mit Hilfe des Schemas aus Bild 2

In dieser Wertschöpfungskette wird der Wasserstoff im Ausland mit Hilfe einer AEL produziert und einschließlich per Pipeline (E-T) im gasförmigen Zustand zum Hafen transportiert. Dort wird der Wasserstoff verflüssigt und in flüssigem Zustand (T-D) nach Deutschland transportiert und dort vom Anlegehafen per Flüssigtank-LKW (D-End) zur Endanwendung transportiert. Dort wird der Wasserstoff regasifiziert, anschließend zwischengespeichert und vor seiner Verwendung aufbereitet. Ebenfalls denkbar wäre, dass der Wasserstoff bereits am Anlegehafen (D) regasifiziert und anschließend per LKW oder Pipeline zur Endanwendung

transportiert wird. In Bild 3 würde dann entsprechend bereits in D die Regasifizierung stattfinden und per LKW oder Pipeline (D-End) in gasförmigen Zustand transportiert werden.

5. Anwendung der Methode

In [12] wurde die entwickelte Methode bereits genutzt, um drei einfache Wertschöpfungsketten miteinander zu vergleichen. Dort wurden die Wasserstoff-Wertschöpfungsketten anhand ihrer CO₂-Emissionen und ihres Gesamtenergieverbrauchs bewertet. In diesem Beitrag sollen dieselben drei Wasserstoff-Wertschöpfungsketten anhand ihrer Investitionskosten und Betriebskosten miteinander verglichen werden. Im Bild 4 wurde das in diesem Beitrag vorgestellte Schema genutzt, um die drei Wasserstoff-Wertschöpfungsketten darzustellen.

- Versorgung per Pipeline
- Versorgung per Tube-Trailer (GH₂-Trailer)
- Versorgung per Flüssigtankwagen (LH₂-Trailer)

In Wertschöpfungskette a wird der in einer AEL produzierte Wasserstoff per Pipeline zu einer Direktreduktionsanlage transportiert, die sich in 100 km Entfernung befindet. Vereinfacht wird angenommen, dass zwischen Produktionsort und Verbrauchsort eine direkte Rohrleitung besteht. In Wertschöpfungskette b und c erfolgt der Transport des Wasserstoffs dann per LKW, gasförmig bzw. flüssig.

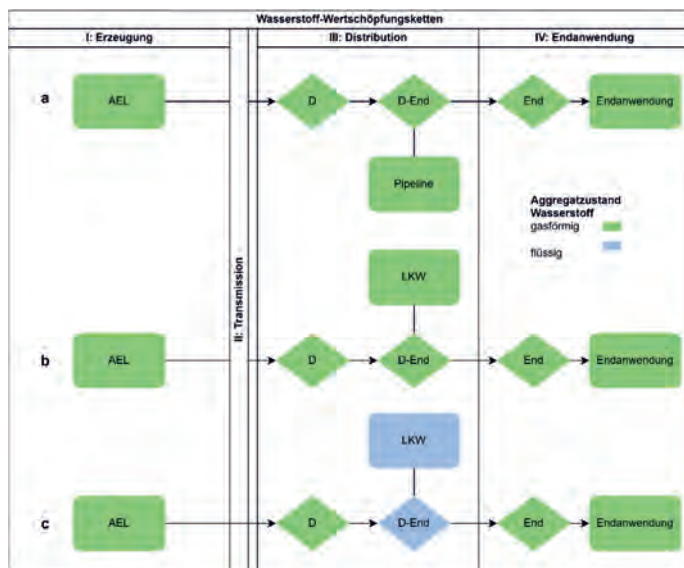


Bild 4: Drei Wasserstoff-Wertschöpfungsketten entsprechend der Szenarien a, b und c

In Tabelle 1 sind die Investitionskosten und die Betriebskosten für verschiedene Konditionierungssysteme und Transportsysteme aufgelistet. Zum Teil werden Kostenfunktionen für die Berechnung der Investitionskosten angegeben, die meist in Abhängigkeit von der Nachfrage x in Tonnen Wasserstoff pro Tag t_{H_2}/d ermittelt werden. Lediglich die Berechnung der Investitionskosten für den Kompressor unterscheidet sich, da dort in Abhängigkeit der Leistung des Kompressors x_{komp} in kW die Kosten ermittelt werden. Eine weitere Besonderheit für die Berechnung der Investitionskosten für den Kompressor ist, dass ein Installationsfaktor verwendet wird, um zu berücksichtigen, ob der Kompressor genutzt wird für den anschließenden Transport per Pipeline $f_{inst,pipe}$ oder per Tube-Trailer f_{inst,GH_2} . Die in der Tabelle gezeigten Kostenfunktionen bilden die Grundlage für die Berechnung der Gesamtinvestitionskosten und der Gesamtbetriebskosten pro Jahr für die drei Wasserstoffwertschöpfungsketten. Die Betriebskosten sind hier in Prozent/Jahr bezogen auf die Investitionskosten angegeben.

Tabelle 1: Kostenfunktionen/ Investitionskosten für Konditionierungssysteme und Transportsysteme nach [13] mit Daten von [5, 14, 15]

System	Investitionskosten	Betriebskosten	Installationsfaktoren
Kompressor	$15.000 \frac{\text{€}}{\text{kW}} * x_{komp}^{0,6089} * f_{inst}$	4 %/a	$f_{inst,pipe} = 2,5$ $f_{inst,GH_2} = 3$
Verflüssiger	$105 \text{ Mio. €} * \left(\frac{x}{50 \frac{t_{H_2}}{d}}\right)^{0,66}$	4%/a	-
Regasifizierung	$3.000 \text{ €} * \frac{x}{\frac{t_{H_2}}{d}}$	3 %/a	-
Pipeline	$292,152 * e^{\frac{0,0016*d}{mm}} \text{ €/m}$	5 €/ (m * a)	-
Zugmaschine	160.000 €	12 %/a	-
GH ₂ -Trailer	660.000 € für 1100 kg	2 %/a	-
LH ₂ -Trailer	860.000 € für 4300 kg	2 %/a	-
AEL	4000 €/ (Nm ³ /h)	18 €/a je kW	-

Die Gesamtinvestitionskosten $Invest_{gesamt}$ berechnen sich nach Tabelle 1 als Summe aller Investitionskosten für die Konditionierung $Invest_{kond}$ addiert mit der Summe aller Investitionskosten für den Transport $Invest_{trans}$.

$$Invest_{gesamt} = \sum Invest_{kond} + \sum Invest_{trans} \quad (1)$$

Die Gesamtbetriebskosten $Betrieb_{gesamt}$ berechnen sich entsprechend ebenfalls wie die Investitionskosten nach Tabelle 1 als Summe aller Betriebskosten für die Konditionierung $Betrieb_{kond}$ addiert mit der Summe aller Betriebskosten für den Transport $Betrieb_{trans}$.

$$Betrieb_{gesamt} = \sum Betrieb_{kond} + \sum Betrieb_{trans} \quad (2)$$

Für die Wertschöpfungskette a wurden die Investitionskosten und die Betriebskosten für den Kompressor und die Pipeline berücksichtigt. Für die Wertschöpfungskette b die Kosten für den Kompressor, für die Zugmaschinen und für die GH₂-Trailer. Für die Wertschöpfungskette c sind es die Kosten für Verflüssiger, Regasifizierung, Zugmaschine und LH₂-Trailer.

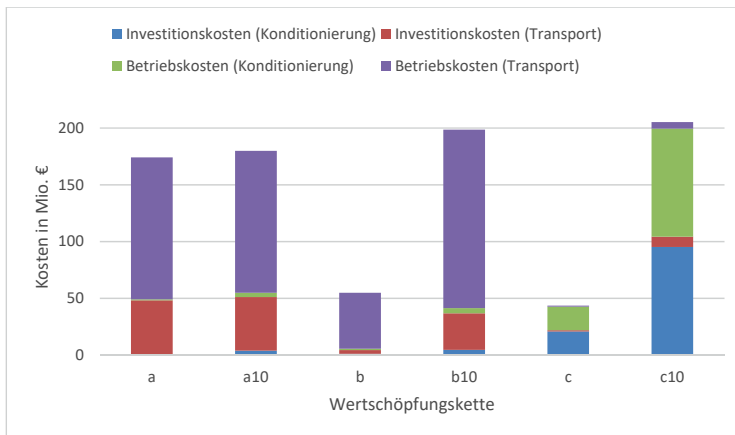


Bild 5: Investitionskosten und Betriebskosten (bezogen auf 25 Jahre) der Konditionierung- und der Transportsysteme in Mio. € für die drei Szenarien

In Bild 5 sind die Investitionskosten und die Betriebskosten (bezogen auf 25 Jahre) der Konditionierungs- und der Transportsysteme in Mio. € für alle drei Wertschöpfungsketten ersichtlich. Da in allen Szenarien die AEL als Erzeugungstechnologie verwendet wurde, wurde in den Bildern darauf verzichtet, die Kosten der AEL mit einzutragen, da alle Gesamtkosten nur gleichmäßig ansteigen würden. Die Grundlage für die Berechnung der AEL-Kosten ist dennoch in Tabelle 1 hinterlegt. Für die Pipeline wurde eine DN-300-Leitung angenommen. Als täglicher Bedarf an Wasserstoff wurden $4,3 \text{ t}_{\text{H}_2}/\text{d}$ angenommen. Diese Menge entspricht genau der Kapazität eines LH₂-Trailers. Um die gleiche Menge mit GH₂-Trailern zu transportieren, werden entsprechend vier LKWs benötigt. Die Pipeline ist natürlich in der Lage, deutlich mehr Wasserstoff zu transportieren, aber dennoch sollte hier für ein Distributionsnetz eine sinnvolle Leitungsgröße angenommen werden, die für wachsende zukünftige Bedarfe

ausgelegt ist. Die Pipeline ist immer die kostenintensivste Transportmöglichkeit. Zusätzlich wurden noch die Wertschöpfungsketten a10, b10 und c10 in Bild 5 zum Vergleich eingefügt. Bei diesen Wertschöpfungsketten handelt es sich um die drei bisherigen Wertschöpfungsketten. Der Tagesbedarf von $4,3 \text{ t}_{\text{H}_2}/\text{d}$ wurde allerdings um den Faktor 10 erhöht. Bei den Wertschöpfungsketten a10, b10 und c10 wird deutlich, dass sobald größere Mengen transportiert werden müssen, die Pipeline die wirtschaftlichste Option darstellt. Außerdem wird der große Einfluss der Investitionskosten der Verflüssigung für große Mengen an Wasserstoff in c10 deutlich. In b10 wird der Einfluss der Bedarf an mehr LKWs deutlich. Dadurch steigen die Investitionskosten und Betriebskosten für den Transport erheblich an. Zu betonen ist, dass hier nicht die 10-fache Menge an Zugmaschinen angenommen wurde, sondern lediglich die 3-fache Menge. Es wird davon ausgegangen, dass die Lkws die Strecke von 100km dann mehrmals pro Tag fahren.

6. Zusammenfassung und Ausblick

In diesem Beitrag wird eine Methode zur Bewertung von Wasserstoff-Wertschöpfungsketten vorgestellt und für eine Betrachtung der Investitionskosten und Betriebskosten von drei Wertschöpfungsketten eingesetzt. Außerdem wurde ein Schema vorgestellt, mit dem sich beliebige Wasserstoff-Wertschöpfungsketten einheitlich abbilden lassen.

Neben der wirtschaftlichen Bewertung der Wasserstoff-Wertschöpfungsketten ist auch die technische und ökologische Bewertung von Bedeutung. In einem anderen Beitrag [12] konnte gezeigt werden, dass in Bezug auf die CO₂-Emissionen und den Energieverbrauch die Wertschöpfungskette a sich besonders vorteilhaft darstellt. Die vorgestellte Methode wird künftig noch weiterentwickelt, und es werden weitere Bewertungskriterien herangezogen, um aufzuzeigen, wie sich die Gesamtbetrachtung dann verändert. Beispielsweise werden zukünftig auch noch die variablen Kosten mitberücksichtigt werden, damit mit Hilfe der Gesamtkosten auf einen entstehenden Wasserstoffpreis geschlossen werden kann.

References

- [1] BMWi. Nationales Reformprogramm 2020 - Die Nationale Wasserstoffstrategie.
- [2] YANG, C., OGDEN, J.: Determining the lowest-cost hydrogen delivery mode. International Journal of Hydrogen Energy 2007; 32(2): S. 268–286, doi: 10.1016/j.ijhydene.2006.05.009.
- [3] Mintz, M., Gillette, J., Elgowainy, A., *et al.*: Hydrogen Delivery Scenario Analysis Model for Hydrogen Distribution Options. Transportation Research Record 2006; 1983(1): S. 114–120, doi: 10.1177/0361198106198300116.

- [4] Reuß, M., Dimos, P., Léon, A., Grube, T., Robinius, M., Stolten, D.: Hydrogen Road Transport Analysis in the Energy System: A Case Study for Germany through 2050. *Energies* 2021; 14(11): S. 3166, doi: 10.3390/en14113166.
- [5] Reuß, M., Grube, T., Robinius, M., Preuster, P., Wasserscheid, P., Stolten, D.: Seasonal storage and alternative carriers: A flexible hydrogen supply chain model. *Applied Energy* 2017; 200: S. 290–302, doi: 10.1016/j.apenergy.2017.05.050.
- [6] Wulfs, C.: Wasserstoff im motorisierten Individualverkehr – Möglichkeiten und deren Bewertung. Dissertation. Technische Universität Hamburg. 2017.
- [7] Seydel, P. Entwicklung und Bewertung einer langfristigen regionalen Strategie zum Aufbau einer Wasserstoffinfrastruktur - auf Basis der Modellverknüpfung eines Geografischen Informationssystems und eines Energiesystemmodells. ETH Zurich 2008.
- [8] Almansoori, A., Shah, N. Design and Operation of a Future Hydrogen Supply Chain. *Chemical Engineering Research and Design* 2006; 84(6): S. 423–438, doi: 10.1205/cherd.05193.
- [9] Agnolucci, P., McDowall, W.: Designing future hydrogen infrastructure: Insights from analysis at different spatial scales. *International Journal of Hydrogen Energy* 2013; 38(13): S. 5181–5191, doi: 10.1016/j.ijhydene.2013.02.042.
- [10] Samsatli, S., Staffell, I., Samsatli, N. J.: Optimal design and operation of integrated wind-hydrogen-electricity networks for decarbonising the domestic transport sector in Great Britain. *International Journal of Hydrogen Energy* 2016; 41(1): S. 447–475, doi: 10.1016/j.ijhydene.2015.10.032.
- [11] Bolat, P., Thiel, C.: Hydrogen supply chain architecture for bottom-up energy systems models. Part 1: Developing pathways. *International Journal of Hydrogen Energy* 2014; 39(17): S. 8881–8897, doi: 10.1016/j.ijhydene.2014.03.176.
- [12] Beck, J.-P., Fay, A.: Methode zur Bewertung von Wasserstoff- Wertschöpfungsketten. *atp magazin* 6/2022.
- [13] Reuß, M.: Techno-ökonomische Analyse alternativer Wasserstoffinfrastruktur. Dissertation. RWTH Aachen. 2019.
- [14] Gas2energy.net: Systemplanerische Grundlagen der Gasversorgung. 2., überarbeitete und erweiterte Auflage. München: DIV Deutscher Industrieverlag 2015.
- [15] Smolinka T. et al.: Studie IndWEde – Industrialisierung der Wasserelektrolyse in Deutschland: Chancen und Herausforderungen für nachhaltigen Wasserstoff für Verkehr, Strom und Wärme; 2018.

Chancen und Herausforderungen für Künstliche Intelligenz in kleinen und mittelständischen Unternehmen

Best Practices zur erfolgreichen Integration von Künstlicher Intelligenz für den produzierenden Mittelstand

Daniel Vranješ, Philip Topalis, Prof. Dr. Oliver Niggemann,
Helmut-Schmidt-Universität, Hamburg

Kurzfassung

Künstliche Intelligenz ist ein allgegenwärtiger Bestandteil unseres Lebens. Die Anwendung und Verbreitung dieser Technologie wird dabei primär von großen IT-Unternehmen vorangetrieben, während traditionelle Unternehmen stark auf ihre Domänenexpertise beruhen und kleinen Unternehmen häufig die Ressourcen fehlen, um neue Technologien signifikant zu fördern. EU-weit haben kleine und mittlere Unternehmen wichtige technologische und ökonomische Positionen, sodass die Bereitstellung und Integration von Künstlicher Intelligenz für diese ein wichtiger Baustein für die Gesamtwettbewerbsfähigkeit der europäischen Wirtschaft ist. Wir zeigen Einsatzmöglichkeiten für Künstliche Intelligenz auf, behandeln sowohl Hürden als auch Enabler für die Umsetzung von Vorhaben im produzierenden Mittelstand und verdeutlichen dies anhand eines Fallbeispiels.

1. Einführung

Künstliche Intelligenz (KI) ist heutzutage in aller Munde und zurecht eine Schlüsseltechnologie in vielen Bereichen. So unterstützt KI z.B. bei autonomem Fahren [1], medizinischen Diagnosen [2], der Steuerung von Robotern [3] und der Anomalieerkennung in Produktionsprozessen [4][5]. Eine Umfrage aus dem Jahr 2020 unter Industrieunternehmen aus Deutschland zeigt, dass KI insbesondere in den Bereichen Predictive Maintenance, Steigerung der Produktivität sowie Optimierung von Produktions- und Fertigungsprozessen wichtige Vorteile zugeschrieben werden [6]. Trotz dieses positiven Meinungsbildes über KI, trifft man häufig auf Skepsis bezüglich der tatsächlichen Anwendung KI-basierter Methoden. Eine Umfrage aus dem Jahr 2019 unter Industrieunternehmen aus Deutschland zeigt, dass Vorhaben im Bereich Industrie 4.0, welche eine Basis für KI Anwendungen darstellen, häufig durch Investitionskosten, Anforderungen an Daten und den Mangel an Fachkräften gehemmt werden [7]. Die genannten Hemmnisse nehmen in Ihrer Wirkung stark zu, je kleiner ein Unternehmen ist, da häufig die verfügbaren Ressourcen, insbesondere bei kleinen Unternehmen, wenig Spiel-

raum für Projekte lassen, deren Return on Investment (ROI) nicht genau bestimmt werden kann oder mit zeitlichem Versatz zu erwarten ist. Entsprechend der Empfehlung 2003/361/EG der Europäischen Kommission fallen Unternehmen unter die Gruppe der Kleinen und Mittelständischen Unternehmen (KMU), sofern Sie weniger als 250 Beschäftigte und höchstens 50 Mio. € Jahresumsatz oder höchstens 43 Mio. € Bilanzsumme pro Jahr aufweisen. Die meisten Unternehmen dieser Klasse liegen hierbei deutlich unterhalb der gesetzten Grenzen. Kumulativ betrachtet stellen KMUs jedoch einen signifikanten Anteil der Wirtschaftsleistung dar. In 2019 erwirtschafteten KMUs ca. 30% des Umsatzes in Deutschland [8], hielten einen Anteil von ca. 56% der insgesamt in Deutschland Beschäftigten [9] und machten einen Anteil von 99,4% der deutschen Unternehmen aus [10].

Dass KMUs andere Anforderungen und Bedürfnisse innehaben, als große Unternehmen, ist allgemein bekannt. Die Defizite in Organisationsgröße und verfügbaren Budgets wurden häufig in Literatur zum Management solcher Unternehmen aufgegriffen, um angepasste Unternehmensstrategien und Prozesse zu entwickeln [11] [12].

Bei klassischen produzierenden Unternehmen äußern sich Defizite jedoch auch sehr greifbar in schwach automatisierten Prozessen und Produktportfolios mit geringem Digitalisierungsgrad, z.B. überwiegend manuelle Fertigungsstraßen. So kommt es dazu, dass insbesondere KMUs, die Potenziale, welche sich durch KI Anwendungen ergeben könnten, ungenutzt lassen, obwohl in KI allgemein Chancen für verschiedenste Anwendungen gesehen werden [13]. Aufgrund der starken KMU-Prägung der deutschen und europäischen Wirtschaft, ist es wichtig, auf die Bedürfnisse von KMUs einzugehen und Lösungen für die Hürden bei der Anwendung von KI auszuarbeiten.

In unserem Beitrag beschreiben wir Anwendungsmöglichkeiten für KI und typische Hürden bei der Implementierung dieser. Wir zeigen darüber hinaus auf, welche KI-Enabler genutzt werden können, um KI effizient nutzbar zu machen. Unser Beitrag setzt hierbei einen Fokus auf Unternehmen des produzierenden Mittelstandes. Anhand eines Fallbeispiels basierend auf einem KMU aus dem Maschinenbau verdeutlichen wir den lösungsorientierten Umgang mit typischen Hürden bei der initialen Verwendung von KI in neuen Produkten. Basis unseres Beitrags sind persönliche Erfahrungen und Studien dritter. Die aufgeführten Aspekte sind weder hinreichende noch notwendige Bestandteile eines erfolgreichen KI-Vorhabens und eher als potenzielle Handlungsempfehlungen anzusehen.

Der Beitrag gliedert sich wie folgt. In Kapitel 2 beschreiben wir typische Anwendungsfelder für KI. In Kapitel 3 gehen wir auf wichtige Herausforderungen ein, die insbesondere KMUs an der Verwendung von KI hindern. In Kapitel 4 stellen wir den aktuellen Stand der Technik bezüglich Enabler für KI dar. In Kapitel 5 zeigen wir anhand eines Fallbeispiels einen potenziel-

len Lösungsansatz für die erfolgreiche Integration von KI durch ein KMU. In Kapitel 6 fassen wir die Erkenntnisse und Handlungsempfehlungen abschließend zusammen und geben einen Ausblick.

2. Anwendungsgebiete für KI

Die potenziellen Anwendungsgebiete für KI sind mannigfaltig und variieren stark mit Unternehmensgrößen, Branchen und Produkten. Häufige Anwendungen sind Datenanalyse für Entscheidungsprozesse, Prozessautomatisierung, Chatbots, neue digitale Geschäftsmodelle, Sprachverarbeitung und KI als Bestandteil von Produkten und Dienstleistungen [14]. Hier möchten wir folgend zwei Aspekte hervorheben, die für produzierende KMUs von besonderer Bedeutung sind.

Optimierung der Produktion mittels KI

Die Optimierung der Produktion hat einen direkten Einfluss auf die Maximierung der internen Wertschöpfung und ist wichtiger Aspekt für den Erhalt der Wettbewerbsfähigkeit, insbesondere in globalen Märkten. KI kann hierbei sowohl bei Logistik- und Planungsprozessen unterstützen [15], als auch mittels Predictive Maintenance den Zustand der Produktionsmaschinen überwachen, um die Ausfallraten gering zu halten, Degradationen oder Defekte frühzeitig zu erkennen und eine optimale Auslastung der Produktion zu ermöglichen [16]. KI kann darüber hinaus verwendet werden, um Anomalien im Produktionsprozess zu erkennen [5]. Anomalien äußern sich hierbei als Abweichungen vom Normalzustand einer Maschine oder Anlage mit ggf. Folgeschäden für die Produktionsanlage oder für das Produkt selbst. Systeme zu Anomalieerkennung erhöhen die Sicherheit der Produktion und ermöglichen eine instantane Qualitätskontrolle.

Neue Produkte und Geschäftsmodelle mittels KI

Die Erweiterung analoger Produkte um Sensoren, Datenverarbeitung und digitale Schnittstellen eröffnet gänzlich neue Anwendungsfälle und Vermarktungsmöglichkeiten. KI kann hier gezielt für die Umsetzung von einzelnen Funktionalitäten genutzt werden. Darüber hinaus können passende Infrastrukturen und darauf aufbauende Services wie das genannte Predictive Maintenance z.B. lokal oder in Form von Software as a Service Produkten angeboten werden. Unser Fallbeispiel in Kapitel 5 behandelt diese Aspekte ausführlich.

3. Hürden für die Implementierung von KI durch produzierende KMU

Wie bereits in der Einführung aufgezeigt, sehen die meisten Unternehmen in mehreren Anwendungsfeldern großes Potenzial in KI. Dennoch ist die tatsächliche Anwendung von KI häufig mit Hürden behaftet, welche effektiv zu einer geringen Anwendungsrate führen. Eine weitere Studie aus 2021 mit einem Fokus auf KMUs aus dem Maschinen- und Anlagenbau bekräftigt das Bild, dass viele Personen in produzierenden Unternehmen KI als eine Technologie mit großen Chancen einschätzen, jedoch selbst wenig Kontakt mit KI haben und nur wenig Expertise zu KI vorhanden ist [17]. Nachfolgend zeigen wir einige Hürden im Detail.

Mangelnde Expertise:

Grundlage für die erfolgreiche Anwendung von KI-Methoden sind digitale Prozesse oder Produkte sowie Daten in hoher Qualität. Eine Studie aus April 2021 gibt an, dass die Mehrheit der befragten KMUs Reifegrade der Stufen 1 oder 2 von 5 bezüglich der Entwicklung digitaler Kompetenzen, darunter insbesondere in den Kategorien Prozessverständnis, Datenanalyse und Programmierung, aufweist [18]. Eine weitere Studie aus dem Jahr 2020 bekräftigt dies mit Fachkräftemangel als größter Herausforderung bezüglich der Umsetzung von KI, wobei der Anteil an Change-Spezialisten zurückgegangen ist, was zumindest ein positives Zeichen für die allgemeine Akzeptanz von KI ist [19]. Der allgemeine Fachkräftemangel in der IT-Branche setzt sich hierbei auch in 2022 fort [20] und ist als signifikante Hürde anzusehen. Aufgrund des Fachkräftemangels besteht gerade für KMUs die Schwierigkeit, Fachwissen im Bereich KI zu erhalten [21]. Daher sollte ein internes KI-Training bzw. Kompetenzaufbau ein Teil der KI-Anwendungsstrategie sein.

Mangelnde Infrastruktur:

Der Mangel an Expertise im Bereich der Digitalisierung spiegelt sich häufig direkt in einer mangelnden Infrastruktur für digitale bzw. automatisierte Prozesse und Produkte wieder. Ohne eine ganzheitliche Betrachtung, können digitale Ansätze häufig nur zu Insellösungen werden, bei denen ein digitaler bzw. automatisierter Teil an analoge bzw. manuelle Teile angrenzt und somit sein Potenzial nicht entfalten kann. Zudem ist die Einstiegshürde für neue Lösungen deutlich höher, sofern diese nicht einfach in eine übergeordnete Infrastruktur integriert werden können. Möchte man z.B. eine Produktionsanlage mittels KI überwachen, aber verfügt nicht über eine Infrastruktur mit standardisierten Schnittstellen und Datenformaten, z.B. OPC UA, sowie eine zentrale Datenverwaltung, z.B. in Form einer oder mehrerer Datenbanken und darauf aufbauender Services, so muss entweder zunächst die notwendige Infrastruktur aufgebaut werden oder es ist nur eine lokale KI Lösung möglich.

Greifbarkeit der Potenziale:

Neue Technologien verursachen häufig hohe Kosten durch deren Implementierung und den Erwerb von Know-how. Die erreichbaren Potenziale einer KI Lösung sind oftmals nicht im Voraus klar zu definieren, da man nicht vorab bestimmen kann, wie stark eine KI einen Prozess verbessern kann und welche Einsparungen generiert werden können, bevor nicht zumindest mit einer Voruntersuchung erste Daten und Ergebnisse generiert wurden. Da für die Voruntersuchung bereits Ressourcen notwendig sind, können Vorhaben aufgrund der mangelnden Greifbarkeit der Potenziale bereits kurz nach der Ideenphase wieder verworfen werden. Der Grund dafür sind häufig konkurrierende Investitionsvorhaben. Somit besteht aufgrund des relativ hohen anfänglichen Ressourcenaufwandes und der nur ungewissen Potenziale die Gefahr, dass Unternehmen auf Basis einer Kosten-Nutzen-Betrachtung entscheiden, nicht in KI zu investieren.

Datenverfügbarkeit:

Die Basis vieler KI Modelle sind Daten. Mit der Qualität und Menge der verfügbaren Daten, steigen Qualität und Robustheit der zu entwickelnden Modelle. Die Grundlage für eine erfolgreiche Einführung von KI liegt somit in einem grundlegend gut funktionierenden Datenökosystem [22]. Dies erfordert die Kenntnis über die Verfügbarkeit vorhandener Daten und dem Erhalten zusätzlicher Daten. Häufig sind die benötigten Daten zu Beginn jedoch noch nicht in benötigter Güte (z.B. Auflösung der Messgrößen zu gering, inkonsistente Datenstrukturen, fehlende Einträge, nicht maschinen-interpretierbare Formate) oder überhaupt vorhanden, sodass Aufwände für die Erstellung oder Aufbereitung der Daten getätigt werden müssen. Der Grund für den Mangel an benötigter Güte ist, dass KMU im normalen Betrieb nicht an datengesteuerte Ansätze gewöhnt sind [23]. Insbesondere bei Daten aus physischen Systemen ist bei der Erstellung oder Aufbereitung der Daten mit relativ hohen Aufwänden zu rechnen und es muss darauf geachtet werden, dass Daten repräsentativ für die geplanten Anwendungen sind. Die mangelnde Datenbasis in KMU hängt wiederum mit dem bereits angesprochenen geringeren digitalen Reifegrad zusammen. Dies betont die Notwendigkeit einer IT-Infrastruktur als Voraussetzung für die Anwendung von KI.

4. Enabler & Best Practices für die Nutzung von KI

Das Potenzial von KI wurde auf technischer, kommerzieller und politischer Ebene erkannt. Hierdurch ergeben sich verschiedenste Enabler, welche den Zugang zu KI erleichtern.

Open Source Software:

Open Source Software (OSS) ist ein fundamentaler Bestandteil vieler KI Projekte. Mit Python, scikit-learn, TensorFlow, PyTorch und vielen weiteren Frameworks existiert professionelle Software für Data Science und KI Applikationen, welche kostenlos genutzt werden kann und Komplexitäten von einfachen Voruntersuchungen bis hin zu produktionsreifen Systemen unterstützt.

Hardware und Cloud-Dienste

Das Training einer KI auf einer neuen Problemstellung basiert in der Regel auf einer Vielzahl an Trainingsschritten über tausende von Datenpunkten. Durch immer leistungsfähigere Prozessoren und Grafikkarten kann die hierfür benötigte Rechenleistung für kleine bis mittlere Aufgaben durch lokale Systeme bereitgestellt werden. Ist die Anschaffung von Hardware keine Option oder übersteigen die zu trainierenden Modelle die verfügbare Rechenleistung, so können Cloud-Dienste zum Einsatz kommen. Dienste wie die von Amazon [24], Google [25] und Microsoft [26] stellen Nutzern flexibel Rechenleistung zur Verfügung und bieten sowohl grafische als auch codebasierte Tools zur Erstellung von Modellen. Für einen leichteren Einstieg sind einige der Funktionen kostenlos verfügbar und Anwender können offizielle Schulungen für fortgeschrittene Funktionen belegen.

AutoML & Low-Code Frameworks:

In dem Begriff AutoML werden verschiedene Ansätze zusammengefasst, die den Erstellungsprozess von ML und KI Modellen automatisieren sollen. Hierzu gehören unter anderem Datenpipelines, Modellarchitektur, Modelloptimierung und Modellvalidierung. Ein Beispiel für solch ein Werkzeug ist CreateML [27], welches dem Anwender ermöglicht, verschiedene Klassifikations- und Detektionsmodelle ohne Programmierkenntnisse zu erstellen, da es lediglich gelabelte Daten benötigt. Die oben genannten Cloud-Dienste bieten mittlerweile ebenfalls low-code Nutzeroberflächen für Endanwender, um Modelle mittels grafischer Oberflächen zu erstellen. Insbesondere für Datenanalysen im Bereich Business Intelligence eignen sich auch grafische Werkzeuge wie Tableau [28] oder RapidMiner [29].

Förderprogramme & Kooperationen:

Förderprogramme bieten eine exzellente Basis, um Themen wie Digitalisierung und KI anzugehen. Einerseits werden die internen Kosten durch die externe Förderung von Personal und Materialkosten reduziert, andererseits bedingen die meisten Förderprogramme die Bearbeitung eines Projektes durch ein Projektkonsortium. Dieses Konsortium besteht bei KI-Projekten häufig aus einem Domänenexperten, welcher in der Regel den Anwendungsfall prägt, und einem KI-Experten (häufig in Form einer Forschungs- oder Bildungseinrichtung), der die Datenanalyse und Modellbildung leitet. Ergänzend können nach Bedarf noch weitere Partner inkludiert werden. Förderprogramme dieser Art eignen sich ideal für den Aufbau von internem Wissen mittels Wissenstransfer und werden in regelmäßigen Abständen veröffentlicht. Förderprogramme haben häufig einen Forschungsschwerpunkt und eignen sich aufgrund von Projektlaufzeiten und Vorbereitungsphasen besonders für langfristig angesetzte Vorhaben.

Externe KI-Experten:

Außerhalb von geförderten Vorhaben können insbesondere für kurz- und mittelfristige Projekte externe KI-Experten in Form von Freelancern oder Agenturen genutzt werden, um schnelle Implementierungen voranzutreiben und ggf. auch Kosten gegenüber einer internen Entwicklung gering zu halten und somit eine kürzere Amortisationszeit zu ermöglichen. Mit dem Outsourcing von Entwicklungsleistungen sind allerdings auch dessen typische Nachteile verbunden, welche in die Planung mit einbezogen werden müssen. Gelegentlich bieten sich auch Forschungs- und Lehreinrichtungen als erster Kontakt an. Diese haben häufig kein kommerzielles Interesse und sind eher an aufbauenden Kooperationen interessiert, sodass Voruntersuchungen teilweise pro bono durchgeführt werden können.

Unternehmensstrategie und Management:

Für die langfristig erfolgreiche Integration von KI sind dessen Akzeptanz und Verständnis essenziell. Angefangen bei der Verknüpfung mit der Unternehmensstrategie oder Digitalisierungsstrategie und einem initialen Change-Management, über die strategische Personalentwicklung hin zum konkreten Design von Prozessen und Produkten muss KI ganzheitlich betrachtet werden. Ohne Unterstützung vom Management werden IT oder KI Experten keine weitreichenden Entwicklungen anstoßen können. Andererseits gewährleistet die Unterstützung des Managements eine ausreichende Zuweisung von Ressourcen und verringert den organisatorischen Widerstand gegen die Einführung. Erfolgserlebnisse sind hierbei ein wichtiger Aspekt für anhaltende Motivation und diese realisiert man initial idealerweise durch Low

Hanging Fruits, also mit relativ wenig Aufwand umsetzbare Vorhaben. Die strategische Personalentwicklung sehen wir hierbei als kritisches Element für langfristige Erfolge. Gerade KMUs können von einer flachen Hierarchie und einem entschlossenen Management profitieren.

Vorgehensmodelle:

Aufgrund der mangelnden Erfahrung und des fehlenden Umgangs mit KI, ist zu Beginn eines solchen Projektes noch wenig Wissen über das Vorgehen zur erfolgreichen Einführung einer KI Anwendung im Unternehmen vorhanden. Es existieren verschiedene Vorgehensmodellen mit jeweils anderem Fokus zur systematischen Einführung datengetriebener Verfahren, wie z.B. CRISP-DM [30], ML4P [31], TSPD [32], EDDA [33] und PAISE® [34]. Durch die Anwendung eines dieser Modelle wird ein strukturiertes Vorgehen für KI Projekte geschaffen, welches die Einführung erleichtert und die Risiken verringert, die sich bei KI Projekten ergeben können. Dadurch wird ein Verständnis darüber geschaffen, welche Schritte und Aufgaben zur erfolgreichen Einführung einer KI Anwendung notwendig sind. Insbesondere das Life-Cycle-Management, d.h. die Überwachung und Aktualisierung der in der Produktion befindlichen KI-Systeme, ist zu beachten. Die Kenntnis über die notwendigen Schritte und Aufgaben bildet die Grundlage für die Entscheidung darüber, an welchen Stellen im Entwicklungsprozess die notwendige Expertise im eigenen Unternehmen vorhanden ist und wo die Zusammenarbeit mit externen Kompetenzträgern notwendig ist.

5. Fallbeispiel

Die A. Liersch GmbH (Liersch) ist ein KMU und hat Schwerpunkte in Entwicklung, Konstruktion und Fertigung von Sondermaschinen und Anlagen für die Herstellung von Textilien aller Art. Kernprodukte sind CNC Nähautomaten, wie der Liersch 606-M CNC für schwere Anwendungen wie die Produktion von Sicherheitsgurten. Endanwender der Nähautomaten sind primär Großserienproduzenten mit mehreren Nähautomaten in einer Produktionsstätte. Bei den Produkten von Liersch selbst handelt es sich jedoch um Kleinserien sowie Sonderanfertigungen, sodass die interne Endproduktion primär manuell abläuft. Die Erweiterung der Produktfunktionen und Services der Nähautomaten und deren Anwendung beim Endkunden im Sinne einer Smart Factory bieten somit für Liersch ein deutlich größeres Potenzial und kleinere Hürden, als die Anwendung von KI im eigenen Produktionsprozess.

Sicherheitskritische Produkte wie Lastengurte oder Sicherheitsgurte erfordern ein hohes Maß an Qualität und erlauben keine Fehler im Endprodukt. Eine gängige Praxis ist hier die manuelle Endkontrolle und der Ausschuss von Produkten mit defekten Nähten. Die automa-

tierte Überwachung (Anomalieerkennung) des Produktionsprozesses technischer Textilien ist für den Endanwender der Nähautomaten ein wichtiger Aspekt der Qualitätssicherung und Kostenreduktion. Um dem Endkunden eine bessere Qualitätskontrolle des Nähprozesses und korrektive Maßnahmen zu ermöglichen, ist ein Nähautomat mit Sensoren und interner Datenverarbeitung zur Überwachung der wichtigsten Prozessparameter definiert worden. Mit den dadurch verfügbaren Prozessdaten können die Nähautomaten als Cyber-Physische-Produktionssysteme (CPPS) [5] angesehen werden und bieten eine Basis für KI Anwendungen. Für die Anomalieerkennung sind insbesondere Deep Learning Methoden von Interesse, da diese das manuelle Feature Engineering für komplexe multivariate Zeitreihen ersetzen und ohne großen Aufwand an verschiedene Nähautomaten und Muster angepasst werden können.

Von der Hürde mangelnder KI-Expertise und Ressourcen war jedoch auch Liersch betroffen, sodass die Entwicklung ohne externe Ressourcen nur schwer möglich gewesen wäre. Da das Management und die Entwicklungsabteilung diesen Umstand rechtzeitig erkannt und in die strategische Produktplanung haben einfließen lassen, konnte für das Vorhaben eine Förderung durch das BMWi gewonnen werden, welche interne Kosten reduziert und gleichzeitig Projektpartner involviert, welche Expertenwissen in den Bereichen Sensorik und KI einbringen können. Diese beiden Aspekte unterstützen signifikant dabei, ressourcenbedingte Hürden bei der Implementierung von KI besser überwinden zu können.

Um eine KI-basierte Anomalieerkennung für Nähautomaten entwickeln zu können, sind Daten von Nähprozessen notwendig, wie sie im späteren Betrieb anfallen werden. Hier stellten sich zwei Hürden. Es waren keine historischen Daten für eine Vorstudie verfügbar und auch im Rahmen des Projektes konnten Daten erst mit lauffähiger Sensorik aufwändig manuell an den Nähautomaten erzeugt und mit Labels versehen werden. Verglichen mit rein digitalen Systemen, ist die Erstellung von Daten durch CPPS mit deutlich mehr Aufwand verbunden, da hier ein physischer Prozess ablaufen oder eine Simulation erstellt werden muss. Dabei ist darauf zu achten, dass das zur Datengenerierung verwendete System repräsentativ für das spätere Einsatzgebiet der Algorithmen ist oder zumindest Methoden wie Transfer Learning unterstützt. Maschinenparameter, Materialeigenschaften und Nahtbilder haben signifikante Einflüsse auf die verwendeten Prozessgrößen, sodass Anwendungsfälle eingegrenzt werden müssen. Die Förderung des Projektes erleichtert hierbei zumindest die finanzielle Last. Da Nähautomaten überwiegend im Normalbereich arbeiten, bilden die erzeugten Datensätze primär diesen Betrieb ab und sind somit nicht ausgeglichen. Für das explizite Klassifizieren von Fehlerfällen, müssten ausreichend Daten dieser Fehlerfälle vorliegen. Bei der KI-Entwicklung muss dieser Aspekt beachtet werden, sodass z.B. Modelle ausschließlich auf

normalen Daten trainiert werden und im späteren Betrieb Abweichungen von dem Normalbetrieb pauschal erkannt werden können. Dadurch ist zwar nur eine Unterscheidung in Normalbetrieb und Anomaliebetrieb möglich, aber dies ist ein erster wichtiger Erfolg (Low Hanging Fruit).

Versuche haben zudem gezeigt, dass die Messgenauigkeit der Sensoren essenziell für die Performanz der Algorithmen ist. So waren einige Datensätze unbrauchbar, da mit der verfügbaren Datenqualität, keine ausreichend genauen Modelle erstellt werden konnten. Diese Erkenntnis über tatsächlich benötigte Daten und deren Güte entsteht häufig erst durch die praktische Durchführung eines Projektes. Eine weitere Hürde stellte die nicht vorhandene übergeordnete Dateninfrastruktur dar. Dies hatte zur Folge, dass Datenspeicher und KI lokal auf den Nähautomaten implementiert werden mussten, obwohl seitens des entwickelten Systems externe Schnittstellen verfügbar waren. Die Verfügbarkeit von Mikroprozessoren mit genügend Rechenleistung für die Ausführung von KI-Algorithmen sowie von Open Source Frameworks, welche das Training und die Portierung von Neuronalen Netzen auf solche Hardware ermöglichen, sind hier wichtige technologische Enabler gewesen, um KI lokal in Nähautomaten integrieren zu können.

Das Vorhaben befindet sich zum Zeitpunkt dieses Beitrags in der finalen Phase der Erprobung bei Endkunden, der Erstellung einer neuen Datenbasis und dem letzten Anpassen der Algorithmen. Für den nachhaltigen Aufbau von Daten- und KI-Kompetenz hat Liersch parallel zum Forschungsprojekt mit dem Ausbau von Personal im Softwarebereich begonnen und sensibilisiert gezielt Mitarbeiter für Digitalisierungspotenziale mittels Schulungen. Die wichtigsten Aspekte unseres Fallbeispiels sind hier zusammengefasst wie folgt:

Zusammenfassung der Fallbeispiel-Hürden:

Ressourcen: Innerhalb des Unternehmens war kein Personal mit KI-Expertise vorhanden und es standen nicht genügend Ressourcen zur Verfügung, um (temporär) Personal in benötigtem Umfang einzustellen.

Daten: Für das geplante Entwicklungsvorhaben waren keinerlei historische Daten vorhanden, da das Vorhaben auf einer gänzlich neuen Hardwareplattform beruht. Zunächst musste die Sensorik implementiert und dann in mehreren Versuchsreihen eine Datenbasis geschaffen werden. Die Sensorik musste dabei den Anforderungen an Datengüte gerecht werden.

Infrastruktur: In Ermangelung einer übergeordneten Infrastruktur und dem gewünschten Anwendungsfall einer lokal arbeitenden KI, kann die Anomalieerkennung zunächst noch nicht über eine vorhandene Pipeline mit zentraler Datenverarbeitung implementiert werden, sondern muss lokal auf den Nähautomaten laufen.

Zusammenfassung der Fallbeispiel Enabler:

Management: KI wurde vom Management als Schlüsseltechnologie verstanden und dessen Implementierung vom Entwicklungsteam vorangetrieben. Datenbasierte Produktfeatures und Geschäftsmodelle werden als langfristige Innovationen und Investitionen angesehen.

Kooperation & Ressourcen: Die Identifikation eines geeigneten Förderprogramms ermöglichte die Einsparung interner Ressourcen bei gleichzeitiger Erweiterung der Expertise durch externe Projektpartner.

Open Source Software: Die Nutzung von OSS für alle Teile der ML Pipeline sparte Kosten für Lizenzen und ermöglichte die Flexible Anwendung kompatibler Frameworks.

Low Hanging Fruit: Die Nutzung von KI als ein neues Produktfeature innerhalb eines geschlossenen Systems erlaubte eine gezielte Anwendung bei gleichzeitiger Umgehung der problematischen Infrastruktur.

7. Zusammenfassung und Ausblick

Sowohl unsere subjektive Wahrnehmung als auch empirische Studien zeigen, dass primär der Mangel an Ressourcen die Umsetzung von KI-Vorhaben durch KMUs bremst. Diese Ressourcen sind einerseits IT- bzw. KI-Experten und andererseits finanzielle Mittel für Projekte. Wie wir aufzeigen, bestehen jedoch auch sehr konkrete Mittel, um diesen Mangel durch externe Ressourcen wie geförderte Forschungsprojekte und Kooperationspartner zu decken, sofern ein geeigneter Projektrahmen gefunden wurde und seitens des Managements eine Digitalisierungs- bzw. KI-Strategie festgelegt wurde, in welche sich dieses Projekt eingliedert. Langfristig betrachtet werden produzierende KMUs aufgrund von nationalem und internationalem Wettbewerb interne Ressourcen für daten-basierte Lösungen und KI-Methoden benötigen, um wettbewerbsfähig zu bleiben, da der aktuelle Vorsprung in speziellem Domänenwissen langfristig kleiner und weniger bedeutend werden wird.

Literaturangaben

- [1] Yaqoob, I., Khan, L. U., Kazmi, S. M. A., Imran, M., Guizani, N. u. Hong, C. S.: Autonomous Driving Cars in Smart Cities: Recent Advances, Requirements, and Challenges. IEEE Network 34 (2020) 1, S. 174–181
- [2] Huang, J., Shlobin, N. A., Lam, S. K. u. DeCuyper, M.: Artificial Intelligence Applications in Pediatric Brain Tumor Imaging: A Systematic Review. World Neurosurgery 157 (2022), S. 99–105
- [3] Johannink, T., Bahl, S., Nair, A., Luo, J., Kumar, A., Loskyll, M., Ojea, J. A., Solowjow, E. u. Levine, S.: Residual Reinforcement Learning for Robot Control. 2019 International

- Conference on Robotics and Automation (ICRA). Piscataway, NJ: IEEE 2019, S. 6023–6029
- [4] Niggemann, O. u. Elmers, M.: Künstliche Intelligenz in Produktion und Maschinenbau. Hintergründe, Anwendungsszenarien, Expertentipps. Berlin: VDE VERLAG 2022
 - [5] Eiteneuer, B. u. Niggemann, O.: LSTM for Model-Based Anomaly Detection in Cyber-Physical Systems (2018)
 - [6] Statista: Vorteile Künstlicher Intelligenz in deutschen Industrieunternehmen 2020 | Statista, 2022. <https://de.statista.com/statistik/daten/studie/990505/umfrage/umfrage-zu-vorteilen-kuenstlicher-intelligenz-in-deutschen-industrieunternehmen/>, abgerufen am: 27.04.2022
 - [7] Statista: Industrie 4.0: Anwendungshemmnisse in deutschen Industrieunternehmen 2019 | Statista, 2022. <https://de.statista.com/statistik/daten/studie/990517/umfrage/umfrage-zu-hemmnissen-bei-industrie-40-anwendungen-in-industrieunternehmen/>, abgerufen am: 05.01.2022
 - [8] Statista: Verteilung des Umsatzes der Unternehmen nach Unternehmensgröße | Statista, 2022. <https://de.statista.com/statistik/daten/studie/731913/umfrage/verteilung-des-umsatzes-der-unternehmen-in-deutschland-nach-unternehmensgroesse/>, abgerufen am: 04.01.2022
 - [9] Statista: Beschäftigte in Unternehmen nach Unternehmensgröße | Statista, 2022. <https://de.statista.com/statistik/daten/studie/731962/umfrage/beschaeftigte-in-unternehmen-in-deutschland-nach-unternehmensgroesse/>, abgerufen am: 04.01.2022
 - [10] Statista: Verteilung der Unternehmen nach Unternehmensgröße | Statista, 2022. <https://de.statista.com/statistik/daten/studie/731901/umfrage/verteilung-unternehmen-in-deutschland-nach-unternehmensgroesse/>, abgerufen am: 04.01.2022
 - [11] Haag, P. (Hrsg.): KMU- und Start-up-Management. Strategische Aspekte, operative Umsetzung und Best-Practice. Springer eBook Collection. Wiesbaden: Springer Gabler 2021
 - [12] Kummert, B.: Controlling in kleinen und mittleren Unternehmen. Vom Geschäftsprozessmodell zum Controller-Profil. Zugl.: Graz, Univ., Diss., 2004. Techno-ökonomische Forschung und Praxis. Wiesbaden: Dt. Univ.-Verl. 2005
 - [13] Bitkom Research: Künstliche Intelligenz kommt in Unternehmen allmählich voran. Bitkom e.V. (2021)
 - [14] PwC: Künstliche Intelligenz in Unternehmen, 2019. <https://www.pwc.de/de/digitale-transformation/kuenstliche-intelligenz/studie-kuenstliche-intelligenz-in-unternehmen.pdf>, abgerufen am: 06.01.2022

- [15] Machine Learning in der Logistik – Eine empirische Studie über die Anwendung in deutschen Unternehmen, Witthaut, M. u. Culotta, C., 2021
- [16] Wang, K. u. Wang, Y.: How AI Affects the Future Predictive Maintenance: A Primer of Deep Learning. In: Wang, K., Wang, Y., Strandhagen, J. O. u. Yu, T. (Hrsg.): Advanced Manufacturing and Automation VII. Lecture Notes in Electrical Engineering, Bd. 451. Singapore: Springer Singapore 2018, S. 1–9
- [17] Bunte, A., Richter, F. u. Diovisalvi, R.: Why It is Hard to Find AI in SMEs: A Survey from the Practice and How to Promote It. ICAART 2021. Proceedings of the 13th International Conference on Agents and Artificial Intelligence : online streaming, February 4-6, 2021. Sétubal: SCITEPRESS - Science and Technology Publications Lda 2021, S. 614–620
- [18] Hellge, V.: Readiness-Check Digitalisierung. Ergebnisse April 2021, 2021.
https://kompetenzzentrum-kaiserslautern.digital/wp-content/uploads/2021/04/RC_Kurzbericht_Arpiil2021.pdf, abgerufen am: 05.01.2022
- [19] Deloitte Deutschland: KI-Studie 2020: Wie nutzen Unternehmen Künstliche Intelligenz? | Deloitte Deutschland, 2022. <https://www2.deloitte.com/de/de/pages/technology-media-and-telecommunications/articles/ki-studie-2020.html>, abgerufen am: 06.01.2022
- [20] Bitkom Research: IT-Fachkräftelücke wird größer: 96.000 offene Jobs. Bitkom e.V. (2022)
- [21] Wangermann, T.: KI in KMU: Rahmenbedingungen für den Transfer von KI-Anwendungen in kleine und mittlere Unternehmen. Konrad Adenauer Stiftung, 2020.
- [22] Bughin, J., Hazan, E., Ramaswamy, S., Chui, M., Allas, T., Dahlström, P., Henke, N. u. Trench, M.: Artificial intelligence the next digital frontier. McKinsey and Company Global Institute 47.3.6 (2017).
- [23] Bauer, M. van Dinther, C., Kiefer D. (2020): Machine Learning in SME: An Empirical Study on Enablers and Success Factors. In: Americas Conference on Information Systems (AMCIS 2020).
- [24] Amazon Web Services, Inc.: Amazon Web Services AWS – Server Hosting & Cloud Services, 2022. <https://aws.amazon.com/de>, abgerufen am: 06.01.2022
- [25] Google Cloud: Cloud-Computing-Dienste & Google Cloud, 2021.
<https://cloud.google.com/>, abgerufen am: 06.01.2022
- [26] Cloud-Computing-Dienste | Microsoft Azure, 2022. <https://azure.microsoft.com/de-de/>, abgerufen am: 06.01.2022
- [27] Apple Developer: Create ML - Machine Learning - Apple Developer, 2021.
<https://developer.apple.com/machine-learning/create-ml/>, abgerufen am: 06.01.2022

- [28] Tableau: Tableau: Software für Business Intelligence und Analytics, 2022.
<https://www.tableau.com/de-de>, abgerufen am: 07.01.2022
- [29] RapidMiner: RapidMiner | Best Data Science & Machine Learning Platform, 2021.
<https://rapidminer.com/>, abgerufen am: 07.01.2022
- [30] Wirth, R. u. Hipp, J.: Crisp-dm: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, volume 1. Springer-Verlag London, UK, 2000.
- [31] Eisenträger, M., Frey, C., Herzog, A., Moghiseh, A., Morand, L., Pfrommer, J., Stephani, H., Stoll, A. u. Wessels, L.: ML4P – Vorgehensmodell Machine Learning for Production, Oktober 2020.
- [32] Microsoft: Team Data Science Process, 2022. <https://docs.microsoft.com/en-us/azure/architecture/data-science-process/overview>, abgerufen am: 07.01.2022
- [33] Hesenius, M., Schwenzfeier, N., Meyer, O., Koop, W., u. Gruhn, V.: Towards a software engineering process for developing data-driven applications. In 2019 IEEE/ACM 7th International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering (RAISE), pages 35–41. IEEE, 2019.
- [34] Hasterok, C., Stompe, J., Pfrommer, J., Usländer, T., Ziehn, J., Reiter, S., Weber, M. u. Riedel, T.: PAISE®. Das Vorgehensmodell für KI-Engineering. Fraunhofer IOSB 2021

Bring your Machine Learning Intelligence to the Edge – An Example on Edge Analytics for Cement Strength Prediction

Dr. M. Platenius-Mohr, Dr. J. Schlake, M. Vach,
ABB AG Corporate Research Center, Ladenburg;
R. Koorts, ABB Switzerland Ltd, Daettwil, Switzerland

Kurzfassung

Der Marktanteil von Anwendungen aus dem Bereich Analytics, Künstliche Intelligenz und Maschinelles Lernen wächst – auch im Prozessautomatisierungssektor. Eine relevante Frage ist, wie und wo man diese Anwendungen bereitstellt, denn in vielen Fällen ist die Übertragung von Massen an generierten Daten in die Cloud nicht praktikabel. Stattdessen kann man dem vielversprechenden Trend „Edge Analytics“ folgen, welcher auf Edge Computing basiert, und bei dem die Verarbeitungsaufgaben nah an dem Ort ausgeführt werden, wo die Daten generiert werden, um schnellere Antworten zu erzielen.

Wir haben einen Proof-of-Concept implementiert, der Edge Analytics am Beispiel der Vorhersage einer wichtigen Qualitätseigenschaft von Zement, der 28-Tage-Druckfestigkeit, exemplarisch realisiert. Der Kunde, also der Anlagenbetreiber, profitiert von einer zuverlässigen und akkuraten Vorhersage, weil sie es erlaubt den Produktionsprozess früher zu korrigieren und den Ertrag zu verbessern, während gleichzeitig Kosten verringert werden. In diesem Artikel diskutieren wir diesen und weitere Anwendungsfälle für Edge Analytics sowie eine Architektur für eine Lösung, die Edge Analytics am Beispiel von Zementfestigkeitsvorhersagen zeigt. Weiterhin beleuchten wir Herausforderungen und offene Forschungsmöglichkeiten in diesem Bereich.

Abstract

The market share of applications in the area of analytics, artificial intelligence and machine learning grows –also in the process automation sector. One relevant question is how and where to deploy analytics applications because, in many cases, it is not practicable to transfer the mass of generated data to the cloud. Instead, a promising trend to leverage is edge analytics, which is based on Edge Computing, where processing tasks are executed close to where the data is generated, in order to enable faster responses.

We implemented a proof-of-concept, illustrating edge analytics on an example use case of attempting to predict a key quality of cement, the 28-day strength. The customer, i.e., the plant operator, would benefit from a reliable and accurate prediction as it would allow to correct the process ahead of time to increase product yield while also reducing costs. In this article, we discuss this and further use cases and an architecture for a solution. Furthermore, we highlight challenges and open research opportunities in this area.

1. Introduction

The market of applications in the area of analytics, artificial intelligence (AI) and machine learning (ML) grows, and these trends have also already entered the process automation sector. There is an increasing number of analytics applications processing sensor data and lab sampled data in order to make industrial production processes more intelligent and to increase the level of process autonomy. One relevant question is how and where to deploy such kinds of applications. For process automation, in many cases, it is not practicable to transfer the huge mass of generated data to the cloud. Instead, a promising trend to leverage is edge analytics.

Edge Analytics is based on Edge Computing, where processing tasks are executed close to where the data is generated in order to enable faster responses. In case of edge analytics, this means that the analytics applications run on edge computing nodes located close to the devices that provide the data to be analyzed. And there are many more benefits, including security and cost aspects. Edge computing is already widely driven in the area of consumer products, like mobile phones. Example use cases from process automation include, amongst many more, asset condition monitoring, process function monitoring, and asset performance management,

In this article, we discuss use cases and an architecture for a solution, illustrating edge analytics on an example use case to predict cement strength. For this purpose, an AI/ML model is trained based on historical lab data from a cement production site. The training, testing, and validation is done in the cloud and then the trained ML model is deployed on the edge following the NAMUR Open Architecture (NOA) [2]. Furthermore, we highlight challenges and open research opportunities in the area of edge analytics for applications in process industries.

This article is structured as follows: In the next section, we introduce the area of edge analytics and discuss its benefits for industrial applications. In Section 3, we describe one concrete analytics use case in more detail: cement strength prediction. We also point at related use

cases. Section 4 gives an overview of an architecture for edge analytics in the case of cement strength prediction. Section 5 discusses challenges and future research directions, before the article is concluded in Section 6.

2. Edge Analytics

In the last decade there was a trend towards processing IT applications, as well as advanced industrial analytics applications in the cloud, leveraging its ability for big data processing. However, in many cases, it is not practicable to transfer the huge mass of generated data to the cloud. Building on a more local solution has benefits as well, especially (but not only) in terms of end-to-end latency. This is why Edge Analytics has emerged in industrial applications.

Edge Analytics leverages Edge Computing, where processing tasks are executed close to where the data is generated, in order to enable faster responses [3], [4]. In case of edge analytics, this means that the analytics applications run on edge computing nodes located close to the devices that provide the data to be analyzed. Example use cases from process automation include asset condition monitoring, process function monitoring, and asset performance management, but there are also many applications beyond, e.g., in the areas of smart cities, smart transportation, and public safety [1]. In [1], we surveyed further use cases and architectural alternatives from academic literature. All in all, industry investment as well as research interests in edge computing have grown a lot within the last years [3], including efforts for industrial edge computing reference architectures and standardization (e.g., [6,7]. Accordingly, Edge Computing is on its way to become a leading digital technology in process industries – and so is Edge Analytics.

In addition to the ability to process real-time data, edge computing and edge analytics architectures come with many benefits:

- It is easier to achieve low latency and high bandwidth due to the physical proximity to the devices. This means higher volumes of data can be processed at faster response times.
- The possibility to (pre-)process data on the edge allows to reduce traffic towards the cloud as, depending on the concrete use case, only filtered and/or aggregated information needs to be transmitted to the cloud.

- This also means that scalability of analytics application deployments improves applications can potentially scale over more devices, more data, more users, and more use cases.
- On the edge, it is usually easier to protect applications and data with respect to privacy and cyber-security policies. Since data is processed on-site, before releasing it (or parts of it in filtered/aggregated form) to the cloud via a secured edge gateway, the attack surface is smaller.
- The edge can operate even during times of (short or longer) unavailability of cloud services (e.g., due to network or cloud failures or intentionally, for mobile equipment like trucks which might only have internet connection from time to time, or also for privacy/security reasons) and therefore it improves availability and reliability of key services.
- Legacy devices or very resource-constrained devices that cannot establish their own cloud connections can be integrated into edge-based system architectures as well.
- Edge resources can be less expensive than cloud resources, and costs may be more predictable. Usually, edge sources add to a higher CAPEX, but have a lower OPEX if used in a continuous monitoring case.

Of course, cloud computing has benefits as well, for example, in terms of scalability and elasticity. But Edge computing is not to be understood as an alternative to cloud computing but as a complement to achieve additional capabilities. All in all, the emergence of edge computing and edge analytics opens many novel possibilities for industrial analytics.

3. Cement Strength Prediction

In the cement-making process, the quality of cement is influenced by many factors. Variability in raw feed and fuel types means that cement production is rarely constant. Continuous changes in plant operating points are needed to counter this variability, guided by online and lab measurements. A particularly important quality is “28-day cement strength.” If cement strength is too low, it cannot be sold, or requires blending with a higher-grade of cement. Conversely, cement strength that is too high comes at the cost of either reduced yield or excessive additives (which means excessive costs). Traditionally, there is a delay of 28 days until it is possible to test cement strength. By then it is too late to make corrections in the process. This means plants often over-deliver on the cement strength specification at reduced profit potential.

Having a way to predict the 28-day cement strength during production, preferably many times a day, could enable plants to respond in a much more timely manner. ABB tackles this challenge using machine learning functions within the ABB Ability™ Genix Industrial Analytics and AI Suite¹. Models were built in ABB Ability™ Genix using its Model Fabric component to map sampled data and predict 28-day cement strength, training and testing example shown in Fig. 1.

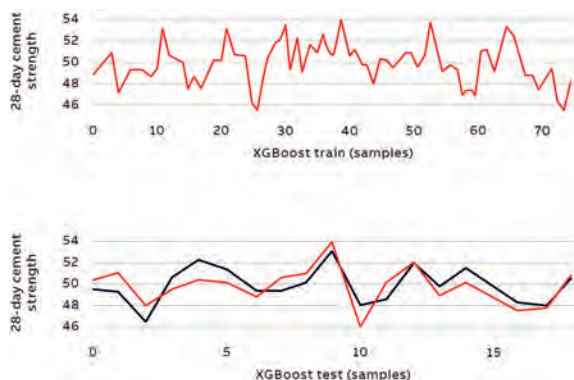


Fig. 1: Training (top) and Testing (bottom) results

¹ <https://new.abb.com/process-automation/genix>

Multiple proposed influencing features (Fig. 2) and models were reviewed in Genix Model Fabric for accuracy and robustness before a final model was selected. This model was then deployed on the edge to predict 28-day strength based on production data provided every 2-3 hours.

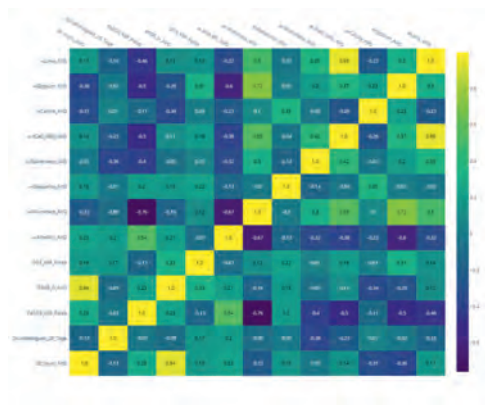


Fig. 2: Defining the selecting of features aided by correlation matrices
(bright = strong positive correlation, dark = strong negative correlation)

The ongoing accuracy of the on-premise model will be monitored before action is potentially transferred to another ABB technology: the ABB Ability™ Expert Optimizer. This technology facilitates advanced process control and automatically makes corrective changes to the production process with the aim to reduce variability and achieve a more constant 28-day cement strength closer to target. Expected benefits of achieving this target are increased yield with reduced additive consumption costs.

Inferential modelling is a technology and methodology employed in several industries where there exist similar challenges to cement strength – where qualities are critical to determining process performance and product salability but are often hard to measure online. These indicators of qualities and salability, like cement strength, are usually requiring sampling from the field several times per day with delays in results for decision making taking several hours due to lab processing. An example of an industry that has successfully used inferentials for over 2 decades is the energy industry focusing on refining oil and gas. Here, inferentials predict (with varying success) key qualities in distillation products such as Kerosene Freezing Point, Kerosene Flash Point and Light Diesel 90% Point which are sampled several times per day

with several hour lag times between lab updates. These inferentials rely on established machine learning methods such as regressions, neural networks and Principal Component Analysis. Success cases in this industry have been realized where the architecture is completely disconnected from any cloud connection and utilizes existing server OPC connectivity to plant control systems and lab interfaces. Modelling of inferentials is done periodically and offline, typically per year or longer, with manual model management and deployment.

4. Architecture of Cement Strength Prediction on the Edge

In order to deploy and install a solution for cement strength prediction, multiple requirements need to be taken into account. For example, the solution needs to be connected in an efficient and secure way with the relevant data source. Furthermore, it should be avoided to overload the network with lots of superfluous data. Moreover, customers often prefer solutions where critical data is kept in the plant or where data transfer into the cloud is well-controlled and secured. At the same time, the costs should be minimal and predictable. For these reasons, an edge analytics architecture is beneficial to running cement strength prediction applications.

Fig. 3 depicts an architecture of a solution to deploy analytics applications in an edge computing setting. In the figure, we distinguish between the training phase in the upper part of the picture and the actual operations phase in the mid and lower part of the picture. The training phase happens “offline” (but usually periodically). There, the cement strength training data usually stored in some data historian is used for training a cement strength prediction model in a cloud environment for development and management of AI/ML models. Such a cloud environment usually includes a model repository, as well as data exploration and training and testing functionality.

The resulting cement strength prediction model (or the best of multiple created models after validation, as explained in Section 3) is then deployed on an edge device and runs in the operations phase. The edge device is installed locally on the plant but is connected to the cloud for (long-term) storage of data and information models. Also, the management of (multiple) edge devices, including the registration and configuration of edge nodes and their applications, is usually done via a service running in the cloud. The edge device receives data, either as a batch of data or as a data stream via a connectivity module from devices or other local systems, e.g., control systems or, as in the cement example case, from a historian. These systems, again, may be connected to further systems, e.g., for advanced process control (see Section 3). Connectivity modules implement communication protocols to integrate with other local systems or devices. In the cement strength prediction data, the connec-

tivity module receives the latest cement values needed for the prediction. The cement strength prediction model is fed into a prediction app running on the edge device's compute engine and is triggered periodically, either time- or event-based. The computed prediction results can either be forwarded to the cloud layer, or stay in the plant, and e.g., be returned back to the originating data historian and connected systems.

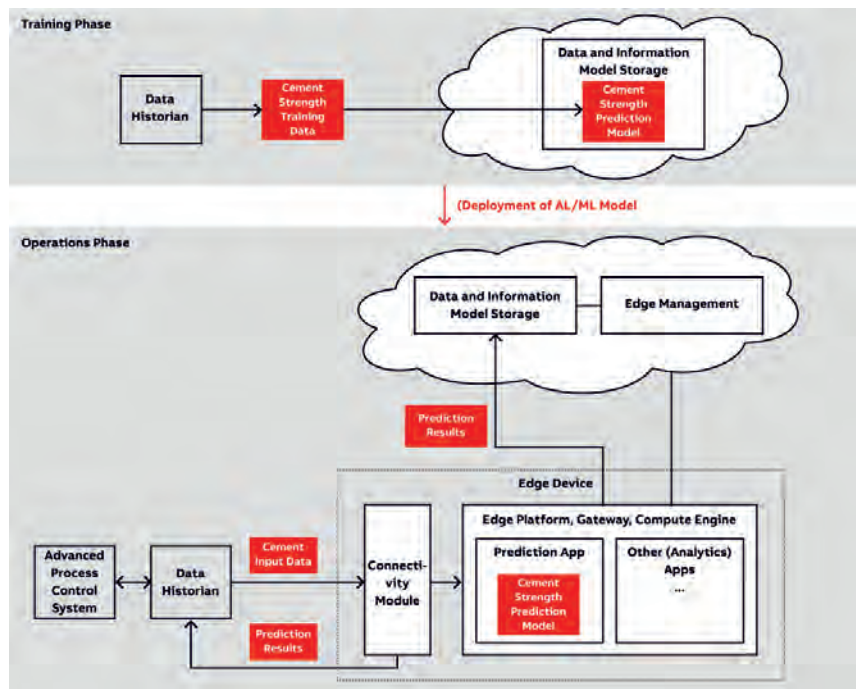


Fig. 3: Architectural Overview

The concrete implementation of such an edge solution ideally goes hand in hand with industrial standard reference architectures, for example, the NAMUR Open Architecture (NOA) [2]. One of the core ideas of NOA is to keep the Core Process Control (CPC) of a process plant unchanged and secured while operating additional (IoT) functionality, especially analytics applications, within the Monitoring and Optimization domain (M+O). Even though data and information flows from the process to the M+O modules, the different security zones ensure that the security goals of the core process automation are not compromised. Another benefit

of NOA is that a separation of the lifecycles of CPC and M+O modules is achieved. Thereby, faster updates can be achieved, and new or improved functionality can be integrated earlier.

One more reason to use the architecture described above for use cases like cement strength prediction is the self-service characteristic: The functionality runs locally on the edge, with all the benefits a local solution comes with, but it can be managed (installed, configured, updated) from the cloud. This also includes the possibility for the application of a continuous optimization process, where the ML model can be retrained on cloud level whenever needed e.g., when the prediction quality starts decreasing because the environmental settings changed. This so called “model drift” or “model degradation” can be overcome with “AIOps” or “MLOps” techniques [8,9]. A good edge computing platform supports the model to be seamlessly, remotely, and securely updated after a retraining without the need to redeploy the whole solution.

At ABB, the ABB Ability™ Genix Industrial Analytics and AI Suite² with its Edgenius edge component fulfills all these requirements and can be used. A prototype based on ABB systems has shown the feasibility of the described architecture [10].

5. Challenges and Open Issues / Future Research Directions

There are plenty of open research challenges in the area of edge analytics in process automation. In the following, we list a selection:

- As mentioned in Section 4, NOA provides an architecture to separate CPC applications and M+O applications. Many ML-based prediction solutions refer here only to a M+O application enabling to gain more insight into the current operation. Expert users always evaluate the solution before a change in the process is executed. In case of the 28-day cement strength prediction, the operator decides the trustfulness of the prediction before the APC is fed with this information. In future, a more integrated solution of the ML-based soft sensor into the APC concept needs to consider not only the NOA-required “Verification of Request” in order to allow ML-based closed-loop control, but also to ensure that operator trust the solution of the ML-based APC / the APC performance does not suffer from wrong prediction performance.
- In order to trigger a retraining of the AI/ML model, to keep the quality of the predictions high, model performance monitoring on the edge is needed. It is key for any

² <https://new.abb.com/process-automation/genix>

type of system that involves ongoing maintenance, but currently there is no standard way to achieve this. There exists the idea of "AIOps" or "MLOps" to include the monitoring task into the model management processes, similar to continuous integration and continuous deployment for DevOps done in the area of software engineering [8,9]. It needs to be evaluated to what extent such concepts can be applied in an edge analytics and a process automation setting.

- The current solution uses the cloud for the resource-consuming task of training the AI/ML model. In a complete offline setup, the training would need to be performed on the edge devices themselves, i.e., in a more resource-constrained setting. There are trends like "tinyML"³ as well as many optimization methods for AI/ML methods, that could be exploited, but the best solution depends a lot on the concrete characteristics of the edge node(s).
- Distributed edge architectures and analytics: The more devices an industrial plant has, the more the need increases for not only one or few edge devices, but for many edge nodes. These edge nodes could be arranged in different architectures, e.g., in an hierarchical setting or in a so-called edge mesh. Distributed analytics methods already exist, e.g., federated learning. It needs to be evaluated how distributed analytics methods can be used in more complex edge architecture and how secure and reliable data and communication management can be ensured in this case.
- When deploying analytics and other applications on an edge device, there are a lot of integration tasks, including functional characteristics as well as quality measures, to be cared for. This becomes even more critical if applications from third-party providers are to be integrated. However, this will become an essential requirement as there are many third-party providers for analytics applications and users could benefit from a big variety of solutions. Thus, standardization of interfaces for integration of third-party analytics apps for the edge is another field to focus on in future.
- In line with the previous item, also from a security point of view, there are many open questions on how to handle the development and deployment of analytics applications best, especially when targeting cross-company settings. These questions include not only how to secure the communication but also, for example, how to protect the intellectual property of the AI/ML models to be deployed. Developing edge analytics applications currently involves both, software engineering/ process engineering and data science know-how. To make this scenario easier, quicker, and more cost-

³ <https://www.tinyml.org/>

efficient, research in the area of “No-code/low-code” platforms for edge analytics apps could become beneficial. Such platforms are equipped with knowledge bases and smart strategies to make the development process more efficient and available to a larger user base.

- Reinforcement Learning involves the continuous experimentation and learning of models and actions to achieve maximum reward. In context of process industries reinforcement learning is considered as a step to increase the autonomy of plants. Practically, this requires a constant transfer from cloud to edge and update of local (action) models.

6. Conclusions

In this article, we discussed use cases and an architecture for an edge-based analytics solution with the process automation domain, illustrating edge analytics on an example use case cement strength prediction. We showed how an AI/ML model trained based on lab data from a cement site can be deployed on an edge computing architecture. The provided benefits include not only latency and bandwidth reduction but also security, privacy and remote management aspects. Thereby, an accelerated decision making can be achieved. Cement strength prediction is only one example, but there are many further use cases that can benefit from an edge analytics architecture.

Open challenges and research opportunities in this area include involving current trends like MLOps, tinyML, edge networks, and no-code/low-code platforms into the world of edge analytics and to evaluate how the process automation domain can benefit these tools in combination with edge-based analytics.

- [1] Platenius-Mohr, Abukwaik, Schlake, Vach: Software Architectures for Edge Analytics: A Survey, European Conference on Software Architecture 2021
- [2] User Association of Automation Technology in Process Industries (NAMUR): NE 175: NAMUR Open Architecture – NOA Concept. 2020, available at: <https://www.namur.net/en/focus-topics/namur-open-architecture.html> (last access: April 2022)
- [3] M. Satyanarayanan: The emergence of edge computing, IEEE Computer, Bd. 50, Nr. 1, pp. 30-39, 2017.

- [4] J. Perino, M. Littlefield und V. Murugesan: Living on the Edge: Edge Computing in the New OT Ecosystem, white paper, LNS Research, Stratus, 2020, available at: <https://resource.stratus.com/whitepaper/edge-computing-in-the-new-ot-ecosystem/> (last access: April 2022)
- [5] Satyanarayanan, Mahadev, et al.: Edge analytics in the Internet of Things. IEEE Pervasive Computing 14.2 (2015): 24-31.
- [6] The Open Industry 4.0 Alliance: Open Industry 4.0 Alliance Technical Solution Design Principles. white paper, version 1.0, 2019, available at: <https://www.automation.com/getattachment/d114d4ee-d9e2-426b-a788-a8f27991340a/OI4-Technical-Whitepaper.pdf?lang=en-US&ext=.pdf> (last access: April 2022)
- [7] User Association of Automation Technology in Process Industries (NAMUR): Reference architecture for Industrial Cloud Federation -- DIN SPEC 92222:2021-12, 2021
- [8] Masood, A., Hashmi, A.: AIOps: Predictive analytics & machine learning in operations. Cognitive Computing Recipes, Springer, 2019
- [9] Tamburri, D.A.: Sustainable MLOps: Trends and Challenges. 22nd Int. Symposium on Symbolic and Numeric Algorithms for Scientific Computing, IEEE, 2020
- [10] Platenius-Mohr, Schlake, Vach, Koorts, Trosten, Ninan: Cementing The Edge – Better decisions in cement strength prediction with artificial intelligence. ABB review, issue 02/2022, 2022 (to be published)

Approach to Data Pre-processing for Predictive Quality of Hydraulic Test Results in a Dynamic Manufacturing Environment

Concept drift approach with ensemble classification for abrupt shift correction

Christian Neunzig,

Bosch Rexroth AG, Ruhr-University Bochum;

Dennis Möllensiep, Simon Fahle, Prof. Dr.-Ing. **Bernd Kuhlenkötter,**
Ruhr-University Bochum;

Dr.-Ing. **Matthias Möller, Jürgen Schulz,**

Bosch Rexroth AG, Homburg

Abstract

Machine learning (ML) is rapidly becoming a competitive technology by addressing challenges such as increasing efficiency and monetizing data storage, but there are still some barriers that prevent mass deployment. Robust and reliable model predictions are essential for industrial applications. The robustness of the predictive models depends crucially on the quality of the data in terms of predictive power, purity, completeness, and equal boundary conditions for the occurrence of the data. System changes and volatile boundary conditions alter the data streams. In this paper, the variability of data streams over time is addressed using the concept drift framework. The concept drift framework consists of drift identification, drift understanding and model implementation. To account these changes in the model implementation, an ensemble classification of different importance metrics is applied. This approach achieves dimensionality reduction and noise filtering of relational discrete input variables, while increasing accuracy in the classification of internal leakage test results in the final inspection of directional control valves.

1. Introduction

Advancing digitalization and high computing power are drivers for the progressive use of ML methods in production use cases [1]. Given training data $\mathcal{D} = \{(x_e, y_e) \in \mathbb{R}^d \times \mathbb{R}\}$, where x_e is the input vector with the dimensionality d , and y_e is the target which are fed to an ML model. The problem to be solved can be described by a function with uncertainty. Examples e of the features or covariates x_e of a training set $\mathcal{D}$ are passed to a model [2]. The model generates

an output in response to the feature vector passed. The learning system modifies its feature-target relationship $\hat{f}$ with respect to the differences in $y_e = \hat{f}(x_e)$ between the generated and true output. In non-stationary manufacturing environments and product ranges with many variants, changes are frequently made to the process and product [3]. This inevitably leads to unpredictable variations in the statistical properties of data streams over time [4]. Along with this, patterns derived from previous data may not be relevant to the new data, leading to poor predictions and decision outcomes [5]. With the change in data patterns over time, it is neither practical nor feasible to restart the model development [6]. With smaller batch sizes and more frequent changes in the system, there is no alternative but to detect changes in the database and take them into account in the modelling, also referenced as adaptive learning [7]. The question is up to what point in the past can historical data be considered if the predictive model has become less accurate over time or which features are independent of those changes [8,9]. Changes in the system lead to a deterioration in the predictability of a target variable. The phenomenon was first proposed as concept drift by SCHLIMMER ET AL. [10]. Different terminologies for concept drift appear in the literature: dataset shift or concept shift [7]. Another way to classify the terms is to consider dataset shift as superordinate and covariate shift, prior probability shift and concept shift as subcategories of dataset shift [2]. In this paper, the suggestion of LU ET AL. will be followed to use only the term concept drift, since concept shift is usually related to covariate shift and prior probability shift, and an increasing number of publications describe the following problem in this way [7].

A major challenge with concept drift is that noisy data can change into non-noisy data at different times [7,11,12]. These changes can be caused by changes in hidden variables that cannot be measured directly. This means that input data for predicting a target variable has different predictive power or feature importance at different points in time. This challenge is also related to the very well-known effect curse of dimensionality [13]. A high dimensionality of features often overfitting of the training data by the algorithm, and not generalizing enough [14]. The final goal in developing a predictive model is to achieve the greatest possible generalizability while reaching the global maximum of predictive accuracy for the target variable. Major challenges of this use case are the large class imbalance and high dimensionality of the input data. The change in the high dimensional input features manifests itself in abrupt, gradual, incremental, or reoccurring drifts over time, which need to be detected and considered in the modelling [15,16]. In this work, the concept drift framework is applied to serial production data of electro-hydraulic directional control valves to encounter this problem. This paper addresses the following research question:

How can data shifts be detected and accounted for in modelling to classify test decisions for a highly imbalanced target variable of directional control valves based on high-dimensional discrete quality data?

2. Problem definition of the industrial use case

Based on real world production data along the value chain of directional control valves from Bosch Rexroth, the physically unavoidable internal leakage between housing and control spool is predicted using machine learning methods. In this industrial use case, the quality prediction of the quality characteristic of the directional control valves manufactured in Homburg is achieved by using the geometric final dimensions from machining, pairing data from assembly, and hydraulic sensor data from end-of-line testing and assigning them to a valve with a time stamp [17]. The test decision is considered a classification problem. The upper limit of leakage (target) is ensured in one of more than 60 test steps as a safety-critical product characteristic within the scope of the final hydraulic inspection [18,19]. In total, more than 1000 alphanumeric, categorical, and time-dependent input covariates are available to the prediction model. The leakage measurement takes place indirectly in a pressure drop measurement, which is reduced by the leakage volume flow through the ring gap into the tank, see Fig. 1 [18]. The target variable is subject to some uncertainty because it depends on state variables of the process and manufacturing tolerances that are not fully captured. In this respect, it is very likely that the immediately prior states to testing also influence the prediction, which makes it impractical to shuffle the data [19]. In addition, regular and irregular changes in the production environment result in concept drifts in the data streams.

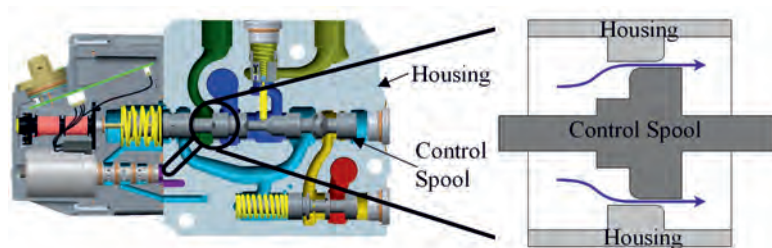


Fig. 1: 2-D cross-section of a ring gap in a directional control valve cf. [18].

Stable production processes enable companies to achieve high good part rates. When using data-driven predictive analytics, the imbalanced ratio of the target also becomes a problem in this use case. The good part/bad part ratios for different time intervals of the total data set are displayed in Table 1. The determination of the number and limits of the time intervals is

explained in chapter 4. The data points grow cumulatively with each time interval. The total number of data points does not increase with the same number of data points per time interval. The same applies to the ratio of good parts (0) to bad parts (1). The proportion of the good part class (0) varies between 96.34 % and 98.55 % in the training data set and between 92.68 % and 99.04 % in the test data set. In summary, there is a strong imbalanced data set with fluctuations in the class ratios in the training and test data set for different time intervals.

Table 1: Overview of class imbalance in production dataset for different time intervals.

Time Interval	Training Data		Test Data	
	class distribution	class ratio	class distribution	class ratio
1	(0):2725, (1):46	98.33%:1.67%	(0):1243, (1):12	99.04%:0.96%
2	(0):3968, (1):58	98.55%:1.45%	(0):1546, (1):122	92.68%:7.32%
3	(0):5514, (1):180	96.83%:3.17%	(0):2058, (1):40	98.09%:1.91%
4	(0):7572, (1):220	97.17%:2.83%	(0):1414, (1):49	96.65%:3.35%
5	(0):8986, (1):269	97.09%:2.91%	(0):1610, (1):100	94.15%:5.85%
6	(0):10596, (1):369	96.63%:3.37%	(0):1826, (1):102	94.70%:5.30%
7	(0):12422, (1):471	96.34%:3.66%	(0):2275, (1):46	98.01%:1.99%
8	(0):14697, (1):517	96.60%:3.40%	(0):601, (1):9	98.52%:1.48%

Within the data set, significant abrupt shifts occur in different intervals. These changes in the dataset significantly affect the predictive power of the input variables and cause the most important input features to fluctuate strongly for different time intervals. This not only complicates the interpretability of the predictions, but also the further pre-processing of the model. As an example, two concept drifts in features over time are shown in Figure 2.

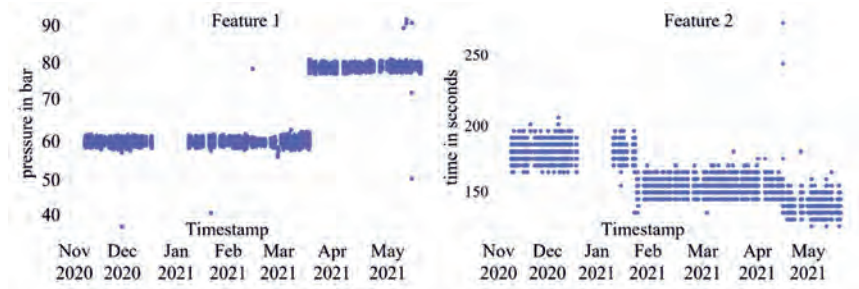


Fig. 2: Concept drift of different features over time cf. [17].

Overall, the challenges of the present production dataset can be summarized as follows: high-dimensional, class-imbalanced, and volatile over time.

3. Related works

The concept drift framework consists of concept drift detection, concept understanding and model implementation [7]. The methods and algorithms for drift detection are classified into three categories based on the test statistics used: error rate-based, data distribution-based and multiple hypothesis test drift detection [11]. Error rate-based detection is the most used [7]. One that is also based on this group of detection is proposed by TRUONG ET AL. They suggest the sliding window method with binary segmentation known from signal processing which is used and defined below [20]: Let $\{s_t\}_{t \in \mathcal{T}}$ denote the recorded values of a variable of the data set indexed to the associated time stamps $\mathcal{T} = \{t_1, t \dots, t_m\}$. Furthermore, let $s_{a:b} \doteq \{s_t\}_{t \in \mathcal{T} \mid a < t \leq b}$ be the partial section of the complete measurement series limited by the time points a and b and let t be a dummy time point smaller than t_m . The aim of binary segmentation is to group the measurement series into disjoint subsections with the help of intermediate points k , so that the sum of the absolute deviation of the sections from their median $\tilde{s}_{t_{k_i}:t_{k_j}}$ becomes minimal, see formula (1):

$$\min \left\{ \sum_{t_0 < t \leq t_{k_1}} |s_t - \tilde{s}_{t_0:t_{k_1}}| + \sum_{t_{k_1} < t \leq t_{k_2}} |s_t - \tilde{s}_{t_{k_1}:t_{k_2}}| + \dots + \sum_{t_{k_{n-1}} < t \leq t_{k_n}} |s_t - \tilde{s}_{t_{k_{n-1}}:t_{k_n}}| \right\} \quad (1)$$

The use of the absolute value and the median increases the robustness of the method against outliers compared to the use of mean and squared deviation. Since the number of intermediate points is not known a priori, starting from the complete measurement series, the intermediate point for further segmentation is determined step by step in a greedy manner, which reduces maximally until a stop criterion is reached. As a stop criterion, it should be implemented that the difference in amount of the medians of the newly found subsections must be greater than the minimum of the empirical standard deviation of both subsections. This means that if a division of $\tilde{s}_{t_{k_i}:t_{k_j}}$ in $\tilde{s}_{t_{k_i}:t_{k_l}}$ and $\tilde{s}_{t_{k_l}:t_{k_j}}$ is planned in the current iteration step, it is terminated if:

$$|\tilde{s}_{t_{k_i}:t_{k_l}} - \tilde{s}_{t_{k_l}:t_{k_j}}| \leq \min \left\{ \sqrt{\widehat{Var}(s_{t_{k_i}:t_{k_l}})}, \sqrt{\widehat{Var}(s_{t_{k_l}:t_{k_j}})} \right\} \quad (2)$$

Concept understanding refers to accessing information about the time of occurrence and duration of the concept drift, the severity of the concept drift and the drift areas of the concept drift. This status information is the output of the drift detection algorithms and is used as input for drift adjustment. Process documentation is often required to explain the status information. [7]

For the model implementation of the concept drift framework, two methods for solving imbalanced learning problems have been described in the literature: Ensemble methods and algorithmic classifier modifications [15,21,22]. Ensemble-based classifiers are widely used in the problem of imbalanced classification since they improve the performance of a single classifier by combining multiple base classifiers that outperform each independent classifier [15]. Non-stationary data streams are challenging since noisy data can transition to non-noisy data at different times. Thus, the feature importance values fluctuate over time. [7,11,12] By using ensemble classification of subsets, stable features can be derived from the individual subsets [8,9,14]. For small samples, feature stability is as important as feature predictive power [9]. The stability of a feature is characterized by the fact that it is important regardless of time and has predictive power for the target variable. [9] Figure 3 shows a matrix $\mathcal{A}_t$ that uses different classifiers c for assessing importance for different subsets s_1 to s_N of features $x_{c,1}$ to $x_{c,d}$ [8].

$$\mathcal{A}_t = \begin{pmatrix} s_1 \\ s_2 \\ \dots \\ s_N \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{1,2} & \dots & \dots & x_{1,d} \\ x_{2,1} & x_{2,2} & \dots & \dots & x_{2,d} \\ \dots & \dots & \dots & \dots & \dots \\ x_{N,1} & x_{N,2} & \dots & \dots & x_{N,d} \end{pmatrix}$$

Fig. 3: Evaluation matrix for the different subsets based on a set of candidate features cf. [8].

4. Application of the concept drift framework using ensemble classification

In this section, the concept drift framework is applied to the industrial use case. In the following, the concept drifts are detected by the binary segmentation approach of TUONG ET AL. in consideration of five data points of a feature dimension per iteration. Since the shifts in the present application can be attributed to abrupt changes in the test parameters, it is sufficient to examine abrupt shifts. Although the way in which other test steps are recorded changes, the underlying physical relationships still hold. By grouping concept drifts, eight time intervals of comparable boundary conditions were identified, which are displayed in Figure 4 with the data points per interval for training-test split.

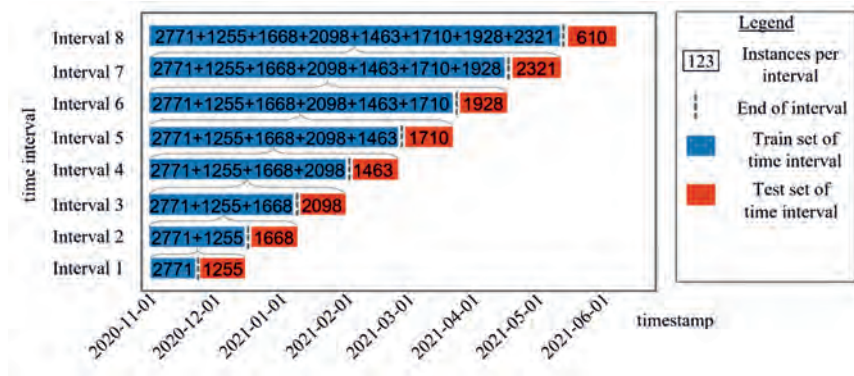


Fig. 4: Timeseries 10-fold-cross-validation for 8 identified intervals of equal boundary condition.

In total, 54 different points in time could be identified with this method at which changes could potentially have occurred. The trade-off is to detect as many concept shifts as possible, but to combine them into the smallest possible amount of time intervals to achieve the largest possible number of data points per time interval. An insufficient amount of data points per time interval prevents meaningful model training or deteriorates prediction accuracy. Of the 54 drifts, 16 were manually classified as relevant by using domain knowledge. This affects 75 features, as concept deviations can be noticed in several measurements.

When concept shifts occur, changing boundary conditions of the system are analyzed. The cause of some concept drifts can be explained by manual interventions such as replacing machine or product components, shortening test steps or changing test parameters. One challenge is to identify different severity of change in the system. This could not yet be solved in this approach. Nevertheless, the time-dependent data variability can be incorporated by using a timeseries 10-fold cross-validation per time interval. Thus, the training data of past time intervals are cumulated, and the data of the most recent time interval are used as an unseen test data set true to reality. The comparability of different time intervals is furthermore supported by the centralization of features, as already demonstrated by NEUNZIG ET. AL. [18]. Due to the concept drift in the high-dimensional data set and the imbalanced target ratio, strongly deviating feature importance ratings are obtained for different time intervals. Accordingly, an ensemble classification approach is applied to develop a feature list that achieves the greatest possible accuracy over all time intervals in a main model for classifying the target variable. The proposed extended approach for the model implementation of the concept drift for data preprocessing consists of six essential steps:

- 1) The entire production data set will be divided into sections with stationary data streams. These are characterized by the same boundary conditions, i.e. no more abrupt data shifts are detected within these sections and the comparability of the data within the time intervals is increased. The candidates of the time intervals are detected in the concept drift detection and summarized to few intervals in the concept drift understanding. [8,9]
- 2) Four models with different feature importance metrics are trained on all detected time intervals from (1). The training strategy used is timeseries 10-fold cross-validation for each time interval, see Figure 4. The models are each hyperparameter-optimized with the in-built Pycaret tuning function to obtain meaningful results. In principle, there are different options for choosing the different models. The choice for the combination of algorithms categorical boosting (catBoost), extreme gradient boosting (XGBoost), light gradient boosting machine (LGBM), random forest (RF) is justified by the fact that tree-based methods are particularly suitable for high-dimensional data sets, low computational costs and high parallelization [22,23]. These candidates of tree classifiers have performed particularly well in preliminary studies [18].
- 3) Feature Importance scores for each of the four models are determined for all time intervals and sorted in descending order of importance.
- 4) Based on the calculated feature importance ranking, all four feature importance scores for each feature are ranked according to whether it is among the x_{bt} number of features within a time interval. If the feature is ranked below the x_{bt} decision boundary, this feature is classified as 1 in this time interval, otherwise it is classified as 0. This results in eight matrices of the form $\mathcal{A}_t$ with $t \in \{1, 2, \dots, 8\}$, as defined in Figure 3. For the preliminary study of x_{bt} , it has been shown that $x_{bt} = 200$ is a sensible definition since this selection covers about 90 percent of the variance of the input vector. Due to the concept drift, the identified time intervals are used for the subsets. As a result, the subsets are not of equal size. An exemplary classification of a time interval is shown in Table 2. This is carried out for all 8 time intervals.

Table 2: Exemplary assessment of feature importance for one of the eight time intervals.

	Feature 1	Feature 2	...	Feature 200	...	Feature 292
XGBoost	0	1	...	0	...	1
CatBoost	1	1	...	1	...	1
LGBM	0	1	...	1	...	0
RF	1	1	...	1	...	0

5) The resulting matrices $\mathcal{A}_i$ must be evaluated to obtain a list of features. In terms of ensemble classification, the first approach is to add the matching positive decision for a feature from all classifications used into a list for a subset. In this application, this means that a feature is added to the feature list for a time interval if it counts among the top 200 of the total 293 features each for the classifiers XGBoost, CatBoost, LGBM and RF. This results in eight lists of so-called stable features. In addition, a majority vote of at least three out of four classifiers with positive classification is also conceivable. This will be discussed further in chapter 5.

6) As a final step, the feature lists obtained are fed into a main model to compare the accuracy of the different models in the different subsets and the overall dataset. Here, the accuracy on the overall dataset is especially crucial. The course of the accuracy over all subsets provides information about which subsets have local minima or maxima. With domain knowledge, the causes of this can be investigated. In this use case, XGBoost classifier was selected as the main model because it performed very well in preliminary studies [18].

5. Experimental results

To validate the advanced procedure for identifying stable features by means of ensemble classification of subsets with the same boundary conditions, a baseline model is used, which is trained with the same training strategy (10-fold timeseries cross validation) and which has the complete scope of the cleaned feature vector (number 293) preprocessed according to [17,18,19] available at all time intervals. Since the baseline model has more information available, the hypothesis is that it only performs worse if there is noise in the feature vector.

Applying the strict criterion that a feature is stable if all four classifiers agree, the main model is completely underrepresented. The average number of features per feature list is only 50 out of the original 293 features. Figure 5 shows the line plot of the F1-score for the advanced and baseline approach only for feature list of interval 3.

There is an increasing difference in F1-score for the training dataset. Furthermore, the baseline model performed better than the advanced model for the entire test data set (four out of four). This relationship is also observed in all other seven time intervals not shown here. Thus, the advanced approach with majority voting (at least three out of four classifiers) is introduced to evaluate the feature lists to represent a larger variance of the feature vector. In Figure 6 the line graph of the F1-score of this approach for feature list from interval 3 converges to that of the baseline.

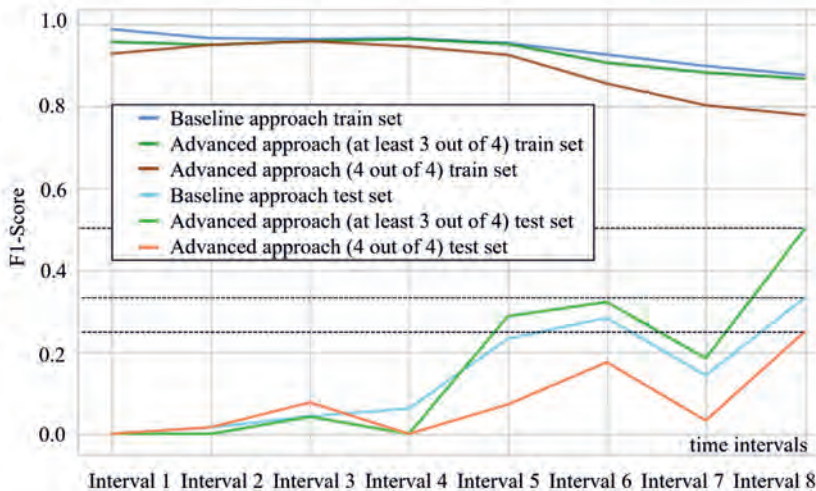


Fig. 5: Line plot of prediction accuracies for two advanced and a baseline approach with interval 3 features.

There is no underrepresentation in this approach and the accuracy on the entire test data set is also clearly superior. The same tendencies can be observed for the other feature lists, but this is not shown here. Since the majority decision (three out of four) produces significantly more accurate predictions, it is examined more closely, see Figure 6. Here, the min-max-range of the F1-scores for all feature lists without list 3 and the line plots show the baseline and advanced approach. The advanced approach with feature list 3 exceeds the remaining feature lists globally and predominantly locally. With an F1-score of about 50 %, the use of the stable features from interval 3 produces by far the most accurate predictions.

In general, for all models in Figures 5 and 6, the prediction accuracy increases with the number of training data. The strong accuracy drop in interval 7 was investigated and explained by domain knowledge with a particularly strong intervention in the process. Here, a test step was shortened, which does not fundamentally change the dependencies between the test steps, but to the extent that it had not been learned before. The degree of concept drift was not carried out in this work due to incomplete documentation and thus challenging explicability of the degrees of concept drift.

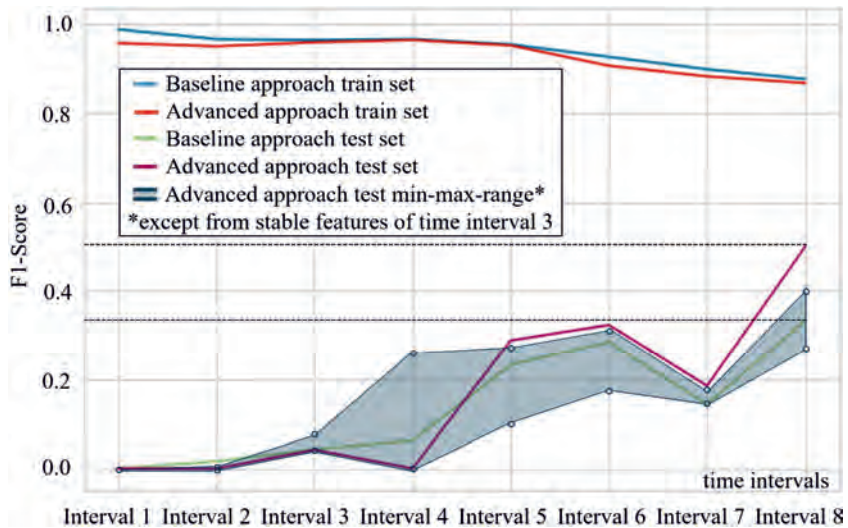


Fig. 6: Line plot of prediction accuracies for advanced (3 out of 4) vs. baseline approach with stable features from time interval 3 including min-max-range of accuracy for the rest of stable features for (3 out of 4) approach.

6. Conclusions and future work

For the prediction of hydraulic test steps of directional control valves, a classification problem with high-dimensional feature vector and imbalanced target vector is present as well as a temporal variability of the database. The hypothesis is confirmed that the application of the concept drift framework with ensemble classifiers leads to stable feature lists. This eliminates noise in the database and increased the accuracy by about 15 percentage points compared to the baseline model with all covariates to an F1-score of 50% for the unseen test data set. Although less features are fed into the model, the predictions are more accurate. The results are encouraging and should be validated by a larger sample size.

Future work should focus on investigating the severity of concept drift and how to weight different degrees of severity in the model implementation. In addition, future work will focus on expanding the current database using synthetic data generation. In terms of knowledge storage and adaptive learning, the solutions developed will be made available in the form of a human-centered assistance system for production engineers to carry out machine learning projects with tabular production data sets.

References

- [1] Liu, J., Cao, X., Zhou, H., Li, L., Liu, X., Zhao, P., Dong, J., 2021. A digital twin-driven approach towards traceability and dynamic control for processing quality. *Advanced Engineering Informatics* 50, pp. 1–17.
- [2] Moreno-Torres, J.G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N.V., Herrera, F., 2012. A unifying view on dataset shift in classification. *Pattern Recognition* 45 (1), pp. 521–530.
- [3] Lughofer, E., Pollak, R., Zavoianu, A.-C., Pratama, M., Meyer-Heye, P., Zörrer, H., Eitzinger, C., Haim, J., Radauer, T., 2018. Self-adaptive evolving forecast models with incremental PLS space updating for on-line prediction of micro-fluidic chip quality. *Engineering Applications of Artificial Intelligence* 68, pp. 131–151.
- [4] Bergler, S., Alaiz-Rodríguez, R., Japkowicz, N., 2008. *Advances in artificial intelligence: Assessing the Impact of Changing Environments on Classifier Performance*. Springer, pp. 23–34.
- [5] Zenisek, J., Holzinger, F., Affenzeller, M., 2019. Machine learning based concept drift detection for predictive maintenance. *Computers & Industrial Engineering* 137, pp. 1–13.
- [6] Wang, L.-Y., Park, C., Yeon, K., Choi, H., 2017. Tracking concept drift using a constrained penalized regression combiner. *Computational Statistics & Data Analysis* 108, pp. 52–69.
- [7] Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., Zhang, G., 2018. Learning under Concept Drift: A Review. *IEEE Transactions on Knowledge and Data Engineering*, pp. 2346–2374.
- [8] Nogueira, S., Brown, G., 2016. Measuring the Stability of Feature Selection, *Machine Learning and Knowledge Discovery in Databases*, pp. 442–457.
- [9] Nogueira, S., Sechidis, K., Brown, G., 2018. On the Stability of Feature Selection Algorithms. *Journal of Machine Learning Research* 18 (2018), pp. 1–54.
- [10] Schlimmer, J.C., Granger, R.H., 1986. Incremental learning from noisy data. *Machine Learning* 1 (3), pp. 317–354.
- [11] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., Bouchachia, A., 2014. A survey on concept drift adaptation. *ACM Computing Surveys* 46 (4), pp. 1–37.
- [12] Lin, C.-C., Deng, D.-J., Kuo, C.-H., Chen, L., 2019. Concept Drift Detection and Adaption in Big Imbalance Industrial IoT Data Using an Ensemble Learning Method of Offline Classifiers. *IEEE Access* 7, pp. 56198–56207.
- [13] Frasconi, P., Landwehr, N., Manco, G., Vreeken, J., 2016. *Machine Learning and Knowledge Discovery in Databases: Measuring the Stability of Feature Selection*. Springer International Publishing, pp. 442–457.

- [14] Kalousis, A., Prados, J., Hilario, M., 2007. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and Information Systems* 12, pp. 95–116.
- [15] Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G., 2017. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications* 73, pp. 220–239.
- [16] Ren, S., Zhu, W., Liao, B., Li, Z., Wang, P., Li, K., Chen, M., Li, Z., 2019. Selection-based resampling ensemble algorithm for nonstationary imbalanced stream data learning. *Knowledge-Based Systems* 163, pp. 705–722.
- [17] Neunzig, C., Fahle, S., Schulz, J., Möller M., Kuhlenkötter, B., 2022. Model Selection for Predictive Quality in Hydraulic Testing. Accepted paper in publication. 55th CIRP Conference on Manufacturing Systems 2022.
- [18] Neunzig, C., Fahle, S., Schulz, J., Möller M., Kuhlenkötter, B., 2022. Approach To A Decision Support Method For Feature Engineering Of A Classification of Hydraulic Directional Control Valve Tests. Accepted paper in publication. Conference on Production Systems and Logistics 2022.
- [19] Neunzig, C., Fahle, S., Kuhlenkötter, B., Möller, M., 2021. Feature Engineering For A Cross-process Quality Prediction Of An End-of-line Hydraulic Leakage Test Using An Experiment Sample, Conference on Production Systems and Logistics 2021, pp. 1–10.
- [20] Truong, C., Oudre, L., Vayatis, N., 2020. Selective review of offline change point detection methods. *Signal Processing* 167 (4), pp. 1–58.
- [21] Farid, D.M., Zhang, L., Hossain, A., Rahman, C.M., Strachan, R., Sexton, G., Dahal, K., 2013. An adaptive ensemble classifier for mining concept drifting data streams. *Expert Systems with Applications* 40 (15), pp. 5895–5906.
- [22] Shan, J., Zhang, H., Liu, W., Liu, Q., 2019. Online Active Learning Ensemble Framework for Drifted Data Streams. *IEEE transactions on neural networks and learning systems* 30 (2), pp. 486–498.
- [23] Lughofer, E., Sayed-Mouchaweh, M., 2019. Predictive Maintenance in Dynamic Systems. Springer International Publishing, pp. 64–93

Methoden der künstlichen Intelligenz für die automatisierte Planung von modularen Produktionsprozessen

Vergleich bestehender Ansätze und Vorstellung einer Benchmark

R. Heesch, N. Widulle, A. Köcher, A. Nordhausen,
L. M. Vieira da Silva, J. Putzke, O. Niggemann,
Helmut-Schmidt-Universität Hamburg

Kurzfassung

Eine Vielzahl von Anforderungen verlangt von produzierenden Unternehmen, dass diese ihre Produktion mit möglichst geringem Aufwand an sich ändernde Gegebenheiten, wie zum Beispiel neue Produktspezifikationen, anpassen können. Zur einfachen Adaption und Rekonfiguration von Anlagen sind maschinenlesbare Beschreibungen der einzelnen Anlagen-Module und der Fähigkeiten dieser Module bereits Gegenstand der Forschung. Darüber hinaus müssen aber auch die Produktionsplanungs-Ansätze entsprechend flexibel sein und mit den Möglichkeiten, die sich aus maschinenlesbaren Beschreibungen ergeben, umgehen können.

In diesem Beitrag wird analysiert, inwiefern sich bestehende Planungsansätze der *Künstlichen Intelligenz* zur Produktionsplanung in flexiblen Produktionssystemen eignen. Dazu werden zunächst Anforderungen an Planungsansätze für die Produktionsplanung in flexiblen Anlagen vorgestellt. Außerdem wird eine *Benchmark* vorgestellt. Diese umfasst ein Anwendungsszenario aus der flexiblen Produktion sowie verschiedene Problemstellungen, für die Produktionspläne generiert werden müssen. Mögliche Lösungen sind ebenfalls Teil des Benchmarks, so dass diese für das Testen neuer Planungsverfahren und den Vergleich verschiedener Verfahren genutzt werden kann. Der Beitrag stellt außerdem zwei erste Ansätze für Planungsverfahren aus unterschiedlichen Bereichen der Künstlichen Intelligenz vor.

Einleitung

Aufgrund von sich immer häufiger wandelnden Kundenanforderungen ist das heutige Produktionsumfeld geprägt von kurzen Produktlebenszyklen und geringen Losgrößen. Um in diesem Umfeld wettbewerbsfähig zu bleiben, ist es wichtig, dass Unternehmen ihre Produktionsprozesse schnell an sich wandelnde Anforderungen anpassen können [1].

Eine Klasse von Produktionssystemen, die für solch schnelle Anpassungen besonders geeignet sind, sind Cyber-physische Produktionssysteme (CPPS). Typischerweise bestehen CPPS aus unterschiedlichen Modulen, die miteinander kombinierbar sind und somit an sich ändernde

Rahmenbedingungen angepasst werden können [2]. Um das Potential dieser Ansätze vollständig zu nutzen, muss jedoch auch die Produktionsplanung, die festlegt auf welche Art und Weise ein Produkt hergestellt wird, automatisiert werden. Zu diesem Zweck ist es notwendig, die Informationen über die einzelnen Module und deren Funktionen maschinenlesbar zu beschreiben.

In der Forschung werden semantische Beschreibungsmittel wie bspw. Ontologien zur Beschreibung der *Fähigkeiten* und *Skills* von Maschinen untersucht. Fähigkeiten stellen dabei eine abstrakte Beschreibung der Funktionen von Ressourcen dar, wohingegen Skills Implementierungen von ausführbaren Funktionen bezeichnen, die beispielsweise über OPC UA aufgerufen werden können [3, 4]. Geht man von einer fähigkeitsbasierten Beschreibung von Maschinen aus, so muss im Rahmen der Produktionsplanung mit einer Menge an vorhandenen Fähigkeiten ein Produktionsprozess geplant werden, der gewisse angeforderte Fähigkeiten umsetzt oder die Produktion eines angeforderten Produktes garantiert.

Im Kontext der Künstlichen Intelligenz (KI) bezeichnet ein Planungsproblem im Allgemeinen das Finden einer Sequenz von Aktionen, die von einem Ausgangs- zu einem Zielzustand führt [5]. Der Ausgangszustand wird in diesem Zusammenhang durch die Eingangsprodukte, der Zielzustand durch das gewünschte Produkt und die Aktionen durch die Fähigkeiten der Module beschrieben. Im Rahmen dieser Ausarbeitung wird untersucht, welche KI-basierten Planungsmethoden für die automatisierte Planung von CPPS geeignet sind. Das Ziel ist dabei das automatisierte Erstellen eines Produktionsplanes, in Form von Fähigkeiten der zur Verfügung stehenden Module.

Zu diesem Zweck wird in einem ersten Schritt ermittelt, mit welchen besonderen Anforderungen das Planen solcher Produktionsprozesse verbunden ist. Anschließend werden verschiedene bestehende KI-basierte Planungsansätze kurz eingeführt, bevor sie anhand der definierten Anforderungen einander gegenübergestellt werden. Da es zurzeit keine beispielhaften Planungsprobleme für CPPS gibt, die in der Lage sind, die definierten Anforderungen abzubilden, stellen wir anschließend eine eigens entwickelte *Benchmark für die Produktionsplanung von CPPS* vor. Abschließend geben wir einen kurzen Ausblick darauf, wie sich zwei unterschiedliche Methoden der KI Planungsansätze für CPPS umsetzen lassen.

Anforderungen

Die Produktionsplanung für modulare Produktionssysteme ist mit einigen Herausforderungen verbunden, die sie von anderen Domänen unterscheiden. Aus diesen Herausforderungen haben wir sechs Anforderungen abgeleitet, die ein Planungsansatz in diesem Anwendungsgebiet erfüllen muss.

Optimierung: Im Produktionsumfeld reicht es oft nicht aus, lediglich herauszufinden, ob ein gültiger Plan existiert. Stattdessen müssen Planungsansätze ebenfalls in der Lage sein, Pläne hinsichtlich verschiedener Kriterien, wie zum Beispiel dem Energieverbrauch oder der Durchlaufzeit, zu optimieren.

Abhängigkeiten: Die Beziehungen zwischen den In- und Output-Produkten eines Produktionsschrittes sind oft komplex und können bspw. mit Hilfe explizit modellierter Verknüpfungen oder auch mittels mathematischer Funktionen abgebildet werden. Ein Planungsansatz für das Produktionsumfeld muss in der Lage sein, solche Beziehungen abzubilden und damit zu arbeiten.

Schleifen: Um den gewünschten Effekt zu erzielen ist es mitunter notwendig, dass Produktionsschritte mit der gleichen Konfiguration wiederholt ausgeführt werden. Die meisten symbolischen Planungsansätze beschränken sich jedoch auf sequentielles Planen, da das Lösen von Planungsproblemen mit Schleifen deutlich komplexer ist. Um beispielsweise das Erstellen eines modifizierten Plans zu erleichtern ist es notwendig, dass ein Planungsansatz in der Lage ist, solche Wiederholungen als Schleifen zu erkennen.

Nachvollziehbarkeit: Eine elementare Voraussetzung für den Einsatz von KI-Ansätzen für das Planen von Produktionsprozessen ist das Vertrauen von beteiligten Personen in Planungsansätze und deren Ergebnisse. Die Entscheidungen eines Planungsansatzes müssen dementsprechend nachvollziehbar und wiederholbar sein.

Aufwand: Für die Wirtschaftlichkeit eines Planungsansatzes ist es notwendig, dass der Aufwand für das Konfigurieren des Planers geringer ist, als der Aufwand für das händische Erstellen des Plans. Daraus leiten sich ein möglichst geringer Aufwand sowie möglichst wenig benötigtes Vorwissen für das Anwenden des Planungsansatzes ab.

Datenintegration: Produktionsprozesse sind komplexe Prozesse, die mit einer Vielzahl von unterschiedlichen, voneinander abhängigen Daten verbunden sind. Diese Daten können Einschränkungen von Produktionsplänen darstellen. Ein Planungsansatz muss folglich in der Lage sein, Daten, wie zum Beispiel Produktionsdaten aus einem Manufacturing Executive System, in den Planungsprozess zu integrieren, um ausführbare Pläne erstellen zu können

Übersicht und Vergleich bestehender Ansätze

Im Folgenden beschreiben wir bestehende Ansätze zur automatisierten Planung von Produktionsabläufen aus zwei Teilgebieten der KI.

Symbolische Planungsansätze

Klassische KI-Planungsansätze basieren auf symbolischen Methoden, wie zum Beispiel Suchalgorithmen oder logischen Verfahren. Im Folgenden werden zwei dieser Ansätze, die *Planning Domain Definition Language* (PDDL) sowie Planungsansätze, die auf Erfüllbarkeitsproblemen basieren, näher betrachtet.

PDDL wurde 1998 im Rahmen der *International Planning Competition* (IPC) entworfen, um Planungsprobleme aus dem Bereich *Classical Planning* zu beschreiben [6]. Seit der Einführung von PDDL wurde die Sprache mit verschiedenen neuen Versionen erweitert, um beispielsweise die Betrachtung von Zeit zu ermöglichen [7]. Für die unterschiedlichen PDDL-Versionen wurden, üblicherweise im Rahmen der IPC, viele verschiedene Solver entwickelt. Diese Solver unterstützen dabei jedoch oftmals nur Teilmengen der Eigenschaften der verschiedenen PDDL-Versionen [8]. Die Beschreibung eines Problems und der möglichen Aktionen ist in PDDL recht intuitiv möglich. Die Sprache ist außerdem sehr verbreitet und wird von einer aktiven Community weiterentwickelt. Dennoch ist die Sprache nicht ausdrucksstark genug, um die Komplexität der realen Welt abzubilden [8].

Die PDDL-Solver basieren zum Großteil auf Deduktion. Alternativ dazu können Pläne jedoch auch dadurch gefunden werden, dass der (erfolgreiche) Plan eine Reihe von gegebenen Gleichungen erfüllt. Sind die Gleichungssysteme in boolescher Algebra gegeben, können für die Lösung sogenannte SAT-Solver eingesetzt werden. Die Grundlage für den Ausdruck von Planungsproblemen in dieser Form wurde in [9] präsentiert. *Satisfiable Modulo Theory* (SMT) ergänzt die booleschen Ausdrücke um sogenannte Theorien, die beispielsweise Integer, Arrays oder Fließkommazahlen hinzufügen. Es existiert eine Vielzahl an effizienten SMT-Solvern, welche durch unterschiedliche Schnittstellen sowie eine standardisierte Sprache namens SMT-Lib angesprochen werden können [10].

Die Formulierung von Planungsproblemen in einer für diese Solver geeigneten Form ist nicht trivial. Ein erster Formalismus wird in [9] vorgestellt.

Subsymbolische Planungsansätze

In Bezug auf Planungsaufgaben verwenden Algorithmen des Maschinellen Lernens (ML) Daten, um einen optimalen Plan, Heuristiken zum Lösen eines Problems oder funktionale Abhängigkeiten zwischen Prozessschritten zu finden.

ML wird typischerweise in die Kategorien *Supervised Learning* (SL), *Unsupervised Learning* (UL) und *Reinforcement Learning* (RL) unterteilt. Ersteres verwendet Domänenwissen in Form von gelabelten Datensätzen zum Training von Modellen. Hierzu muss beispielsweise die Domäne, im Anwendungsfall von Shen et al. [11], durch eine STRIPS-Beschreibung dargestellt

sein. Die Gewichtungen werden iterativ unter Berücksichtigung von Vorwissen angepasst. Weitere Ansätze, wie das GGSNN [12], verzichten auf eine solche Beschreibung und sind hierdurch generalisierbarer darstellbar. Bekannte SL-Ansätze sind bisher lediglich für wenig komplexe Planungsprobleme einsetzbar, welche nicht den Bedingungen von realen Prozessen entsprechen.

UL versucht ungelabelte Datensätze zu verwenden, um Planungsprobleme zu lösen. Ursprung ist hierbei die Tatsache, dass in Produktionssystemen meist ungelabelte Datensätze vorliegen und das Hinzufügen von Labeln sehr arbeitsintensiv ist. Planungsansätze auf Basis von UL befinden sich derzeit noch am Anfang, weshalb der Einsatz in CPPS noch nicht umsetzbar ist. Erste Ansätze versuchen jedoch, über das Clustering nach Ähnlichkeiten in einer Struktur die Zuordnung von Planungsinstanzen zu einem generalisierten Plan zu ermöglichen [13]. Alternativ können solche ML-Ansätze dazu genutzt werden, Heuristiken anstelle von vollständigen Plänen zu lernen, welche das Problem vereinfachen und somit zu einer effizienteren Lösung des Planungsproblems beitragen [14].

RL verfolgt den Ansatz, mittels positiver oder negativer Belohnungen eine optimale Strategie zu erlernen, um Planungsprobleme zu lösen. Beispielsweise können, wie bei [15], durch die Vergabe einer sog. *dichten Belohnung* (dense reward) pro Iteration von Zustand, Aktion und Belohnung die Robustheit und die Ergebnisse bei der Erkennung von gültigen und ungültigen Aktionen gute Ergebnisse erzielt werden. Die Kombination von RL und Graphensuche, wie bei Eysenbach et al. kann den Vorteil bieten, hochdimensionale Aufgaben über längere Zeithorizonte zu planen [16]. Bisher wurden reale, komplexe Anwendungen, auch aufgrund der nicht untersuchten Nachvollziehbarkeit des Verhaltens eines RL-Agenten, allerdings nur selten in Betracht gezogen [15]. Dennoch stellt RL eine Alternative zu mathematischen Ansätzen dar [15].

Vergleich der Ansätze anhand der Anforderungen

In der folgenden Tabelle 1 wird verglichen, inwieweit die betrachteten Ansätze die zuvor identifizierten besonderen Anforderungen aus dem Produktionsumfeld erfüllen. Für einen detaillierten Überblick verweisen wir auf [17].

Da es sich bei PDDL lediglich um eine Sprache mit variierenden Schwerpunkten der Solver handelt, haben wir uns dazu entschieden, ein Kriterium nur als teilweise erfüllt zu betrachten, sofern nicht der Großteil der verfügbaren Solver dieses Kriterium unterstützt. So wurde beispielsweise kein Solver gefunden, der alle Anforderungen umsetzt.

Bei einem direkten Vergleich der symbolischen Ansätze fällt auf, dass SMT lediglich bezüglich des Aufwandes Defizite aufweist, stattdessen jedoch in der Lage ist, die komplexen Zusammenhänge eines Produktionsprozesses abzubilden. Von den Ansätzen aus dem Bereich ML schneidet UL bei einem direkten Vergleich am schlechtesten ab. Die anderen beiden Ansätze erscheinen insgesamt ähnlich gut geeignet, mit dem Unterschied, dass SL das Kriterium der Optimierung vollständig erfüllt, dafür jedoch Defizite bei dem benötigten Aufwand aufweist, da gelabelte Daten benötigt werden.

Zwar gibt es für Planungsalgorithmen insbesondere aus dem Umfeld der IPC verschiedene Problembeschreibungen, anhand derer Planungsalgorithmen getestet und verglichen werden können, jedoch fehlen Planungsprobleme mit realitätsnahen, flexiblen Produktionsumgebungen, welche die oben genannten Anforderungen berücksichtigen. Da dies den Vergleich der Planungsansätze auf eine theoretische Betrachtung beschränkt, haben wir eine Benchmark entwickelt, die einen Produktionsprozess abbildet und aus der verschiedene Planungsprobleme generiert werden können. Basierend auf der theoretischen Betrachtung erscheinen ein SMT- sowie ein SL-basierter Ansatz am geeignetsten, sofern der benötigte Aufwand für die Modellierung vorerst unberücksichtigt bleibt. Im Fall des SMT-basierten Ansatzes bietet sich beispielsweise eine teilautomatisierte Generierung des benötigten SMT-Modells aus einem Fähigkeitsmodell der Module an.

Tabelle 1: Vergleich ob die Ansätze die Anforderungen erfüllen (●), teilweise erfüllen (◐) oder nicht erfüllen (○)

	Optimierung	Abhängigkeiten	Schleifen	Nachvollziehbarkeit	Aufwand	Datenintegration
PDDL	◐	○	◐	◐	◐	○
SMT	●	●	◐	◐	○	○
SL	●	◐	◐	○	○	●
UL	○	○	◐	○	◐	●
RL	◐	◐	◐	○	◐	●

Benchmark für CPPS-Planungsprobleme

Um unterschiedliche Planungsansätze für CPPS testen und miteinander vergleichen zu können, wurde eine Vergleichsmöglichkeit in Form einer Benchmark entwickelt. Wie bereits ein-

leitend erwähnt wurde, gehen wir davon aus, dass für eine Anlage, an der ein Planungsproblem gelöst werden muss, eine maschinenlesbare Beschreibung der Anlage und ihrer Funktionen vorliegt.

Für unsere Benchmark nutzen wir eine sog. *Fähigkeitsbeschreibung* der Anlage in Form einer Ontologie nach dem Ansatz aus [5]. In diesem Modell werden Fähigkeiten als Prozessoperatoren der VDI-Richtlinie 3682 modelliert. Jede Fähigkeit kann über Inputs und Outputs verfügen (Informationen, Energien oder Produkte), die wiederum mit Merkmalen weiter spezifiziert werden können. Die Modellierung der Merkmale erfolgt gemäß IEC 61360 sowie [18]. Für eine detaillierte Beschreibung des verwendeten Fähigkeitsmodells wird auf [5] verwiesen. Das Modell ist darüber hinaus online verfügbar¹.

An dieser Stelle soll die entwickelte Benchmark genauer vorgestellt werden. Ziel dieser Benchmark ist es, unterschiedliche Planungsansätze miteinander vergleichbar zu machen, indem Planungsprobleme auf eine strukturierte Art beschrieben werden. Die entwickelte Benchmark orientiert sich an bestehenden Benchmarks aus dem Bereich klassischer KI-Planungs-Ansätze, welche im Rahmen der IPC verglichen werden². Diese Benchmarks sind allerdings in PDDL formuliert und somit nur für PDDL-Solver nutzbar. Aufgrund der bereits genannten Einschränkungen von PDDL soll diese Benchmark jedoch auch einen Vergleich anderer Ansätze - bspw. auf Basis von SMT oder ML - zulassen. Daher wurde die Benchmark im JSON-Format³ erstellt.

Wie auch in PDDL besteht unsere Benchmark aus einer Kombination von Domänen und zu lösenden Problemen, die gemeinsam das Planungsproblem darstellen. Dabei sind mehrere Probleme einer Domäne zugeordnet. In der Domäne werden die Aspekte der betrachteten Planungssituation definiert, die unabhängig von der zu lösenden spezifischen Situation sind. Hier werden vor allem die vorhandenen Fähigkeiten referenziert. Aktuell verfügt unsere Benchmark über mehrere Domänen, welche unterschiedliche Konfigurationen einer modular aufgebauten Laboranlage darstellen. An dieser Laboranlage lassen sich Thermometer und Druckluftzylinder mit einigen Konfigurationsmöglichkeiten fertigen. Abb. 1 gibt eine Übersicht über die Anlage.

¹ <https://github.com/aljoshakoecher/machine-skill-model>

² <https://github.com/potassco/pddl-instances>

³ Die JavaScript Object Notation (JSON) ist ein vergleichsweise einfach aufgebautes maschinenlesbares Daten-Austauschformat, siehe <https://www.json.org/json-en.html>

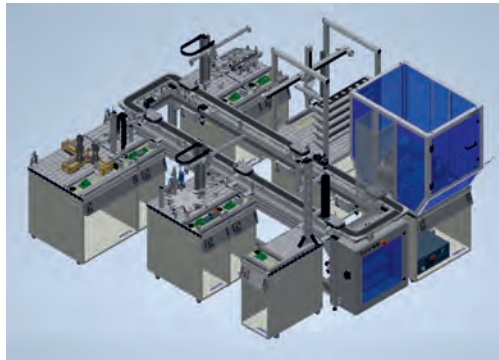


Bild 1: Modular aufgebaute Laboranlage, die die Grundlage der Benchmark darstellt

Unabhängig davon, dass diese Anlage durch weitere Module ergänzt oder bestehende Module ausgetauscht werden können, können die bestehenden Module hinsichtlich ihrer Positionierung getauscht werden. Für diese Anlage wurde ein Fähigkeitsmodell erstellt, das alle Fähigkeiten der einzelnen Module inklusive aller Merkmale und deren Einschränkungen abbildet⁴. Ausgehend von dieser Anlage wurden unterschiedliche Planungsdomänen erstellt, welche sich durch die vorhandenen Fähigkeiten unterscheiden und somit unterschiedliche Anlagenkonfigurationen abbilden. Die entsprechende Domänen-Datei verweist dazu auf das Fähigkeitsmodell der Anlage und listet alle zur Verfügung stehenden Fähigkeiten auf. Damit sind einem Planungsalgorithmus alle möglichen Fähigkeiten bekannt, die zur Problemlösung genutzt werden können. Sämtliche Details dieser Fähigkeiten wie bspw. relevante Merkmale und deren Einschränkungen, sind Teil des Fähigkeitsmodells und müssen von einem Planungsalgorithmus ebenfalls berücksichtigt werden.

Die verschiedenen Probleme werden in unterschiedlichen JSON-Dateien dargestellt. Zunächst enthält jedes Problem einen Verweis auf die Domäne, damit die Zuordnung zu einer Umgebung, in der das Problem gelöst werden soll, eindeutig ist. Darüber hinaus definiert jedes Problem einen Anfangs- und Zielzustand.

Der Anfangszustand definiert die Konfiguration der Anlage. So werden bspw. die Rohteile-Lager der Verteilstation definiert, indem den Eigenschaften (z.B. Farbe, Größe der Rohteile) Werte zugewiesen werden.

⁴ <https://github.com/hsu-aut/MPS500-Capabilities>

Neben dem Anfangszustand wird das Ziel angegeben. Bei dem Ziel kann es sich um ein zu fertigendes Produkt handeln. Zusätzlich oder alternativ dazu lassen sich als Ziel bestimmte Prozessschritte, die ausgeführt werden sollen, modellieren.

Bei Angabe eines zu fertigenden Produkts werden neben dem Typ des Produkts auch die Merkmale des Produkts deklariert. Werden Prozessschritte im Rahmen der Zieldefinition festgelegt, so stellen diese eine deutlich schärfere Vorgabe als die Vorgabe eines Produktes dar, weil sie einen Teil des zu bestimmenden Produktionsplanes festlegen. Jeder Prozessschritt der Zielbeschreibung referenziert eine Fähigkeit, die in diesem Schritt aufzurufen ist. Zusätzlich können die Merkmale des Prozessschrittes festgelegt werden. Neben einzelnen Prozessen können so auch ganze Abfolgen vorgegeben werden.

Der letzte Teil einer Problem-Datei ist die Angabe eines Ergebnisses. Diese Angabe kann wie die Angabe eines *Soll-Ergebnisses* im Rahmen automatisierter Tests genutzt werden, um die berechnete Lösung eines Solvers zu vergleichen. Die Angabe eines Ergebnisses umfasst zunächst eine Aussage darüber, ob überhaupt ein Plan existiert. Falls für das betrachtete Problem kein Plan existieren kann (bspw., weil die benötigten Rohteile nicht in der Konfiguration enthalten sind), so besteht das Ergebnis nur aus der Aussage, dass kein Plan gefunden werden kann. Falls ein Plan existiert, der mit den gegebenen Fähigkeiten und der definierten Anlagen-Konfiguration das geforderte Produkt fertigen bzw. die geforderten Prozesse ausführen kann, so ist dieser Plan im Ergebnis angegeben.

Die entwickelte Benchmark ist inklusive einer etwas umfassenderen Dokumentation offen zugänglich und online unter <https://github.com/hsu-aut/CPPS-Planning-Benchmark> verfügbar.

Ausblick auf zukünftige Planungsansätze für CPPS

Der theoretische Vergleich der KI-basierten Planungsansätze anhand der Anforderungen zeigte, dass im Bereich der symbolischen Verfahren ein SMT-basierter Ansatz und im Bereich der ML-Verfahren ein SL-basierter Ansatz am geeignetsten erscheinen. Aus diesem Grund werden im Folgenden zwei Ausblicke auf mögliche Implementierungen dieser Verfahren im Bereich der Produktionsplanung von flexiblen Produktionssystemen dargestellt.

Ansatz auf Basis von Satisfiability Modulo Theories

Mit Hilfe des SMT-Ansatzes soll eine Sequenz von parametrisierten Skill Aufrufen generiert werden, die das Fertigen eines gewünschten Produktes ermöglicht.

Ein SMT-Solver prüft, ob eine Formel in der Prädikatenlogik erster Stufe unter Berücksichtigung einer zu bestimmenden Theorie T erfüllbar ist oder nicht. Wie oben beschrieben, handelt

es sich bei einem SMT-Problem um ein SAT-Problem welches um verschiedene Hintergrundtheorien ergänzt wurde. Um zu prüfen, ob eine Formel erfüllbar ist, wird in der Regel zuerst eine boolesche Abstraktion der Formel erstellt. Findet der SAT-Solver anschließend eine gültige Zuweisung für die abstrahierte Formel, wird die Konsistenz der Zuweisung bezüglich der Hintergrundtheorie mit Hilfe eines Theorie-Solvers geprüft. Ist dies erfolgreich wurde ein gültiges Modell für das zu betrachtende Problem gefunden.

Übertragen auf das vorliegende Planungsproblem aus dem industriellen Umfeld prüft der Solver folglich, ob es eine gültige Zuweisung von Fähigkeiten gibt, die von einem Ausgangszustand zu einem Zielzustand, dem zu fertigendem Produkt, führen. Zu diesem Zweck werden die Fähigkeiten der Module mit Hilfe von Formeln beschrieben, welche deren Auswirkungen auf die einzelnen Eigenschaften des Systems bzw. die Eigenschaften der Produkte abbilden. Diese Formeln können neben festen Variablen, wie beispielsweise Produkteigenschaften, ebenfalls die zu bestimmenden Parameter enthalten. Am Beispiel eines Ofens kann der Einfluss auf die Temperatur des Produktes beispielsweise mit Hilfe von produktspezifischen Eigenschaften, der zugeführten Leistung sowie der Verweildauer als Formel beschrieben werden. Die produktspezifischen Eigenschaften sind in diesem Fall feste Variablen oder Variablen, die durch andere Fähigkeiten beeinflusst werden können. Die zugeführte Leistung oder die Verweildauer können hingegen Parameter sein, die im Rahmen des Planungsprozesses zu bestimmen sind. Andere Voraussetzungen aus Fähigkeitsmodellen können ebenfalls direkt als Bedingungen hinzugefügt werden. Im Falle des Ofens kann so beispielsweise eine Minimal- und Maximalleistung in Form eines Wertebereichs für die betreffende Variable berücksichtigt werden. Auf diese Art und Weise ist der Planer in der Lage, neben den benötigten Fähigkeiten und deren Reihenfolge, ebenfalls die Parameter zu bestimmen, die für die Skill Aufrufe benötigt werden.

Ansatz auf Basis von Supervised Learning

Das Lernen von Sequenzen ist im Machine Learning schon lange möglich, beschränkt sich aber in den meisten Fällen auf die Bereiche Natural Language Processing (NLP), also das Lernen von Übersetzungen und logischen Zusammenhängen in der Sprache, sowie kombinatorische Problemstellungen. Wie bereits oben beschrieben wird dies im Bereich von CPPS, beziehungsweise in unserem Fall dem Planen von Fertigungsabläufen, noch nicht genutzt. Einerseits aufgrund nicht vorhandener Expertise und andererseits, weil für SL geeignete Beschreibungen von Maschinen meist fehlen. Wir hoffen, dass wir mit der ontologischen Beschreibung und den bestehenden ML-Ansätzen eine Brücke zwischen den Bereichen schlagen können. Die Gesamtontologie einer Anlage stellt einen Graphen dar, weshalb sich bei den

ML Algorithmen auch das *Graph to Sequence Learning* anbietet. Hierbei gibt es zwei verschiedene Ansätze. Einerseits kann man mit *Graph Neural Networks* Heuristiken lernen, die mittels konventioneller Suchalgorithmen die gesuchte Sequenz ausgeben. [11] Andererseits kann die Sequenz auch direkt im Graphen gefunden werden, indem pro Schritt zwei Dinge prognostiziert werden. Erstens die gesuchte Aktion und zweitens der Zustand des Graphen für den nächsten Schritt (siehe Abb. 2) [12].

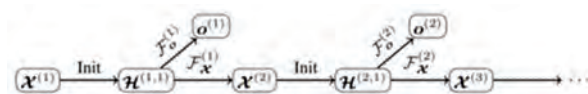


Bild 2: Aktion und Zustand eines Graphen [12]

Unabhängig von der Herangehensweise, benötigt man für das Training eines Modells viele gelabelte Daten. In diesem Fall sind dies gelabelte Beispielsequenzen, in denen zu Beginn nur der Ausgangs- und der Endzustand bekannt ist. Dies beinhaltet sowohl korrekte, als auch inkorrekte Sequenzen. Im Optimalfall sind die Sequenzen aus dem Betrieb des Systems, was aber häufig nur schwer umsetzbar ist. Daher generieren wir die Sequenzen synthetisch, mit Hilfe eines konventionellen Suchalgorithmus. Mit diesen Sequenzen wird das Modell trainiert, womit der Algorithmus anschließend bei vorgegebenen Ausgangs- und Zielzustand die benötigten Schritte bestimmen kann. Der Kernaspekt ist hier, dass auch Sequenzen generiert werden können, die vorher nicht im Training gesehen wurden, da das System die notwendigen Muster gelernt hat.

Zusammenfassung und Ausblick

In diesem Beitrag wurden sechs Anforderungen identifiziert, welche das Planen von Produktionsprozessen in flexiblen Produktionssystemen von anderen Domänen unterscheiden. Der theoretische Vergleich bestehender Planungsansätze zeigt, dass diese beispielsweise Defizite bei dem Abbilden der komplexen Zusammenhänge innerhalb eines Produktionssystems aufweisen. Ein Vergleich anhand eines konkreten Planungsproblems war leider nicht möglich, da KI-basierte Planungsansätze in der Regel nur anhand von einfacheren Problemen getestet und miteinander verglichen werden, welche die Komplexität und Flexibilität von Anlagen nicht abbilden. Um zukünftig eine Aussage darüber zu ermöglichen, inwieweit ein Planungsansatz für ein reales Problem und insbesondere ein Problem auf dem Produktionsumfeld geeignet ist, wurde eine Benchmark mit Beispiel-Problemen unterschiedlicher Komplexität vorgestellt. Diese Benchmark soll in weiteren Arbeiten um zusätzliche Anlagen (bspw. mit verfahrenstechnischen Prozessen) erweitert werden, um gänzlich andere Problem-Domänen untersuchen zu

können und unsere beiden vorgestellten symbolischen und subsymbolischen Planungsansätze weiterzuentwickeln.

Quellen

- [1] H. Kagermann, W. Wahlster, and J. Helbig, Securing the future of German manufacturing industry: Recommendations for implementing the strategic initiative INDUSTRIE 4.0. Berlin: acatech, 2013. [Online]. Available: <https://www.acatech.de/publikation/umsetzungsempfehlungen-fuer-das-zukunftsprojekt-industrie-4-0-abschlussbericht-des-arbeitskreises-industrie-4-0/download-pdf?lang=wildcard>
- [2] L. Monostori, "Cyber-physical Production System: Roots, Expectations and R&D Challenges" *Procedia CIRP*, vol. 17, pp. 9-13, 2014.
- [3] E. Järvenpää, "Capability-based Adaptation of Production Systems in a Changing Environment", Dissertation, Tampere University of Technology, 2012.
- [4] A. Köcher, C. Hildebrandt, L. M. Vieira da Silva, and A. Fay, "A formal capability and skill model for use in plug and produce scenarios," in 2020 25th IEEE ETFA, vol. 1, 2020, pp. 1663–1670, ISSN: 1946-0759.
- [5] S.J. Russell, P. Norvig: Artificial intelligence: A modern approach. Fourth Edition. Hoboken, NJ: Pearson, 2021 (Pearson Series in Artificial Intelligence).
- [6] C. Aeronautiques, A. Howe, C. Knoblock et al., "PDDL| The Planning Domain Definition Language," Technical report, Tech. Rep., 1998.
- [7] A. Anis, W. Schäfer, and O. Niggemann, "A comparison of modeling approaches for planning in cyber physical production systems," in Proceedings of the 2014 IEEE ETFA. 2014, pp. 1–8.
- [8] A. Rogalla and O. Niggemann, "Automated process planning for cyber physical production systems," in 2017 22nd IEEE ETFA, 2017, pp. 1–8.
- [9] Kautz and B. Selman, "Planning as Satisfiability" in Proceedings of the 10th European Conference on Artificial Intelligence (ECAI 92), 1992.
- [10] C. Barrett, L. de Moura, S. Ranise et al., "The SMT-LIB Initiative and the Rise of SMT," in Hardware and Software: Verification and Testing, ser. Lecture Notes in Computer Science, S. Barner, I. Harris, D. Kroening, and O. Raz, Eds. Springer, 2011, pp. 3–3.
- [11] W. Shen, F. Trevizan, and S. Thiébaux, "Learning domain-independent planning heuristics with hypergraph networks," in Proceedings of the International Conference on Automated Planning and Scheduling, vol. 30, 2020, pp. 574–584.

- [12] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, "Gated graph sequence neural networks." [Online]. Available: <http://arxiv.org/pdf/1511.05493v4>
- [13] Javier Segovia-Aguas, Sergio Jiménez, and Anders Jonsson, "Unsupervised classification of planning instances," *Proceedings of the International Conference on Automated Planning and Scheduling*, vol. 27, pp. 452–460, 2017.
- [14] G. Francès, A. B. Corrêa, C. Geissmann, and F. Pommerening, "Generalized Potential Heuristics for Classical Planning," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*, S. Kraus, Ed. California: International Joint Conferences on Artificial Intelligence, 2019, pp. 5554–5561.
- [15] A. Kuhnle, J.-P. Kaiser, F. Theiß et al., "Designing an adaptive production control system using reinforcement learning," *Journal of Intelligent Manufacturing*, vol. 32, no. 3, pp. 855–876, 2021.
- [16] B. Eysenbach, R. R. Salakhutdinov, and S. Levine, "Search on the replay buffer: Bridging planning and reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [17] A. Köcher, R. Heesch, N. Widulle et al., "A Research Agenda for AI Planning in the Field of Flexible Production Systems," in *2022 5th IEEE ICPS*, 2022
- [18] C. Hildebrandt, A. Scholz, A. Fay et al., "Semantic modeling for collaboration and cooperation of systems in the production domain" in *2017 22nd IEEE ETFA*, 2017, pp. 1–8

Auf dem Weg zur Digitalisierung von R&I

Evaluierung von Few Shot Learning Techniken zur einfachen Erkennung bisher unbekannter Elemente

Towards Digitization of P&IDs

Evaluation of Few Shot Learning Techniques to Recognize Previously Unknown Elements Easily

Dr. M. Gärtler, Dr. B. Schmidt, ABB AG, Ladenburg

Kurzfassung

In diesem Beitrag werden die Probleme und Herausforderungen der Digitalisierung von Rohrleitungs- und Instrumentendiagramm (R&I oder auch P&ID) erörtert, die in der Prozessindustrie häufig für die Konstruktion verwendet werden. Während Initiativen wie DEXPI (<https://dexpi.org/>) darauf abzielen, einen allgemeinen Datenaustauschstandard in digitaler Form bereitzustellen, sind die meisten P&IDs nur auf Papier, als gescannte Bilder oder in herstellerspezifischen Formaten verfügbar. Das übergeordnete Ziel ist es, bestehende "nicht-digitale" P&IDs nach DEXPI zu übertragen. Für diese Digitalisierung wurden in den letzten Jahren verschiedene Ansätze, einschließlich der Objekterkennung durch Computer Vision, mit unterschiedlichem Erfolg angewandt. Während es Normen wie ISO 10628 und ISO 14617 gibt, die Nomenklatur und Symbole für P&IDs definieren, werden sie in der Realität von Werkzeugen oft unterschiedlich umgesetzt, was in der Praxis zu vielen Abweichungen führt. Infolgedessen sind die P&IDs für Fachleute immer noch leicht lesbar, während sie für Programme aufgrund der unterschiedlichen Piktogramme schwer zu interpretieren sind, was die Digitalisierung besonders schwierig macht. Hier helfen Meta-Lernen und "few-shot learning", um die Anpassung von zuvor erworbenem Wissen an neue Aufgaben zu nutzen. Aufbauend auf einem (zuvor erlernten) Standard reichen sehr wenige Beispiele (~1-5) aus, um das Modell für einen neuen Satz von Symbolen zu erweitern. Wir wenden eine Auswahl von Techniken an und bewerten ihre Nützlichkeit bei der Erkennung neuer Symbole anhand weniger Beispiele.

Genauer gesagt, betrachten wir die Aufgabe der Erkennung der Elemente in zwei Versionen: Erstens als eine Variante von Omniglot [1], die sich mit der Erkennung von Zeichen in

verschiedenen / unbekannten Alphabeten beschäftigt. Zweitens als eine Variante der Erweiterung der Objekterkennung, wie sie in [2] beschrieben ist.

Abstract

The paper discusses the problems and challenges of digitization of piping and instrumentation diagrams (P&ID), which are commonly used for engineering in the process industry. While initiatives like DEXPI (<https://dexpi.org/>) aim to provide a general data exchange standard in digital form, the majority of P&IDs are only available on paper, as scanned images or in vendor-specific formats. The overall goal is to transfer existing “non-digital” P&IDs to DEXPI. Different approaches for this digitalization have been applied, including object detection in computer vision from recent years, with various degrees of success. While there are standards such as ISO 10628 and ISO 14617 that define nomenclature and symbols for P&IDs, in reality (engineering) tools often implement them differently, which leads to many variations. As a result, the P&IDs are still easily readable by domain experts, while hard to interpret programmatically due to different pictograms, which makes digitization especially difficult. Here, meta learning and few-shot learning help to leverage adaptation of previously gained knowledge to new tasks. Building upon a (previously learned) standard, very few examples (~1-5) are sufficient to extend the model for a new set of symbols. We apply a selection of technique and evaluate their usefulness in recognizing new symbols given a few examples. More precisely, we consider the task of recognizing the elements in two versions: Firstly, as a variation of Omniglot [1], which deals with the recognition of characters in different / unknown alphabets. Secondly, as a variation of the extension of object detection, as described [2].

Introduction

Nowadays, Industry 4.0 and the industrial internet of things (IIoT) with its data-centric view are ubiquitous. Artificial intelligence (AI) and Machine Learning (ML) have been identified as key aspects in the digitization in the industry [3]. While both have been intensively studied in the recent years with many impressive success stories under their belt, there are still many challenges as reported in [4]. We illustrate this on a problem from the digital transformation. Most task in that area are essential to bring established information and knowledge, often in the form of knowing by heart, papers, or legacy systems, to the digital system in (universally) accepted standard formats.

One such task is the transformation of piping and instrumentation diagrams (P&ID), which are commonly used during engineering and often exist only as printouts, images, or scanned PDFs to a format like DEXPI (<https://dexpi.org/>) which aim to provide a general data exchange

standard in digital form. Overall, the task has also been intensively studied, consider [5]–[10] as examples. On the one hand, the task can be directly interpreted as an optical character recognition (OCR) or, more generally, as object detection (OD). Many P&IDs follow established standards such as ISO 10628 and ISO 14617 with a relatively small number of primitives (less than 500 different pictograms). Furthermore, the humans look at P&ID, even without intensive training, they can easily read the diagram; thus, the task should be easy. On the other hand, the used pictograms often vary slightly from the defined standard, or use composed annotation, where two similarly looking elements have very different meaning due to small distinct differences. For example, sensor types are often indicated by a letter such as F, L, P or T for flow, level, pressure, or temperature, respectively, which cover a very small area compared to the overall symbol. Furthermore, a typical number of related data sets (per case) is usually very small, thus making it a rather hard problem.

These aspects make the task an ideal candidate for meta learning [11]–[13] and few shot learning [2], [14], in particular. One related and prominent example, where the methods have been highly successfully applied, is Omniglot, where different symbols and characters from different alphabets need to be identified. The task is split into an initial phase, where the training uses a subset of the data, and the transfer phase, where re-training or adjustment was done for previously unknown alphabets and characters.

Next, we give a detailed overview of meta and few shot learning, followed by the description of our use case. We conclude with our evaluation of different methods and the assessment of quality and usability.

Metalearning - State-of-the Art/Related Work

Metalearning is a branch of machine learning which investigates the learning process itself. Understanding and ultimately improving the learning process for certain tasks belongs to the domains of metalearning. This can be seen in a typical metalearning architecture as described by [15] which involves a learning system and a learning subsystem. The learning system extracts metadata about the learning process of the learning subsystem. The learning system, on the other hand, makes use of the metadata about the learning process and improves its learning process. The problems learned by the system belong to a class of similar tasks (e.g., similar types of images or timeseries). The metadata influences the so-called learning bias, prior knowledge which modifies the learning of the sub process by, e.g., changing the search space or the initial parameterization.

One sub-field of metalearning is few-shot learning. Here, the interplay between the learning system and the subsystem is used to reduce the data requirements of the learning process of

the subsystem. The learning system generates a variety of different tasks of the same task class, which are learning input for the subsystem. Every task includes a very small amount of learning data, as little as 1 to few dozen examples. With every training cycle, the ability to learn a model for the task class with few examples improves. The overall training process requires access to many task class instances. Once this process is finished, generation of a model for a new instance of the same task class is possible with few examples.

Typically, three different types of meta learning are distinguished: model-based, metric-based and optimization-based. Firstly, the model-based approach takes a given model and updates the parameters based on a few training steps. A realization of this for neural networks is given in [2]. In a first step, a model is trained using base training with instances of the considered class. In their realization, the base model is a mask r-cnn [16] generated using the Detectron 2 platform [17]. Few shot learning is realized as fine tuning only the last layer of the model using few shot data for the base images and the new classes. Secondly, the metric-based approach learns a projection of the task data into an embedding space. In this space, the distance between objects is used to decide whether they are similar or not. Two methods which apply this technique are Siamese networks [18] and prototypical networks [19]. Siamese networks calculate sample similarity. Prototypical networks learn class prototypes and calculate the distance of a sample to each prototype. Finally, optimization-based techniques adjust the optimization algorithm of the model to enable learning with few examples. The Reptile [20] algorithm realizes it. Reptile copies a main model to the learning subsystem, executes few shot learning in the subsystem and updates the main model parameters based on the adjustment of the parameters which happened in the learning subsystem. Another example of this group of algorithms is Model-agnostic Meta Learning [21].

Use Case – Pictogram Identification in P&IDs

The detection of diagram symbols is a relevant and active field of research. One reason is that most complex human-made structures, assets or setups are designed based on diagrams. Especially for structures, assets and setups created in times before 2010, the design documents typically do not exist in a human readable form. In this paper, we focus on piping and instrumentation diagrams (P&ID) diagrams used in process plants. Most brownfield plant P&ID diagrams exist as paper documents, which often are digitalized as a scan. Digital P&ID diagrams (e.g., the DEXPI format) have many advantages over static documents, as they allow the extraction of information from the specified structure. A variety of techniques have been proposed [22], [23], most recently also many machine learning-based techniques [24], [25]. In

all cases, the instrument symbols are extracted, their type is identified based on the form and the pipes between the instruments are identified. The result is a structured digital representation of the P&ID diagram. These techniques have one specific challenge in common: the variety of symbol sets. As mentioned above, instrumentation symbols build on standards, with sometime custom- or tool-specific variations. Furthermore, details are often an important factor, because small annotations can change the meaning. Assuring robustness of the approaches to a variety of different symbol sets is difficult to achieve. Especially the machine learning-based approaches require a large set of examples to be trained for a specific shape. Taking one model for P&ID symbols of different companies will require many annotated examples, which typically are not given.

Here, few shot learning is an interesting approach. Few shot learning acts in the domain of learning with a relatively small number of labeled examples – often between 1 and 10.

For the activity described in this paper, we explored the application of few shot learning for symbol robustness on P&ID diagrams. Not to be mistaken – few shot learning still requires large amounts of data, to generate a model which is capable of few shot learning. Therefore, the ability to generate many tasks of the same task class is important. We investigated the performance on 1) basic symbols, 2) P&ID symbols. The foundation for the basic symbols was the Omniglot dataset [1] (containing 1623 characters of 50 alphabets with 20 examples each). The P&ID symbols were taken from the standard.

These datasets were applied in tests for the two metalearning types: metric-based and optimization-based. For the third, model-based techniques first experiments were executed, building on the work in [2] based on [16] but not followed-up due to time-constraints. For the metric-based approach prototypical networks were chosen, for the optimization-based approach reptile was chosen. To assure reproducibility a frequently used implementation of the algorithms was used, as provided in the learn2learn¹ framework [26].

Few shot learning setups are specified in terms of n-way, k-shots. Way specifies the number of classes to be compared and k specifies the number of labeled examples. For example, a 3-way, 1-shot problem is a few shot learning setup which trains a 3-class classifier using 1-single example of each class per inner training loop.

Experiments were designed as follows. First, a learning system and learning subsystem architecture was used, with a prototypical network and with the reptile algorithm. The respective architecture was used to train a model based on symbols from one of the three

¹ <https://github.com/learnables/learn2learn>

aforementioned symbol classes (basic symbols, P&ID symbols). In all cases the learn2learn implementations for task sampling were used to generate the training tasks with shot-many examples and way-many classes. Data augmentation was applied to increase the amount of data for the model training – this was achieved by changing size, position, and orientation of the available images of the datasets. Furthermore, techniques like cropping and adding noise to the images were tested with varying effect.

Once a model was created, it was tested using P&ID diagrams. In the case of training with P&ID diagrams it was assured that P&ID diagram classes used for training were not used in the testing.

The described setup has two main targets. First, the general performance of few-shot learning on symbol data is investigated. Second, and more specific, the task-transfer capability of the few shot learning model is investigated. Learning on hand-written characters and applying few-shot learning on P&ID diagrams is a much broader understanding of a similar task class, than training on P&ID diagrams and testing on different P&ID diagrams. Here, Omniglot to P&ID seems to include the highest difference with P&ID to P&ID to be the easiest.

Experiments

The experimental setup was as follows. The algorithms were tested for all three data classes in a range of 2 to 10 ways (classes) and 5 to 7 shots (labeled examples). In some cases, additional transformation techniques of cropping, blur and adding noise were used, following the intuition that increased variety and complexity in the train data increases robustness. As the report will show, this did not work with the used level of cropping up to 40 percent of an image and gaussian noise.



Fig. 1: Training Example with data augmented using blur

In the following, an overview of the covered experiments is provided, and results are reported. In all cases, it is possible to create a model with acceptable performance for 5 to 7 shots (labeled examples) in the tested range of 2 to 10 classes. The effect of cropping and adding noise transformation techniques for data augmentation is negative, pure resizing re-orientation and re-positioning shows the best results.

While the prototypical network works best if trained directly on PID data, reptile shows performance gains for training on Omniglot – here the broader and more generic shapes of Omniglot seem to support creation of a more robust model, at least for the case of PID detection.

Often a performance increase results from adding few more examples. The gain is typically between 2 to 5 percent per added example. In some cases, the difference can be over 10 percent per added labeled example. Thinking of an interactive application for automatic annotation, this could be used to ask experts for additional labels for areas with weak performance to have a general improvement. This observation goes together with the observation that some examples do not show this trend – we expect that the smaller number of examples makes it much more important to have “good” labeled examples. In our case examples are randomly selected, without taking the information gain or example quality into account.

We further observed that the algorithms soon in the training could capture general shapes well—only the initially mentioned subtle differences between shapes often complicated detection. This effect could also be useful in a support tool. While the reported numbers here focus on having the exact correct classification, it is also possible for a metric-based approach like prototypical network to list the best guesses and let a human expert do the final selection—the person would not need to check one of dozens of shapes, but only select across few shapes.

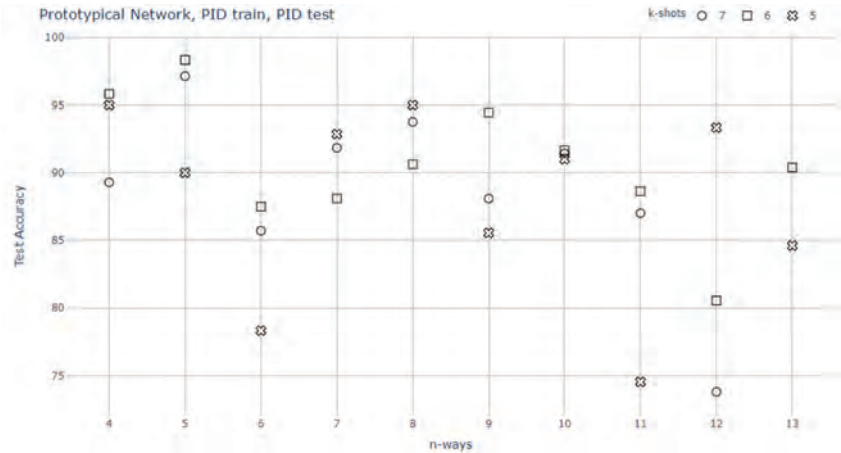


Fig. 2: Prototypical Network, PID train, PID test (no crop and noise transformations for data augmentation)

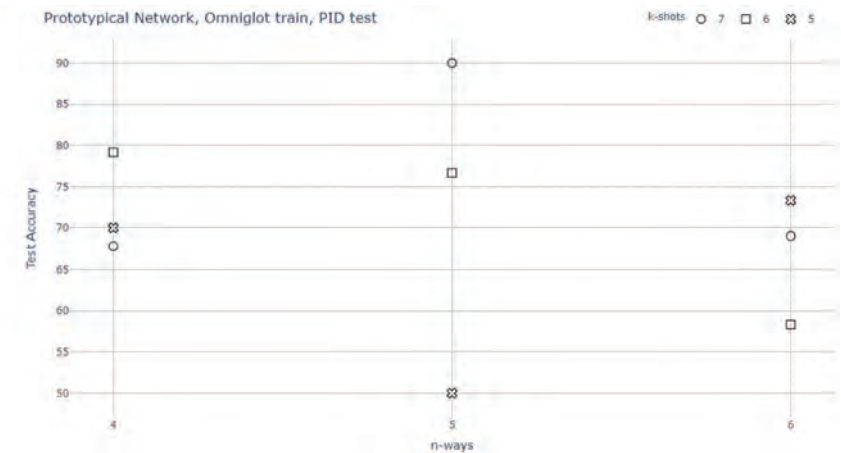


Fig. 3: Prototypical Network, Omniglot train, PID test (no crop and noise transformations for data augmentation)

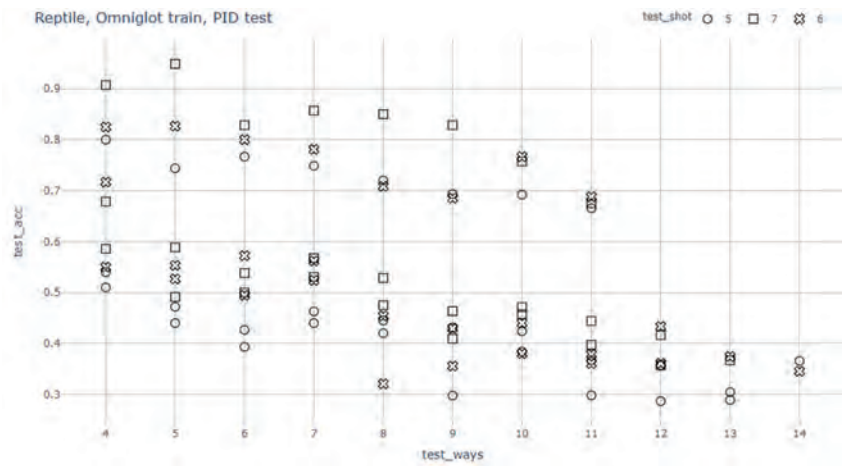


Fig. 4: Reptile, Omniglot train, PID test (crop and noise transformations for data augmentation for results below 0.6)

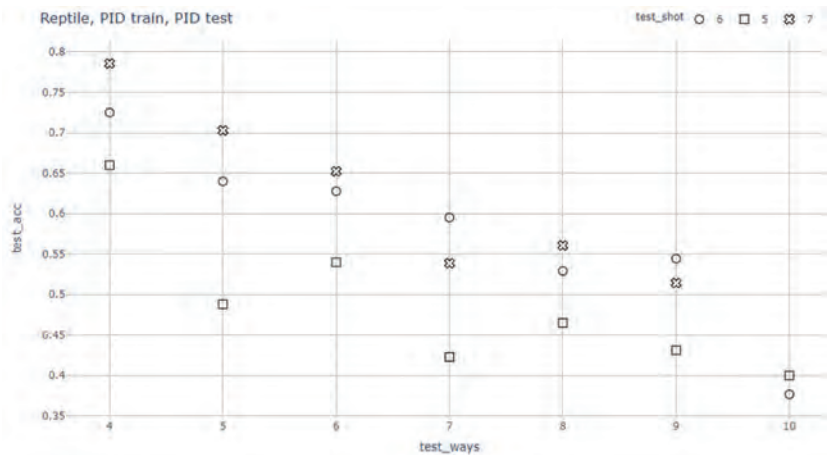


Fig. 5: Reptile, PID train, PID test (crop and noise transformations for data augmentation)

Conclusion

Existing methods using meta and few shot learning, previously used for Omniglot and similar task, could be successfully applied to symbol recognition in P&IDs, even with previously unknown symbols. This is also highly suitable to be integrated in workflows with human, for examples an expert marks custom symbols, which are likely unknown or custom-specific or refine obtained identification. In summary, we believe few shot learning techniques are a good extension of techniques to solve the digitalization effort of brownfield P&ID diagrams. It especially allows for an easy integration of human domain expertise with system support for data annotation. For an actual application, an extended test is required which investigates robustness on more task classes.

As general learnings, the analysis has shown that the specification of task similarity is a crucial step for few shot learning approaches, while it also shows that flexibility exists. Here, we have seen that Omniglot characters provide a good foundation for creating a few shot learning base model which can be used with few labeled examples for P&ID symbol recognition. The development of standardized frameworks is of great help—here learn2learn allowed for an easy setup and configuration of the learning tasks.

Bibliography

- [1] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "The Omniglot challenge: a 3-year progress report," *arXiv:1902.03477 [cs]*, May 2019, Accessed: May 03, 2022. [Online]. Available: <http://arxiv.org/abs/1902.03477>
- [2] X. Wang, T. Huang, J. Gonzalez, T. Darrell, and F. Yu, "Frustratingly Simple Few-Shot Object Detection," in *Proceedings of the 37th International Conference on Machine Learning*, Nov. 2020, pp. 9919–9928. Accessed: Nov. 02, 2021. [Online]. Available: <https://proceedings.mlr.press/v119/wang20j.html>
- [3] J. Diemer *et al.*, "KI in der Industrie 4.0: Orientierung, Anwendungsbeispiele, Handlungsempfehlungen," *undefined*, 2020, Accessed: Nov. 09, 2021. [Online]. Available: <https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/ki-in-der-industrie-4-0-orientierung-anwendungsbeispiele-handlungsempfehlungen.html>
- [4] M. Gärtler, V. Khaydarov, B. Klöpper, and L. Urbas, "The Machine Learning Life Cycle in Chemical Operations – Status and Open Challenges," *Chemie Ingenieur Technik*, vol. 93, no. 12, pp. 2063–2080, 2021, doi: 10.1002/cite.202100134.
- [5] H. K. Yuen, J. Princen, J. Dlingworth, and J. Kittler, "A Comparative Study of Hough Transform Methods for Circle Finding," in *Proceedings of the Alvey Vision Conference 1989*, Reading, 1989, p. 29.1-29.6. doi: 10.5244/C.3.29.

- [6] J.-M. Cha, T. Lee, S. Salim, and B. Ryu, "Object Detection for P&ID Images," in *International Congress and Conferences on Computational Design and Engineering*, Jul. 2019, vol. 3.
- [7] M. Gada, "Object Detection for P&ID Images using various Deep Learning Techniques," in *2021 International Conference on Computer Communication and Informatics (ICCCI)*, 2021, pp. 1–5. doi: <https://doi.org/10.1109/ICCCI50826.2021.9402386>.
- [8] S.-O. Kang, E.-B. Lee, and H.-K. Baek, "A Digitization and Conversion Tool for Imaged Drawings to Intelligent Piping and Instrumentation Diagrams (P&ID)," *Energies*, vol. 12, no. 13, Art. no. 13, Jan. 2019, doi: 10.3390/en12132593.
- [9] R. Rahul, S. Paliwal, M. Sharma, and L. Vig, "Automatic Information Extraction from Piping and Instrumentation Diagrams," *arXiv:1901.11383 [cs]*, Jan. 2019, Accessed: Sep. 06, 2021. [Online]. Available: <http://arxiv.org/abs/1901.11383>
- [10] E.-S. Yu, J.-M. Cha, T. Lee, J. Kim, and D. Mun, "Features Recognition from Piping and Instrumentation Diagrams in Image Format Using a Deep Learning Network," *Energies*, vol. 12, no. 23, Art. no. 23, Jan. 2019, doi: 10.3390/en12234425.
- [11] M. Huisman, J. N. van Rijn, and A. Plaata, "A survey of deep meta-learning," *Artif Intell Rev*, Apr. 2021, doi: 10.1007/s10462-021-10004-4.
- [12] A. Balte, "Meta-Learning With Landmarking: A Survey," *International Journal of Computer Applications*, vol. 105, no. 8, p. 5.
- [13] H. Peng, "A Comprehensive Overview and Survey of Recent Advances in Meta-Learning," *arXiv:2004.11149 [cs, stat]*, Oct. 2020, Accessed: May 20, 2021. [Online]. Available: <http://arxiv.org/abs/2004.11149>
- [14] V. Garcia and J. Bruna, "Few-Shot Learning with Graph Neural Networks," *arXiv:1711.04043 [cs, stat]*, Feb. 2018, Accessed: Jun. 30, 2021. [Online]. Available: <http://arxiv.org/abs/1711.04043>
- [15] C. Lemke, M. Budka, and B. Gabrys, "Metalearning: a survey of trends and technologies," *Artificial intelligence review*, vol. 44, no. 1, pp. 117–130, 2015.
- [16] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 2980–2988. doi: 10.1109/ICCV.2017.322.
- [17] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2." 2019. [Online]. Available: <https://github.com/facebookresearch/detectron2>

- [18] G. Koch, "Siamese Neural Networks for One-Shot Image Recognition," University of Toronto, 2015. [Online]. Available: <http://www.cs.toronto.edu/~gkoch/files/msc-thesis.pdf>
- [19] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical Networks for Few-shot Learning," *arXiv:1703.05175 [cs, stat]*, Jun. 2017, Accessed: May 20, 2021. [Online]. Available: <http://arxiv.org/abs/1703.05175>
- [20] A. Nichol and J. Schulman, "Reptile: a scalable metalearning algorithm," *arXiv preprint arXiv:1803.02999*, vol. 2, no. 3, p. 4, 2018.
- [21] C. Finn, P. Abbeel, and S. Levine, "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks," *arXiv:1703.03400 [cs]*, Jul. 2017, Accessed: May 20, 2021. [Online]. Available: <http://arxiv.org/abs/1703.03400>
- [22] E. Arroyo, M. Hoernicke, P. Rodríguez, and A. Fay, "Automatic derivation of qualitative plant simulation models from legacy piping and instrumentation diagrams," *Computers & Chemical Engineering*, vol. 92, pp. 112–132, Sep. 2016, doi: 10.1016/j.compchemeng.2016.04.040.
- [23] S. Paliwal, A. Jain, M. Sharma, and L. Vig, "Digitize-PID: Automatic Digitization of Piping and Instrumentation Diagrams," *PAKDD*, 2021, doi: 10.1007/978-3-030-75015-2_17.
- [24] H. Kim *et al.*, "Deep-learning-based recognition of symbols and texts at an industrially applicable level from images of high-density piping and instrumentation diagrams," *Expert Systems with Applications*, vol. 183, p. 115337, Nov. 2021, doi: 10.1016/j.eswa.2021.115337.
- [25] S. Mani, M. A. Haddad, D. Constantini, W. Douhard, Q. Li, and L. Poirier, "Automatic Digitization of Engineering Diagrams using Deep Learning and Graph Search," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2020, pp. 673–679. doi: 10.1109/CVPRW50498.2020.00096.
- [26] S. M. R. Arnold, P. Mahajan, D. Datta, I. Bunner, and K. S. Zarkias, "learn2learn: A library for Meta-Learning research," Aug. 2020, [Online]. Available: <http://arxiv.org/abs/2008.12284>

Digitale Zwillinge als Enabler für Agentensysteme zur Realisierung autonomer Wertschöpfungsnetzwerke mit erhöhter Produktivität

E. Bayrhammer, Fraunhofer IFF, Magdeburg;
Univ.-Prof. **B. Vogel-Heuser**, **F. Ocker**,
Technische Universität AIS, München;
H. Avgoustinos, **R. Thrift**,
EXPO21XX GmbH, Hannover

Kurzfassung

Agentenbasierte Produktionssysteme versprechen ein hohes Maß an Flexibilität für Wertschöpfungsnetzwerke. Allerdings sind ihre Inbetriebnahme sowie die Kopplung mit physischen Assets im Netzwerk aufwendig. Eine wesentliche Vereinfachung für beides versprechen Digitale Zwillinge. Dieser Beitrag beschreibt, wie Digitale Zwillinge, konkrete in der Form der Verwaltungsschale, für die Implementierung von agentenbasierten Produktionssystemen genutzt werden können.

1. Motivation

Der stetig zunehmende Wunsch nach Individualisierung in Kombination mit stark umkämpften Märkten führt zu einem hohen Druck auf produzierende Unternehmen. Dies wird durch globale Krisen, wie z.B. COVID, verschärft. Ein wichtiges Ziel ist daher Resilienz zu schaffen [1, 2], z.B. durch resiliente Anlagen und flexible Wertschöpfungsnetzwerke, die auch in solchen Situationen anpassungsfähig bleiben und in der Lage sind eine hohe Produktivität im Sinne der Overall Equipment Effectiveness (OEE) aufrechtzuerhalten. Agentensysteme können hierfür ein zentraler Baustein sein. Agenten als autonome, kooperierende Entitäten repräsentieren konkrete Produktionsressourcen und Produkte. Sie können proaktive Eigenschaften aufweisen und handeln die Nutzung von Ressourcen nach Zielvorgaben (Betriebswirtschaftlichkeit, Nachhaltigkeit etc.) untereinander aus. Bisher scheiterte die breite industrielle Nutzung agentenbasierter Ansätze jedoch unter anderem am Aufwand für die Erstellung von Wissensbasen für die einzelnen Agenten, der Initialisierung des Agentensystems, sowie der Kopplung der Agentenebene mit der physischen Ebene der Produktionssysteme.

Zur Lösung dieser Probleme wird die Nutzung von digitalen Zwillingen als ein weiterer Baustein vorgeschlagen. Damit liegen die notwendigen Beschreibungen vor, um Agentensysteme (teil-)automatisiert in Betrieb zu nehmen und mit den physischen Komponenten zu koppeln.

Für den gemeinsamen Betrieb wird eine Laufzeitumgebung benötigt. Diese Kombination stellt eine proaktive Verwaltungsschale [3] dar, mit der die Idee der Industrie 4.0-Komponente [4] umgesetzt werden kann. Um den Aufwand bei Erstellung und Einbindung von Digitalen Zwillingen und Agenten zu reduzieren, wird ein Ökosystem benötigt, welches eine entsprechende Tool Chain anbietet, um den Nutzer mittels Automatisierung und verständlichen Nutzerschnittstellen bei der Anwendung der Technologiekombination aus Agenten und Digitalen Zwillingen zu unterstützen. So kann die Industrie, insbesondere auch KMU, zur Nutzung moderner Produktionssysteme auf Basis von Digitalen Zwillingen und Agenten befähigt werden.

Basierend auf akademischen Vorarbeiten wird in diesem Beitrag eine Cloud- und Edge-Computing-Architektur inklusive eines Ansatzes für die Inbetriebnahme des Agentensystems unter Nutzung von Digitalen Zwillingen vorgestellt. Mit Abschnitt 2 erfolgt die Einordnung in den Stand der Wissenschaft und Technik. Anschließend werden die Konzepte dargestellt, um Agentensysteme, Digitale Zwillinge und das reale Produktionssystem zu koppeln. Dies umfasst eine Laufzeitumgebung zur Verbindung von Agentensystemen und Verwaltungsschalen (VWS) in einem resilienten Wertschöpfungsnetzwerk, das Design des Agentensystems basierend auf eigenen Vorarbeiten [5, 6] sowie die Implementierungsmöglichkeiten von agentenbezogenen Informationen im Digitalen Zwilling. In den Folgeabschnitten wird die Umsetzung beschrieben und ein Ausblick. Gegebenen.

2. Stand der Wissenschaft und Technik

Dieser Abschnitt gibt einen Überblick über verwandte Arbeiten im Kontext Digitaler Zwillinge und Wertschöpfungsnetzwerke, agentenbasierte Systeme sowie Plattformen und Ökosysteme im Kontext von Industrie 4.0.

2.1 Digitale Zwillinge und Wertschöpfungsnetzwerke

Der Digitale Zwilling (DZ) ist "eine der vielversprechendsten Basistechnologien zur Realisierung von Smart Manufacturing und Industrie 4.0 [...], die sich durch die nahtlose Integration zwischen der virtuellen und der physischen Welt auszeichnet" [7]. Die Idee ist dabei, mit Hilfe der Abbildung der Eigenschaften eines Assets (hier Maschine, Komponente, etc.) in der digitalen Welt, eine Verbindung zwischen Informations- und Operationstechnologie herzustellen. Die Abbildung der Eigenschaften erfolgt in Form entsprechender Daten, die zum realen Asset aktuell gehalten werden. Je nach Use Case kann auf dieser Basis eine vertikale und/ oder horizontale Integration erfolgen. Von vertikaler Integration spricht man bei der durchgängigen Nutzung des Digitalen Zwillingen bei einem Stakeholder über mehrere Ebenen hinweg, das heißt von der Akteur-/Sensorebene, die Steuerungsebene, über die Produktionsleitebene, bis

in die Unternehmensplanungsebene [8]. Eine horizontale Integration wird erreicht, wenn der Digitale Zwilling über den Lebenszyklus des Assets hinweg, d.h. über Organisations- und Firmengrenzen hinaus zwischen verschiedenen Stakeholdern genutzt wird. Über beide Richtungen wird ein Wertschöpfungsnetzwerk aufgespannt (Abbildung 1). Der Informationsgehalt des Digitalen Zwillings kann darin nach dem Add- und Reuse-Prinzip [9] wachsen und so zahlreiche Use Cases ermöglichen. Die Herausforderungen für eine solche Vision sind vielfältig. Um dies auf eine einheitliche Basis zu stellen, laufen verschiedene Standardisierungsbestrebungen zum Thema (z.B. [10, 11]).

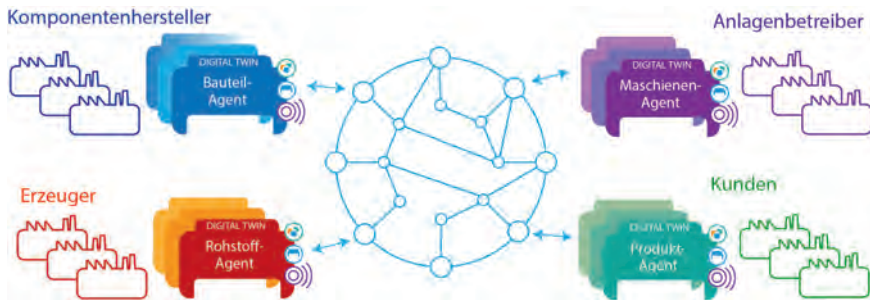


Bild 1: Wertschöpfungsnetzwerk

Das in Abbildung 1 dargestellte Wertschöpfungsnetzwerk kann durch die Verwendung verteilter Digitaler Zwillinge [12] in gewissen Grenzen (Parameter) auf Kundenwünsche wie auf Variationen in zugelieferten Komponenten reagieren. Bei größeren Einflüssen fehlt jedoch weitergehende automatisierbare Flexibilität. Hier kommen Agenten ins Spiel.

2.2 Multiagentensysteme und Flexibilität

Agenten als intelligente Entitäten sind gekennzeichnet durch Autonomie, Kapselung und ihre eigenen Ziele [13]. Dies deckt sich mit der Definition des VDI von Agenten als gekapselte Hardware- oder Softwareentität mit definierten Zielen hinsichtlich der Steuerung eines technischen Systems [14]. Speziell im Produktionskontext wird zwischen Ressourcenagenten, Prozessagenten, Kommunikationsagenten und dem Agentenmanagementsystem unterschieden [15]. Ressourcenagenten repräsentieren Produktionsressourcen, also Maschinen oder ganze Fabriken, während Prozessagenten den Prozess steuern. Kommunikationsagenten können genutzt werden, um auch Bestandssysteme anzubinden. Das Agentenmanagementsystem

verwaltet alle Agenten im System. Zudem können Produktagenten genutzt werden um intelligente Produkte zu realisieren [16]. In Kombination ergeben mehrere Agenten ein sogenanntes Multi-Agenten-System (MAS), welches sich besonders durch seine Dezentralisierung und Parallelisierung von Aktivitäten auszeichnet [17]. Somit sind MAS inhärent geeignet sind, um die Flexibilität und Adaptivität von Produktionssystemen zu steigern.

Zur Umsetzung von MAS sind verschiedene Frameworks verfügbar, welche von generischen Implementierungen (z.B. JADE, PADE, akka¹) bis zu Architekturen für die Produktion (z.B. ADACOR [18]) reichen. Agenten in modernen Architekturen sind in der Regel mit einer Wissensbasis ausgestattet, die Annahmen, Ziele und Intentionen (en: beliefs, desires, intentions) abbildet [19]. Die für diese Wissensbasen benötigten Informationen sind zum Teil in Form von Entwicklungsdokumenten und Kundenbestellungen implizit vorhanden, eine automatisierte Nutzung ist aber aufgrund der Heterogenität der Dokumente und Schwierigkeiten bei der Interpretation problematisch. Bisher gibt es nur wenige Ansätze, um diesen Prozess zu unterstützen. So können die Wissensbasen von Agenten beispielsweise aus Ontologien initialisiert werden [20], welche allerdings keine Anbindung an das physische System bieten. DZ haben das Potential beide Herausforderungen, hinsichtlich der Erstellung der Wissensbasen und der Anbindung von Agenten an das physische Asset, zu adressieren und somit bisher hohe manuelle Aufwände zu reduzieren.

2.3 Industrie 4.0 Plattformen und Ökosysteme

Laut einer aktuellen Studie [21] sind mehr als 450 verschiedene IoT-Plattformen verfügbar. Eine kleine Auswahl wurde zum Beispiel in anderen Vorarbeiten [22] näher betrachtet. Sie umfasst einerseits die umsatzstärksten Hersteller und andererseits Plattformen, die für den Kontext des Beitrags relevant sind. Von dieser Auswahl arbeiten 57 % mit dem Konzept des digitalen Zwillings in verschiedenen Formen, für konkrete Anwendungsfälle. Eine einheitliche, standardisierte technologische Umsetzung, etwa durch die Verwendung der Verwaltungsschale, ist für die Auswahl jedoch nicht gegeben. Bei einigen Plattformen werden die Begriffe „Things“ und „Things Modeling“ verwendet, wodurch eine plattformweit einheitliche Umsetzung erfolgt.

Was die Integration von Agenten betrifft, so wird bei einzelnen der betrachteten Plattformen der Begriff Agent verwendet. Mit diesen Softwaremodulen wird jedoch eher die Funktionalität der Plattform an sich hergestellt, z.B. mit der Crypto Extension für PTC-ThingWorx [23]. Bei Cumulocity IoT werden Agenten verwendet, um Kommunikationsprotokolle und Modelle in die

¹<https://jade.tilab.com/> ; <https://pade.readthedocs.io/en/latest/> ; <https://www.akka-technologies.com/>

eigene, plattformweit einheitliche Form zu transformieren und eine sichere Kommunikation herzustellen [24]. Eine Verwendung von Agenten im Sinne der Repräsentanz von Produkten, Prozessen oder Ressourcen, die miteinander interagieren und entsprechende Entscheidungen treffen, liegt bei diesen Plattformen nicht vor.

Eines der bekanntesten Industrie 4.0-Frameworks ist BaSys 4.0² zusammen mit der Referenzimplementierung Eclipse BaSyx³. Damit werden Konzepte der Plattform Industrie 4.0⁴ implementiert. Als technologische Umsetzung des digitalen Zwillings wird hier die Verwaltungsschale verwendet. Als Open Source Plattform bietet es die Möglichkeit spezielle Funktionalitäten, wie die Einbindung von Agenten, darauf aufzubauen.

3. Konzepte

Dieses Kapitel stellt das entwickelte Agentensystem, die genutzten Digitalen Zwillinge, die entwickelte Laufzeitumgebung sowie das integrierte Ökosystem vor.

3.1 Design des Agentensystems

Der vorgestellte Ansatz fokussiert Agenten, die Ressourcen und Produkte repräsentieren. Für beide Arten von physischen Entitäten gibt es in der Regel Digitale Zwillinge (DZ), während Prozess-DZ weniger verbreitet sind. Ressourcenagenten (RA) sind geeignet, alle Arten von Ressourcen zu repräsentieren, von einzelnen Maschinen bis hin zu ganzen Anlagen. Produktagenten (PA) hingegen repräsentieren die Produkte. Auf allen Ebenen muss ein RA den PA auf Anfrage Produktionsfähigkeiten anbieten. Dabei muss der RA den aktuellen Status der physischen Ressource berücksichtigen, z.B. wenn die Fähigkeiten aufgrund von Degradation eingeschränkt sind. Darüber hinaus muss der RA die eigentlichen Produktionsprozesse initiieren, indem er die physische Ressource ansteuert. Falls ein RA untergeordnete RAs hat, dient er zudem als Directory Facilitator. PAs müssen daher die Fähigkeiten, die zur Realisierung der gewünschten Produkteigenschaften erforderlich sind, bei den RAs anfordern. Um ihre Aufgaben zielgerichtet zu erfüllen, müssen die unterschiedlichen Agenten den Status des Produkts berücksichtigen und Entscheidungen über die Planung der nächsten Schritte im Produktionsprozess treffen. Dazu benötigen sie verschiedene Informationen, die von der Beschreibung ihrer Fähigkeiten (RA) bis zum Produktionsstatus (PA) reichen. Diese müssen im Digitalen Zwilling abgebildet werden.

² <https://www.basys40.de/>

³ <https://www.eclipse.org/basyx/>

⁴ <https://www.plattform-i40.de/>

Um der hierarchischen Struktur von Produktionssystemen gerecht zu werden wird eine rekursive Architektur für RAs genutzt, welche von holonischen Systemen [25] inspiriert ist. Diese wird durch PAs ergänzt [6]. Ein RA kann aus einer beliebigen Anzahl von untergeordneten RAs bestehen. Beispielsweise fasst der RA einer ganzen Fabrik die RAs der einzelnen Maschinen zusammen. Als Einstiegspunkt für die RA-Struktur wird ein einzelner übergeordneter RA, der Koordinator, festgelegt. Dieser Koordinator ist dafür verantwortlich, die Erstellung aller ihm untergeordneten RAs zu initiieren, dient aber auch als Directory Facilitator für PAs, das heißt er fungiert als Verzeichnis der RAs und bietet einen Überblick über vorhandenen Fähigkeiten. Folglich können PAs Dienste von diesem RA der obersten Ebene anfordern, der die Adressen der entsprechenden untergeordneten RAs liefert. Außerdem können PAs die Reihenfolge der Ressourcennutzung auf der Grundlage der ihnen zugewiesenen Priorität aushandeln, falls sie dieselbe Ressource benötigen. Alle Agenten innerhalb des MAS sind über ihren jeweiligen DZ mit der übergeordneten Anlage verbunden und verfügen über eine Wissensbasis, die aus dem DZ generiert wird. Die Wissensbasis jedes Agenten enthält das Umweltmodell des Agenten, d. h. seine Überzeugungen und eine Zielfunktion zur Entscheidungsfindung.

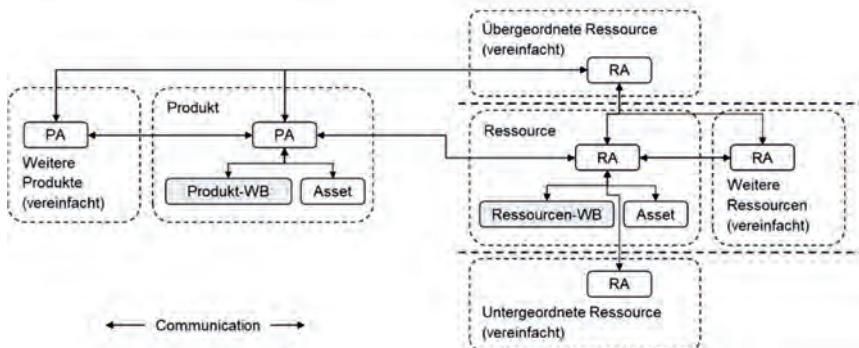


Bild 2: Rekursive Architektur für das MAS [6]

Im Vergleich zu vorherigen Arbeiten werden die Agenten nicht mehr aus den statischen Digitalen Zwillingen i.S.v. AASX-Dateien initialisiert. Stattdessen werden die reaktiven Digitalen Zwillinge aus dem Ökosystem genutzt, welche an die physischen Assets angebunden sind.

3.2 Digitaler Zwilling

Hiermit wird ein Überblick gegeben, wie die von den Agenten benötigten Informationen im Digitalen Zwilling, in Form der Verwaltungsschale, abgebildet werden können (Tabelle 1). Das VWS-Metamodell bietet explizit eine Klasse zur Darstellung von Ressourcenfähigkeiten, um

"implementierungsunabhängige Beschreibungen des Potenzials eines Assets, einen bestimmten Effekt in der physischen oder virtuellen Welt zu erzielen" [26]. Die allgemeinere Klasse *SubmodelElement* kann zur Beschreibung und Unterscheidung von Assets verwendet werden und eignet sich daher zur Erfassung der Zielfunktion eines Agenten oder eines für ein Produkt erforderlichen Prozesses. Mehrere *SubmodelElements*, welche benötigte Prozesse repräsentieren, können über eine *SubmodelElementCollection* aggregiert werden. Einzelne Attribute können im Metamodell der VWS durch *Properties* und *Ranges* repräsentiert werden. Zusammen sind diese Klassen ausreichend, um die von RAs und PAs benötigten Informationen zu erfassen.

Tabelle 1: Übersicht der von den verschiedenen Agenten benötigten Informationen [6]

Agententyp	Information	VWS-Klasse
Ressourcenagent	Fähigkeit der Ressource	Capability
	Zielfunktion	SubmodelElement
	Parameter: Kosten, Status, Information	Property / Range
Produktagent	Menge der benötigten Prozesse	SubmodelElementCollection
	Statische Parameter wie Priorität, Variablen, Produktionsstatus	SubmodelElement
		Property / Range

3.3 Laufzeitumgebung

Die Laufzeitumgebung soll drei Kernanforderungen erfüllen. Zum einen soll sie die Basis für Vernetzung zwischen Produkten und Ressourcen schaffen, Offenheit der genutzten Schnittstellen gewährleisten und zudem flexibel einsetzbar sein.

Um ein Wertschöpfungsnetzwerk, wie in Abschnitt 2.1 beschrieben, technisch umzusetzen, muss eine Vernetzung von Komponenten, Anlagen, IT-Systemen und -Applikationen, sowie Agenten erfolgen. Basis hierfür sind Verwaltungsschalen als technologische Umsetzung des digitalen Zwillings. Dabei muss die Laufzeitumgebung mit jeder Variante von Verwaltungsschalen (passiv, reaktiv, proaktiv) umgehen können. Die passive Verwaltungsschale, beschreibt dabei u.a. die Vernetzungsmöglichkeiten der einzelnen Entitäten sowie die beschriebenen Informationen, vgl. Abschnitt 3.2, für die Einbindung von Agentensystemen. Sie ist zunächst nicht mehr als ein statisches Austauschformat. Kommt zur passiven Verwaltungsschale Software hinzu, die eine Schnittstelle zu den Daten bietet, wird von einer reaktiven Verwaltungsschale gesprochen. Prinzipiell enthält die reaktive Verwaltungsschale eine eigene Laufzeitumgebung als Softwareebene, das heißt hier muss eine Vernetzung auf Software-Ebene

erfolgen. Erweitert man die Reaktive Verwaltungsschale um Softwaremodule zur Entscheidungsfindung oder Autonomie, wie z.B. Agenten, wird von einer proaktiven Verwaltungsschale gesprochen. Hier erfolgt die Vernetzung ebenfalls auf einer Softwareebene, allerdings über die Agenten.

Verwaltungsschalen sind ein Baustein für die Offenheit der Laufzeitumgebung. Ein weiterer Baustein sind SDKs. Diese können in verschiedenen Sprachen bereitgestellt werden, so dass die Laufzeitumgebung flexibel um spezielle Funktionalitäten erweitert werden kann. Zum Beispiel könnten so Frameworks Dritter integriert (z.B. PADE für Agenten), oder Softwaresysteme (z.B. ERP) angebunden werden.

Mit flexibler Einsetzbarkeit ist hier gemeint in welchem Kontext die Laufzeitumgebung konkret eingesetzt wird, das heißt ob von einer Edge- oder einer Cloud-Computing Architektur die Rede ist. Anlagenkomponenten können ihre eigene Datenverarbeitungstechnik mitbringen. Auch darauf könnte eine proaktive Verwaltungsschale implementiert werden. Assets, die keine Datenverarbeitungstechnik mitbringen, könnten in der Cloud gehostet werden. Solche Assets sind nicht nur Anlagenkomponenten, sondern zum Beispiel auch Produktzwillinge, für die es in der Entwicklungsphase noch kein reales Gegenstück gibt.

4. Umsetzung des Ökosystems für einen agenten- und DZ-basierten Anlagenbetrieb

Basierend auf den beschriebenen Konzepten, wird in diesem Abschnitt die Implementierung der Laufzeitumgebung beschrieben. Dabei wird auf die Architektur, den Aufbau der notwendigen DZ, die Umsetzung des MAS, und die Anbindung der DZ an die realen Ressourcen sowie der DZ an die Agenten eingegangen.

Die Laufzeitumgebung basiert auf einer flexiblen, granularen "cloud native" Microservices-Architektur. Jeder Microservice, jede App und generell jede Softwarekomponente wird in Containern ausgeführt. So können Agenten die von Kunden oder Integratoren programmiert wurden sicher und resilient in eigene Container eingebunden werden. Orchestriert werden die Container durch Kubernetes⁵. Je nach Bedarf müssen nur die benötigten Komponenten installiert werden. Diese Eigenschaften erlauben es die Laufzeitumgebung sowohl auf kleinsten Edge-Computern als auch in unternehmenseigenen oder Public-Clouds zu betreiben. Diese Architektur ist besonders interessant für KMUs, da sie es ihnen ermöglicht mit wenig Know-how und marktüblichen Tools ihre eigenen Agenten zu entwickeln und sicher zu integrieren.

⁵ <https://kubernetes.io/>



Bild 3: Anbindung der einzelnen Agenten über die Digitalen Zwillinge an die physischen Assets

Die Anbindung von Agenten, Verwaltungsschalen und Assets erfolgt wie in Abbildung 3 dargestellt über API-Calls bzw. OPC-UA. Als MAS-Framework wurde PADE⁶ angebunden. Die relevanten Typen von Agenten (Produktagent, Ressourcenagent) sind in Form von Klassen implementiert, die auf der generischen Agentenklasse aus PADE basieren.

Um das MAS zu initialisieren werden Agenten für die im System vorhandenen Produkte und Ressourcen angelegt. Dafür werden die Agentenklassen entsprechend den vorhandenen DZ instanziiert und die jeweils relevanten Informationen aus den DZ werden in ihren Wissensbasen bereitgestellt. Durch die hierarchische Struktur der Ressourcen ist hier ein kaskadierender Prozess möglich, bei dem der Koordinator die Initialisierung der ihm untergeordneten Agenten anstößt. Der Ablauf des Initialisierungsprozesses ist in Abbildung 4 zusammengefasst. Als initiale Information ist neben dem Zugriff auf die Laufzeitumgebung lediglich eine Liste mit den eindeutigen IDs der DZ notwendig, für die Agenten initialisiert werden sollen.

Als Proof of Concept wurden für einen Demonstrator (xPPU) die notwendigen Digitalen Zwillinge exemplarisch mit dem AASX Package Explorer⁷ erstellt, vgl. Abbildung 5. Hierbei wurde der Fokus auf die für die Agenten relevanten Teile des DZ gelegt, insbesondere also die Fähigkeiten der Ressourcen und die benötigten Prozessschritte aus Sicht des Produkts.

⁶ <https://pade.readthedocs.io/en/latest/>

⁷ <https://github.com/admin-shell-io/aasx-package-explorer>

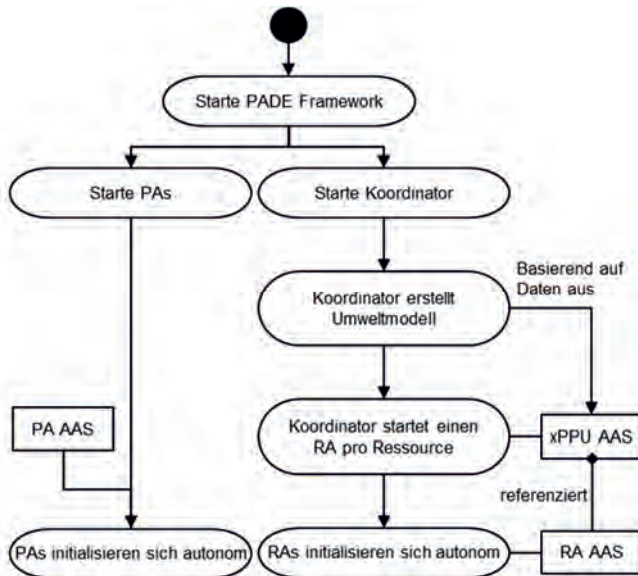


Bild 4: Ablauf der Initialisierung des MAS durch den Koordinator

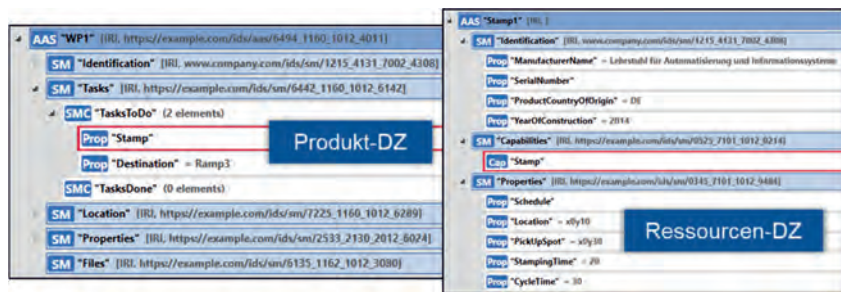


Bild 5: Ausschnitte der Verwaltungsschalen für ein Werkstück (Produkt-DZ) und einen Stempel (Ressourcen-DZ) in der xPPU [6]

5. Zusammenfassung und Ausblick

Es ein Ansatz zur Realisierung moderner Produktionssysteme durch die Kombination von Digitalen Zwillingen mit Agenten vorgestellt und die Umsetzung einer entsprechenden Laufzeitumgebung beschrieben. Letztere stellt das bisher fehlende Bindeglied zwischen der Maschine,

ihrem Digitalen Zwilling und ihrem Agenten dar. Sie ermöglicht zudem die Kommunikation zwischen Agenten und bietet APIs zu Softwaresystemen eines jeden Unternehmens, wie z.B. ERP-Systemen. Dadurch ist es integraler Bestandteil des Wertschöpfungsnetzwerks. Der Fokus des Beitrages lag darauf, wie Digitale Zwillinge zum Enabler für Agentensysteme werden können. Die entwickelte Laufzeitumgebung bietet zusammen mit ergänzend entwickelten Applikationen viele weitere Möglichkeiten zum Erstellen und Betreiben von Digitalen Zwillingen. Zum einfachen Erstellen wurde u.a. bereits der Import aus Systemen wie SAP-AIN umgesetzt. Insgesamt bildet dies ein Ökosystem, welches hochflexible Wertschöpfungsnetzwerke ermöglicht und der Industrie deren Nutzung leicht zugänglich macht.

6. Projektreferenzen

Diesem Beitrag zugrundeliegende Arbeiten wurden mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS18075A-F gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren.

7. Literaturverzeichnis

- [1] J. Arlinghaus, „Interview: Resilienz ist wichtiger als Effizienz“, *atp*, Bd. 11-12/2021.
- [2] Plattform Industrie 4.0, „Resilienz im Kontext von Industrie 4.0“. April 2022. [Online]. Verfügbar unter: <https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/Resilienz.html>
- [3] A. Belyaev u. a., „Modelling the Semantics of Data of an Asset Administration Shell with Elements of ECLASS“. 29. Juni 2021. [Online]. Verfügbar unter: https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/Whitepaper_Plattform-Eclass.pdf
- [4] M. Hoffmeister, „Die Industrie 4.0 Komponente“. April 2015.
- [5] F. Ocker, C. Urban, B. Vogel-Heuser, und C. Diedrich, „Leveraging the Asset Administration Shell for Agent-Based Production Systems“, *IFAC-Pap.*, Bd. 54, Nr. 1, S. 837–844, Jan. 2021, doi: 10.1016/j.ifacol.2021.08.186.
- [6] B. Vogel-Heuser, F. Ocker, und T. Scheuer, „An approach for leveraging Digital Twins in agent-based production systems“, - *Autom.*, Bd. 69, Nr. 12, S. 1026–1039, Dez. 2021, doi: 10.1515/auto-2021-0081.
- [7] F. Tao, H. Zhang, A. Liu, und A. Y. C. Nee, „Digital Twin in Industry: State-of-the-Art“, *IEEE Trans. Ind. Inform.*, Bd. 15, Nr. 4, S. 2405–2415, 2019, doi: 10.1109/TII.2018.2873186.
- [8] Plattform Industrie 4.0, *Ergebnispapier: Forschungsagenda Industrie 4.0 - Aktualisierung des Forschungsbedarfs*. Berlin, 2016. [Online]. Verfügbar unter: <https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/forschungsagenda-i40.pdf>

- [9] S. Himstedt, „Offene Systemgrenzen für das digitale Abbild“, *IT&Production*, 17. Mai 2019. <https://www.it-production.com/fertigungsnahe-it/werksnahe-it-und-der-digitale-zwilling/> (zugegriffen 11. Mai 2020).
- [10] „ISO/IEC AWI 30172 - Digital Twin - Use cases“, *International Organization for Standardization*. <https://www.iso.org/standard/81578.html> (zugegriffen 2. Mai 2022).
- [11] „ISO/IEC AWI 30173 - Digital Twin - Concepts and terminology“, *International Organization for Standardization*. <https://www.iso.org/standard/81442.html> (zugegriffen 3. Mai 2022).
- [12] E. Bayrhammer u. a., „Verteilte Digitale Zwillinge – der Stand der Technik, Anwendungsfälle und Zielstellung“, gehalten auf der VDI Automation 2020, Juni 2020.
- [13] M. Wooldridge, „Intelligent Agents: The Key Concepts“, in *Multi-Agent Systems and Applications II*, Berlin, Heidelberg, 2002, S. 3–43. doi: https://doi.org/10.1007/3-540-45982-0_1.
- [14] VDI/VDE-Gesellschaft Mess- und Automatisierungstechnik, Hrsg., *VDI/VDE 2653 Blatt 1 - Multi-agent systems in industrial automation - Fundamentals*. 2018. Zugegriffen: 16. Mai 2022. [Online]. Verfügbar unter: <https://www.vdi.de/richtlinien/details/vdivde-2653-blatt-1-multi-agent-systems-in-industrial-automation-fundamentals>
- [15] L. A. Cruz Salazar, D. Ryashentseva, A. Lüder, und B. Vogel-Heuser, „Cyber-physical production systems architecture based on multi-agent's design pattern - comparison of selected approaches mapping four agent patterns“, *Int. J. Adv. Manuf. Technol.*, Bd. 105, Nr. 9, S. 4005–4034, Dez. 2019, doi: 10.1007/s00170-019-03800-4.
- [16] I. Kovalenko, D. Tilbury, und K. Barton, „The model-based product agent: A control oriented architecture for intelligent products in multi-agent manufacturing systems“, *Control Eng. Pract.*, Bd. 86, S. 105–117, Mai 2019, doi: 10.1016/j.conengprac.2019.03.009.
- [17] B. Vogel-Heuser, J. Lee, und P. Leitão, „Agents enabling cyber-physical production systems“, - *Autom.*, Bd. 63, S. 777–789, Jan. 2015, doi: 10.1515/auto-2014-1153.
- [18] S. Borgo, A. Cesta, A. Orlandini, und A. Umbrico, „Knowledge-based adaptive agents for manufacturing domains“, *Eng. Comput.*, Bd. 35, Juli 2019, doi: 10.1007/s00366-018-0630-6.
- [19] A. Rao und M. Georgeff, „BDI agents: From theory to practice“, in *ICMAS-95*, Juni 1995, S. 312–319.
- [20] F. Ocker, I. Kovalenko, K. Barton, D. Tilbury, und B. Vogel-Heuser, „A Framework for Automatic Initialization of Multi-Agent Production Systems Using Semantic Web Technologies“, *IEEE Robot. Autom. Lett.*, Bd. 4, Nr. 4, S. 4330–4337, Okt. 2019, doi: 10.1109/LRA.2019.2931825.

- [21] „IIoT platforms and ecosystems: Industrial IoT platform evolutions“, *i-SCOOP*. <https://www.i-scoop.eu/internet-of-things-iiot/industrial-iiot-platform-ecosystem/> (zugegriffen 5. Mai 2022).
- [22] S. Christian, E. Holger, A. A. Shayan, Dewes Andreas, und J. Jan, „Aktuelle Industrie 4.0 Plattformen – Eine Übersicht“, Jan. 2021, doi: 10.5281/ZENODO.4485756.
- [23] „IoT Security Platform | Our IoT Security Features | Device Authority“, *Device Authority Ltd*, 28. August 2021. <https://www.deviceauthority.com/platform/> (zugegriffen 6. Mai 2022).
- [24] „Interfacing devices - Cumulocity IoT Guides“. <https://cumulocity.com/guides/concepts/interfacing-devices/> (zugegriffen 6. Mai 2022).
- [25] A. Giret und V. Botti, „Holons and agents“, *J. Intell. Manuf.*, Bd. 15, S. 645–659, Okt. 2004, doi: 10.1023/B:JIMS.0000037714.56201.a3.
- [26] Plattform Industrie 4.0, Details of the Asset Administration Shell: Part 1 - The exchange of information between partners in the value chain of Industrie 4.0 (Version 3.0RC01). 2020. [Online]. Verfügbar unter: https://www.plattform-i40.de/PI40/Redaktion/DE/Downloads/Publikation/Details_of_the_Asset_Administration_Shell_Part1_V3.html

Hot Spot Detection in Electrical Equipment using the Inverse Pixel Count Algorithm

Dr. Ralf Gitzel, ABB, Ladenburg;
Tomas Kozel, ABB, Brno, Tschechien

Kurzfassung

Infrarotkameras sind ein nützliches Werkzeug um Defekte in elektrischen Geräten wie z.B. Schaltschränken zu entdecken. Dieser Artikel beschreibt einen Algorithmus, der heiße Stellen in Infrarotbildern erkennen kann. Diese heißen Stellen sind Indikatoren für versteckte Defekte. Der in diesem Paper beschriebene Algorithmus verwendet Bilder aus mehreren Bereichen des Schaltschranks und setzt einen intelligenten Schwellwert an um eine Kennzahl (den sog. Inverse Pixel Count) zu ermitteln. Mit diesem als Eingangsgröße erreicht eine Support Vector Machine eine Genauigkeit von 93.5%.

Abstract (optional)

Infrared cameras are a useful tool to monitor electrical equipment such as switchgear and to detect faults. This paper describes an algorithm that can detect hot spots in infrared images that indicate a large class of faults. The algorithm uses images from several switchgear compartments and applies an intelligent threshold to compute a feature called the 'inverse pixel count'. When used as input for a Support Vector Machine, an average accuracy of 93.5% was achieved.

1 Introduction

Faults in mission-critical electrical equipment can lead to downtime in production processes or technical infrastructure. For example, if a switchgear panel has a faulty connection, it can fail and interrupt an electrical distribution network. However, if potential faults are detected early, e.g., via condition monitoring, overall system failure can be prevented.

Thermal monitoring is one approach to condition monitoring in this context. It is common knowledge that many technical problems in electrical equipment show up as hot spots in an infrared thermography (IRT) image. For example, if a single phase in a switchgear panel suffers from a loose screw or corrosion, the phase will appear hotter than normal (cf. Fig. 1; the hot spot is visible near the upper edge of the image).

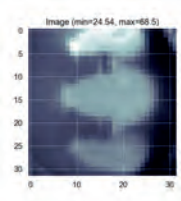


Fig. 1: Breaker compartment with a fault in the "upper" phase

A typical approach to automatically interpreting IRT images is to define regions of interest (ROIs) and monitor them. If a region becomes too hot, especially in comparison to other ROIs, there is probably a fault. A drawback of this approach is that a manual definition of the regions has to be performed. This requires a configuration phase where infrared images of the operating equipment are studied and annotated.

We propose an approach that automatically recognizes a single region of interest based on a size estimate provided by the user. Without comparison to other regions, the temperature difference found in the ROI can be used to determine the state of faulty or healthy.

We show how this feature can be used for machine learning with high accuracy. While the example data is from medium voltage switchgear, the algorithm can also be applied to other electrical equipment such as motors or cabling. The work done here is part of the Fleming public funded project, which is explained in detail in [20].

2 Infrared Monitoring of Electrical Equipment

Condition monitoring of electrical equipment using infrared cameras is not a new idea and there is a lot of prior art in the field. In this section, we quickly survey the existing approaches and how they compare to the solution described in this paper.

While it is easy for a human to detect faulty systems by identifying hot spots in an infrared image, most automated algorithms need features that can be used to classify such images. Probably the simplest feature is the absolute temperature at a given position in the image ([9], [45]). One problem with such approaches is that the temperature estimates in the IRT image are wrong by a certain unknown offset, so a dangerous temperature might not be recognized correctly. However, since this offset is similar at all positions in the image, a better classification accuracy is achieved by using the temperature difference between different zones in the image (e.g. [7], [8], [9], and [10]). A similar approach is to take a crosscut of the image, i.e., a graph that shows the pixel temperatures along a line. Typically, the analysis is based on a comparison of the peaks on that curve ([11], [12], [13]).

Other solutions involve calculating more complex metrics based on the image data. For example, the statistical moments of the temperature histogram can be used as input for classification algorithms [13]. Sometimes these features are applied not to the histogram of the image as a whole but to one or more regions of interest (see [3], [13], [14]). The second-order statistical features identified by Haralick et al (see [15]) support the analysis of IRT images [14]. Similar metrics are image moments, e.g., the Zernike moments discussed by Rahmani et al [16].

Features can be combined with hand-crafted thresholds but the majority of approaches seems to rely on machine learning algorithms to do the classification. Many papers use Support Vector Machines (e.g. [1], [2], [3], [4], [5]). There are also examples of neural networks ([1], [4]). Convolutional neural networks would seem like a natural match but they are not widely used in this context with the exception of [18] and [19]. Other ML-algorithms encountered in the related work are K-Nearest Neighbours ([2], [5]), fuzzy logic [6], and Random Forest [5].

While these solutions often give good results, we see two shortcomings that we would like to address with this paper (which is a more efficient version of the algorithm in our prior work [21]). First of all, the related work uses high-resolution infrared images. The cost of high-resolution cameras is such that it prohibits a permanent installation for most electrical equipment. (This is definitely true for the switchgear panels in our experiments.) Second, any equipment with a heterogeneous geometry for different models and manufacturers will make solutions involving a manual definition of regions of interest quite cost-intensive, especially in brown-field solutions. (Again, for switchgear, this is the case.) While some papers propose automatic recognition mechanisms (e.g. [5]), these are intended to address point-of-view issues and not different types of equipment. Also, there are algorithms that do not use features, e.g. SIFT [17]. However, it is not likely that these would work well on low-resolution images. The algorithm proposed in this paper does not require a high resolution nor does it require a fully defined region of interest. Instead of exact boundaries, it needs only an estimate of the region's size. As such, it addresses the two open issues we have identified. It extends our prior work [21] in that it is more intuitive to configure and use.

3 The Inverse Pixel Count Algorithm

The inverse pixel count is based on the pixel count algorithm introduced in a prior paper [21]. We will start this section with a high-level explanation of the pixel count algorithm's working principle. Based on this, we will then describe the inverse pixel count algorithm introduced in this paper.

An IRT image is essentially a matrix of temperature values that can be visualized as a grayscale image (see Fig. 2). The example image shows the phases of a medium-voltage switchgear which heat up because current is flowing through them.

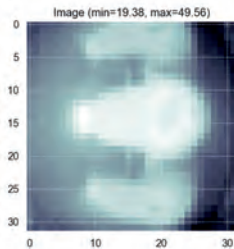


Fig. 2: Example IRT Image

When a threshold is applied to the image, it is separated into two regions – a region where the temperature is equal to or above the threshold t and the region where it is below t . In Fig. 3, the IRT image is shown again on the right. Its maximum temperature in the image is about 50°C. The image in the middle shows the result of applying a threshold of 42°C. All pixels above 42°C are marked white, anything else is marked black. In the image on the right, a lower threshold (this time 35°C) is used, so more pixels are included.

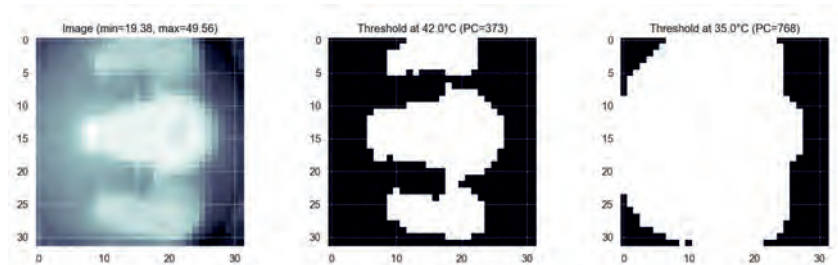


Fig. 3: Applying a threshold

In the example shown above, a threshold of about 42°C (ca. 8°C below the maximum temperature) roughly reproduces the phases while removing the background (see middle image). When the phases are fully visible, the number of pixels above the threshold is 373 in this example. This value is called the **pixel count**. The lower threshold on the right has a higher pixel count – 768.

In the middle image, all three phases and none of the background is part of the white region. This means that the lowest temperature of the phases is 42°C and their hottest point is 50°C , the maximum temperature of the image. Thus, the temperature interval on the phases is 8°C . The difference between phase temperatures is a common classification feature found in related work.

When we apply the same threshold used in Fig. 3 to an image with roughly the same maximum temperature but with a hot spot (Fig. 4, left), the pixel count will be lower than for the healthy case in Fig. 3. The reason is that the “healthy” phases are still cooler than the ones in the completely healthy image. The hot spot by definition has a higher temperature than the other phases, so the total temperature range must be greater than 8°C to include all phases. This is the basic working principle of the original pixel count algorithm described in [21] – if for the same maximum temperature and the same temperature interval the pixel count is lower, there is a hot spot.

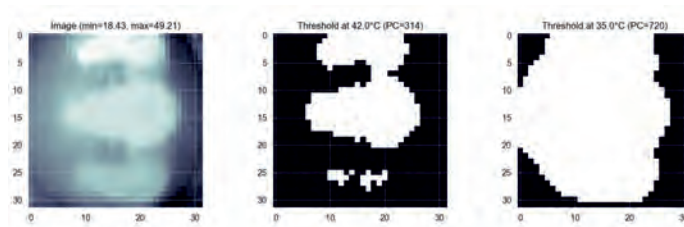


Fig. 4: Applying the same thresholds to a hot spot image

Additional experiments with the pixel count algorithm have shown some of its shortcomings. In the algorithm, the threshold is always set relative to the maximum temperature. However, finding a good rule for this temperature interval's size is not easy. A fixed interval will cause problems in low-current scenarios. We have proposed an interval that changes with the maximum temperature of the image. However, the proper parameterization is non-intuitive. Thus, we propose another algorithm we call the *inverse pixel count (IPC)*. Instead of defining a temperature interval and count the pixels we take the opposite approach. For the IPC, we define a fixed pixel count target number (based on the size of the phases) and determine the size of the interval that results in that value for the pixel count. In Fig. 5, which shows the same hot spot image as Fig. 4, the threshold has been gradually lowered until the pixel count is as close to the target value that fully shows all three phases. In this process, the phases emerge like an object in a tub of water. The resulting image will look roughly the same for healthy and

for faulty cases. In fact, the image looks similar to the middle one in Fig. 3, which shows a healthy image.

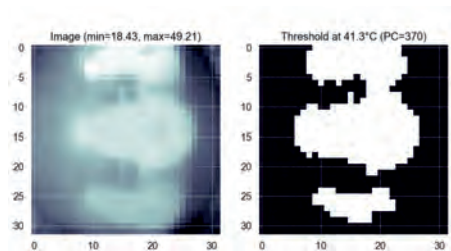


Fig. 5: Inverse Pixel Count (I)

The main difference to Fig. 3, however, is that the temperature interval is slightly larger (9°C vs 8°C). As the hot spot grows in intensity, the required interval to include the target number of pixels increases as well. In our example, it goes up to 21°C as the maximum temperature reaches 68.5°C (Fig. 6). Thus, the size of the interval can be used to classify faulty and healthy cases. If the temperature interval is low, the temperature of the phases is homogeneous. If the interval is high, there is a hot spot.

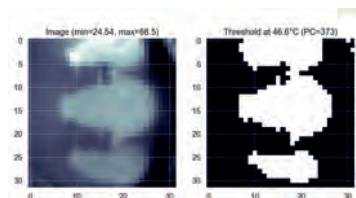


Fig. 6: Inverse Pixel Count (II)

The algorithm can be summarized by the five steps shown below. While more efficient and accurate implementations are definitely possible, this description best illustrates the concept:

1. Define a starting interval size (e.g. 0.5°C), a maximum interval size (e.g. 25.0°C), and the step size (e.g. 0.5°C). The interval size is called X in this description. The step size is dx .
2. Define a target mask size t that corresponds to the size of the three phases.
3. Compute the t_x , the pixelcount value for X .
4. If t_x is greater than t , use the previous X as result and stop. (In other words, the last pixel count that was smaller than the target mask size t .)
5. If t_x is less than or equal to t , increase X by dx and go back to step 3.

4 Machine Learning

While a manual definition of value ranges that indicate healthy and faulty cases is possible, the nature of the classification problem makes the use of machine learning attractive. Experimental data was used to train a Support Vector Machine (SVM). The 8 features used where the maximum temperature in the image and the inverse pixel count for each compartment for all time steps.

4.1 Data and Separation

The data available to us in this paper for training and testing the SVM consisted of 73 experiments, where each is a time series of images from 4 compartments of a switchgear. Since we had access to a line-up including two panels with suitable cameras this gave us two groups of experiments to work with. Some experiments contain a fault (so they are consistently labeled as faulty) others use an intact switchgear and are labeled healthy.

In order to avoid data leakage, we separated the data into training and testing data by full experiments. Also, since the data from the two panels is different, we trained separate SVMs for each panel.

4.2 Results

We did not do a full crossvalidation but we trained a total of 6 SVMs using different experiments as test data. Due to the 2 panels, this means 3 SVMs per panel. We used Python's `sklearn.svm.SVC` with default parameters.

The table below shows a more detailed analysis of experiment performance. The experiment name indicates where the fault was found in the experiment used for testing. Another, healthy experiment was added to the test data as well. The remaining 71 experiments were used as training data. Note that the busbar experiment in panel 2 suffers from heavy class imbalance, so the precision and recall are not very meaningful.

Table 1: Precision and Recall of Experiments

Experiment	Actual Faults	True Positives	False Positives	Precision (Faults)	Recall (Faults)
Panel 1, bb	683	622	22	96,6%	91,1%
Panel 1, bc l2	295	285	23	92,5%	96,6%
Panel 1, bc l3	230	225	23	90,7%	97,8%
Panel 2, bb	34	15	173	8,0%	44,1%
Panel 2, bc l2	1154	1143	170	87,1%	99,0%
Panel 2, bc l3	509	509	173	74,6%	100,0%

The median of these results is shown in Fig. 7. **Accuracy** is a measure to determine how well the algorithm classified the test data. 1 means that all data was classified correctly, 0 that all

were done wrong. **Precision** looks at all elements that were classified as faulty and tells us the percentage of those that were actual faults. So, a high precision means that most or all of the alarms for faulty are justified. **Recall** tells you how many of the actual faults were recognized. The ideal algorithm has a high precision and recall. An algorithm that only spots very strong and clear faults will have a high precision but low recall. An algorithm that does not miss any fault but will give false alarms has higher recall but lower precision.

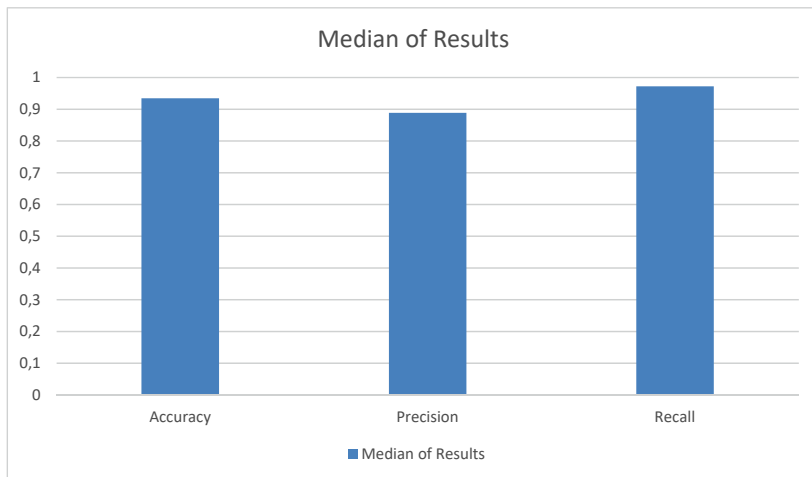


Fig. 7: Overall SVM algorithm performance

The algorithm's performance is quite good. The one "outlier" in the otherwise good performance of the algorithm is "panel 2, broken bus bar". The accuracy is the lowest and both precision and recall are unacceptable. The confusion matrix (Fig. 8) shows that there is a strong class imbalance, i.e., there are far more healthy than faulty cases. Second, there are very few faults in absolute numbers, so it is probably unrealistic to expect a good performance.

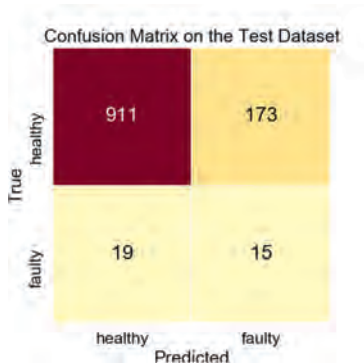


Fig. 8: Worst machine learning performance - confusion matrix

5 Conclusions

In this paper, we have discussed a new algorithm to detect hot spots in medium voltage switchgear. The *inverse pixel count* allows to identify hot spots in low resolution IRT images without the exact definition of the region of interest. The algorithm achieves good results on the test data which was derived from multiple experiments in the laboratory. Future work includes the automatic identification of the proper pixel count value to identify the phases without human intervention to make the algorithm fully adaptable.

6 Acknowledgements

The authors acknowledge Dietmar Treibert, Vadim Migunov, Adrian Hozoi, and Holger Kaul for technical help with the experiments and Armin Gasch, Oliver Becker, and Aydin Boyaci for help with the project management.

- [1] A. Rahmani; J. Haddadnia; O. Seryasat (2010): Intelligent fault detection of electrical equipment in ground substations using thermo vision technique. In : 2010 2nd International Conference on Mechanical and Electronics Engineering, vol. 2, pp. V2-150-V2-154.
- [2] N.T. Moja, A.J. Willis, Classification of thermally condition-monitored components using statistical and neural network techniques, in: A.G. Tescher (Ed.), Proceedings of SPIE 3164: Applications of Digital Image Processing XX, San Diego, CA, USA, 1997, pp. 573–581.

- [3] Bai, Tangbo; Zhang, Laibin; Duan, Lixiang; Wang, Jinjiang (2016): NSCT-based infrared image enhancement method for rotating machinery fault diagnosis. In *IEEE Transactions on Instrumentation and Measurement* 65 (10), pp. 2293–2301.
- [4] Jadin, Mohd Shawal; Taib, Soib; Ghazali, Kamarul Hawari (2014): Feature extraction and classification for detecting the thermal faults in electrical installations. In *Measurement* 57, pp. 15–24.
- [5] Janssens, O.; Rennuy, M.; Devos, S.; Loccufier, M.; R. Van de Walle; van Hoecke, S. (2016): Towards intelligent lubrication control: Infrared thermal imaging for oil level prediction in bearings. In : 2016 IEEE Conference on Control Applications (CCA), pp. 1330–1335.
- [6] A.T.P. So, W.L. Chan, C.T. Tse, K.K. Lee, Fuzzy logic based automatic diagnosis of power apparatus by infrared imaging, in: Proceedings of the Second International Forum on Applications of Neural Networks to Power Systems, 1993, pp. 187–192.
- [7] Bagavathiappan, S.; Lahiri, B. B.; Saravanan, T.; Philip, John; Jayakumar, T. (2013): Infrared thermography for condition monitoring – A review. In *Infrared Physics & Technology* 60, pp. 35–55.
- [8] Niancang Hou (1998): The infrared thermography diagnostic technique of high-voltage electrical equipments with internal faults. In : POWERCON '98. 1998 International Conference on Power System Technology. Proceedings (Cat. No.98EX151), vol. 1, pp. 110-115 vol.1.
- [9] Smedberg, Magnus (2006): Thermographic decision support-detecting and classifying faults in infrared images. In *Master of Science Thesis, Royal Institute of Technology, Stockholm, Sweden*.
- [10] Snell Jr, John R.; Spring, Robert W. (2003): The new approach to prioritizing anomalies found during thermographic electrical inspections. In : Thermosense XXV, vol. 5073. International Society for Optics and Photonics, pp. 222–230.
- [11] Jadin, Mohd Shawal; Taib, Soib (2012): Recent progress in diagnosing the reliability of electrical equipment by using infrared thermography. In *Infrared Physics & Technology* 55 (4), pp. 236–245.
- [12] Ryuichi Ishino (2002): New detection method of faulty distribution power apparatus using thermal images. In Xavier P. Maldague, Andres E. Rozlosnik (Eds.): Thermosense XXIV, vol. 4710. International Society for Optics and Photonics: SPIE, pp. 353–362. Available online at <https://doi.org/10.1117/12.459584>.

- [13] Huda, A. S. Nazmul; Taib, Soib (2013): Application of infrared thermography for predictive/preventive maintenance of thermal defect in electrical equipment. In *Applied Thermal Engineering* 61 (2), pp. 220–227.
- [14] A.S. Nazmul Huda; Soib Taib; Mohd Shawal Jadin; Dahaman Ishak (2012): A semi-automatic approach for thermographic inspection of electrical installations within buildings. In *Energy and Buildings* 55, pp. 585–591.
- [15] R. M. Haralick; K. Shanmugam; I. Dinstein (1973): Textural Features for Image Classification. In *IEEE Transactions on Systems, Man, and Cybernetics* 3 (6), pp. 610–621.
- [16] A. Rahmani; J. Haddadnia; O. Seryasat (2010): Intelligent fault detection of electrical equipment in ground substations using thermo vision technique. In : 2010 2nd International Conference on Mechanical and Electronics Engineering, vol. 2, pp. V2-150-V2-154.
- [17] E. Borenstein; E. Sharon; S. Ullman (2004): Combining Top-Down and Bottom-Up Segmentation. In : 2004 Conference on Computer Vision and Pattern Recognition Workshop, p. 46.
- [18] Du, X.; Si, S.; and Li, Y. 2019. Fault Diagnosis of Gear-box based on Convolutional Neural Network and InfraredThermal Imaging. In 2019 Prognostics and System HealthManagement Conference (PHM-Qingdao), 1–4. IEEE.
- [19] Fanchiang, K.-H.; Huang, Y.-C.; and Kuo, C.-C. 2021.Power Electric Transformer Fault Diagnosis Based on In-frared Thermal Images Using Wasserstein Generative Ad-versarial Networks and Deep Learning Classifier.Electron-ics, 10(10): 1161
- [20] Hoffmann, M. W.; Wildermuth, S.; Gitzel, R.; Boyaci, A.;Gebhardt, J.; Kaul, H.; Amihai, I.; Forg, B.; Suriyah, M.;Leibfried, T.; Stich, V.; Hicking, J.; Bremer, M.; Kaminski, L.; Beverungen, D.; zur Heiden, P.; and Tornede, T. 2020.Integration of Novel Sensors and Machine Learning for Pre-dictive Maintenance in Medium Voltage Switchgear to En-able the Energy and Mobility Revolutions.Sensors, 20(7):2099.
- [21] Gitzel, R.; Gebhard, J.; Kozel, T. (2021): Automatic Analysis of Thermograms – Challenge in Thermal Monitoring of Switchgears Using Infrared Cameras. In: proceedings of CIRED 2021, Geneva

Automatisierung des manuellen Löschvorgangs bei Holzbränden

Fabian Stoller, Prof. Dr.-Ing. **Alexander Fay**,
Helmut-Schmidt-Universität, Hamburg;
Felix Kümmerlen, Wehrwissenschaftliches Institut für
Schutztechnologien – ABC-Schutz (WIS), Munster

Kurzfassung

Ein zentraler Bestandteil des Brandschutzes sind tragbare Feuerlöscher. Ihr Löschvermögen ist im Rahmen der Zulassung von einem Menschen in einem genau definierten Versuch zu überprüfen. Die Untersuchung und Automatisierung dieses Versuchs versprechen die Zertifizierung tragbarer Feuerlöscher in Zukunft nachvollziehbarer und besser wiederholbar zu machen. Zudem bildet dieser Versuch eine gute Grundlage für die wissenschaftliche Untersuchung des automatisierten Löschsens von Feststoffbränden. Dafür wird in diesem Beitrag der Löschvorgang untersucht, wie ein Mensch ihn durchführt. Dies bildet eine Basis, um den Löschvorgang so zu modellieren und anzupassen, dass er von einem automatisierten Löschsystem durchgeführt werden kann. Dafür werden die Höhe von Flammen sowie die Temperatur und Größe von Glutnestern als essenzielle Merkmale für die Steuerung des Löschvorgangs identifiziert. Sowohl die Flammen als auch die Glutnester werden im Rahmen realer Brandversuche zu für den Löschvorgang entscheidenden Zeitpunkten untersucht.

1. Einleitung

Tragbare Feuerlöscher sind ein zentraler Bestandteil von Brandschutzmaßnahmen an Arbeitsstätten und in öffentlichen Gebäuden. Die Menge an Feuerlöschern, die bereitzustellen ist, richtet sich nach der jeweiligen Brandgefahr und nach dem Löschvermögen der eingesetzten Feuerlöscher [1]. Die DIN EN 3-7 [2] beschreibt die Anforderungen sowie die für eine Zertifizierung erforderlichen Prüfungen von tragbaren Feuerlöschern. Für die Prüfung des Löschvermögens eines tragbaren Feuerlöschers schreibt die Norm das erfolgreiche Löschen eines genau beschriebenen Prüfobjekts vor. Die DIN EN 3-7 [2] definiert für verschiedene Brandklassen jeweils eigene Prüfobjekte. Diese Arbeit bezieht sich auf das Löschen des Normbrands der Brandklasse A (Feststoffe, die unter Bildung von Glut brennen [3], wie z. B. Holz). Die Norm definiert neben dem Aufbau des Prüfobjekts auch die Umgebungsbedingungen, unter denen der Versuch durchzuführen ist, und den Ablauf der Versuchsdurchführung. Das Löschen selbst ist jedoch von menschlichen Löschmeistern durchzuführen, die auf Basis ihrer

jeweiligen Erfahrung ihre Vorgehensweise beim Löschen des Testbrandes an die begrenzt verfügbare Menge an Löschmittel in einem Feuerlöscher angepasst haben. Damit sie das Feuer effektiv bekämpfen können, differenzieren sie zwischen Flammen und Glut und wenden jeweils unterschiedliche Löschtechniken an, um diese zu bekämpfen. Mit dem Ziel, die Qualität dieses Versuchs und damit der Zertifizierung zu verbessern, soll dieser in der DIN EN 3-7 [2] definierte Versuch automatisiert werden. Zusätzlich bietet dieser Versuch durch seine engen Vorgaben für die Durchführung eine reproduzierbare Grundlage für die weitere Untersuchung des automatisierten Löschens von Holzbränden.

In diesem Beitrag wird hergeleitet, wie der manuelle Vorgang der Brandbekämpfung bei einem Brandversuch auf ein automatisiertes System übertragen werden kann. Dafür wird untersucht, wie sich ein Brand mit einem Kamerasystem nach [4] überwachen lässt, um normgerechte Prüfbjekte in einem gezielten und koordinierten Löschvorgang erfolgreich zu löschen. Eine zentrale Randbedingung ist dabei die Verwendung handelsüblicher tragbarer Feuerlöscher mit ihrer begrenzten Menge an Löschmittel.

Dieser Beitrag ist wie folgt strukturiert: Zuerst wird in Abschnitt 2 der Stand der Forschung, aufgeteilt auf die Bereiche der Löschrategie bei automatisierten Löschsystemen sowie die Bewertung der Intensität von Bränden, präsentiert. Anschließend wird in Abschnitt 3 beschrieben, welche Messgrößen für die Steuerung eines Löschvorgangs im Falle des Normbrandversuchs geeignet sind. Auf deren Basis wird anschließend ein Modell für einen entsprechenden Löschvorgang präsentiert. Abschnitt 4 enthält eine Untersuchung der ausgewählten Messgrößen anhand realer Brandversuche. Abschließend wird in Abschnitt 5 ein Fazit gezogen und ein Ausblick auf zukünftige Arbeiten gegeben.

2. Stand der Forschung

2.1. Das automatisierte und gezielte Löschen von Bränden

Es gibt verschiedene Ansätze zur automatisierten und gezielten Bekämpfung von Bränden. Ein wiederholt verwendeter Ansatz ist die Ausstattung einer steuerbaren Wasserkanone mit einem geeigneten Kamerasystem [5, 6]. In [6] werden zum Beispiel mehrere Aufnahmen einer einzelnen Kamera genutzt, um die Position des Brandes im Raum relativ zur Kamera zu ermitteln und diese Position als Zielpunkt für die Wasserkanone zu verwenden. Während des Löschens wird die Wasserkanone um $\pm 5^\circ$ um den so bestimmten Zielpunkt herumbewegt. Auch in [5] wird nur eine einzelne Kamera verwendet. Doch hier wird neben dem sichtbaren Spektrum des Lichts auch ein Teil des Infrarotspektrums genutzt, um das Feuer und den Landepunkt des Löschmittelstrahls zu detektieren. Die Ausrichtung der Wasserkanone wird in der

Folge so geregelt, dass der Landepunkt des Löschmittels und die Position des detektierten Feuers zusammenfallen.

Ein Stereokamerasystem zur Lokalisierung von Bränden im Raum wird in [7] eingesetzt. Das Stereokamera paar besteht aus zwei Infrarotkameras, um auch in verrauchten Umgebungen klare Sicht zu ermöglichen. Die hier eingesetzte Wasserspritze wird mithilfe eines Modells für die Flugbahn des Löschmittels so eingestellt, dass sie das Ziel den Berechnungen zufolge treffen sollte. Verändert sich der Brand für eine bestimmte Zeit nicht mehr oder hat er seine Größe ausreichend verringert, wird der Löschmittelstrahl um den errechneten Zielpunkt herum bewegt. Sollte dies weiterhin nicht zum Löscherfolg führen, wird die Zielposition basierend auf einer Heuristik schrittweise angepasst, bis der Brand nicht mehr detektiert wird und als gelöscht betrachtet wird.

Zusammenfassend lässt sich feststellen, dass es verschiedene Ansätze zur gezielten Löschung von Bränden gibt, die jedoch stets für deutlich höhere Entfernungen vorgesehen sind, als sie beim Normbrand auftreten. Eine besondere Vorgehensweise oder Taktik zum Löschen findet sich nur in [7]. Diese ist jedoch ebenso auf die dort zu erwartenden Entfernungen angepasst und nicht für den Einsatz mit tragbaren Feuerlöschern geeignet. Außerdem ist allen Ansätzen gemein, dass sie auf eine kontinuierliche Versorgung mit einer beliebigen Menge an Wasser zurückgreifen können. Dies kann für tragbare Feuerlöscher nicht gelten.

2.2. Bewertung der Intensität von Bränden

Die zentrale Kenngröße für Intensität und Gefährdungspotenzial eines Brandes in der aktuellen Forschung ist die Wärmefreisetzungsrate (HRR) [8]. Diese ist jedoch nur mit einem erheblichen Messaufwand in sogenannten Kalorimetern messbar. Aufgrund der erforderlichen Messprinzipien eignen sich Kalorimeter allerdings nur bedingt für eine Online-Messung [9]. Daher wurde daran geforscht, die HRR empirisch mit den geometrischen Eigenschaften von Flammen, wie zum Beispiel ihrer Länge, Höhe oder Querschnittsfläche, in Zusammenhang zu bringen [10 bis 13]. So lässt sich die HRR auf vereinfachte Weise zumindest näherungsweise bestimmen. Der Zusammenhang der Länge von Flammen und der HRR, welcher in [10] beschrieben ist, bildet dabei eine zentrale Grundlage. Wobei zu beachten ist, dass die Schätzung sehr anfällig für Messrauschen ist, da sie auf einem Potenzgesetz basiert [14].

Im Bericht [15] wird festgestellt, dass sich sowohl die Oberflächentemperatur des Holzes als auch die HRR zur Bestimmung des Löscherfolgs eignen, die Oberflächentemperatur aber erheblich einfacher zu bestimmen und daher zu bevorzugen sei [15]. Dieses Kriterium wird in [15] allerdings nicht weiter spezifiziert. Das Ziel ist lediglich die Einschätzung des Löscherfolgs.

Ein Ansatz, Buschbrände anhand eines Multikamerasystems zu vermessen und geometrische Eigenschaften zu bestimmen, wird in [12, 13] vorgestellt. Hier wird die Korrelation jeweils zwischen Flammenhöhe, -fläche, -volumen und dem Wärmefluss berechnet. Es zeigt sich dabei, dass der Wärmefluss die größte Korrelation mit dem geschätzten Volumen aufweist. Für die Bestimmung des Volumens wird ein 3D-Modell des Brandes mithilfe eines Multikamerasystem erzeugt [11]. Dieser Ansatz erfordert einen entsprechend hohen Aufwand. Außerdem lassen sich diese Verfahren nur einsetzen, sobald Flammen vorhanden und sichtbar sind.

Insgesamt stellt sich also heraus, dass sich die Intensität eines Brandes zur Laufzeit von Brandversuchen insbesondere über die Geometrie der Flammen bestimmen lässt. Da diese Metrik aber nur mit Flammen funktioniert, reicht sie alleine nicht aus. Eine Bewertung darüber hinaus wird nur in Bezug auf den Löscherfolg betrachtet.

3. Löschvorgang für die Automatisierung der Überprüfung des Löschvermögens tragbarer Feuerlöscher

3.1 Löschen eines normgerechten Prüfobjekts von einem erfahrenen Löschmeister

Damit im Folgenden eine Vorgehensweise für das automatisierte Löschen im Rahmen der Prüfung des Löschvermögens nach DIN EN 3-7 [2] hergeleitet werden kann, wird hier zunächst der Löschvorgang, wie er von einem erfahrenen Menschen durchgeführt wird, analysiert. Ein erfahrener Mensch geht bei der Löschung des Prüfobjekts in zwei Phasen vor: In der ersten Phase werden ausschließlich die Flammen bekämpft. Die Verbrennungsreaktion findet in dieser Phase mit hoher Intensität statt und die Menge freigesetzter Wärme überschreitet die für die Fortsetzung der Reaktion erforderliche Aktivierungsenergie um ein Vielfaches. Ziel des Löschvorgangs in dieser Phase ist es, diese Kettenreaktion einzuschränken und zu unterbrechen. Feuerlöscher setzen dafür je nach Löschmittel auf unterschiedliche Prinzipien: Abkühlung, Unterbrechung der Sauerstoffzufuhr oder chemische Unterbindung der Verbrennungsreaktion [16]. Beim Löschen von Holzbränden der Brandklasse A mit tragbaren Feuerlöschern wird auf die ersten beiden dieser Prinzipien gesetzt. Das Löschmittel wird hierfür von der löschenden Person sequenziell auf die vier äußeren Seitenflächen des Prüfobjekts aufgetragen. Dabei ist darauf zu achten, dass möglichst das gesamte eingesetzte Löschmittel auf dem Prüfobjekt auftrifft. Mit diesem Vorgehen wird die abkühlende Wirkung des typischerweise wasserbasierten Löschmittels genutzt und mit der Verdrängung des Luftsauerstoffs durch den entstehenden Wasserdampf ergänzt. Diese Löschphase zeichnet sich auch durch eine erhöhte Dynamik aus: Solange das Prüfobjekt in Flammen steht, entzünden sich Glutnester durch die freigesetzte Wärme nach kurzer Zeit wieder. Werden keine Flammen mehr detektiert, nimmt diese Dynamik ab, die zweite Phase des Löschvorgangs beginnt. Zu diesem

Zeitpunkt verbleiben weiterhin Glutnester, die sich durch einfallende Strahlungswärme wieder entzünden können. Derartige Glutnester werden im Rahmen dieser Arbeit als kritische Glutnester bezeichnet. Ein menschlicher Löschmeister erkennt sie anhand ihrer dunkelrot leuchtenden Glut oder anhand von schnell ziehenden Rauchfahnen. In der zweiten Phase des Löschvorgangs werden auf diese Weise identifizierte Glutnester mit gezielten Sprühstößen gelöscht. Hierbei ist die primäre Löschwirkung das Abkühlen. Dies wird fortgesetzt, bis keine kritischen Glutnester mehr gefunden werden. Zu diesem Zeitpunkt wird der Brand als im Rahmen des Brandversuchs ausreichend gelöscht angesehen. Nun dürfen für den Erfolg der Prüfung innerhalb der nächsten drei Minuten keine Rückzündungen stattfinden. Dafür wird das Prüfobjekt in dieser Phase des Löschvorgangs weiterhin von der durchführenden Person überwacht, bis der Versuch beendet ist. Sollte eine Rückzündung auftreten, so wird diese gelöscht. Danach sind die drei Minuten erneut abzuwarten, bevor der Versuch als beendet gilt.

3.2 Anforderungen an die automatisierte Bekämpfung eines Brandes unter Normbedingungen

Das Löschen eines Brandes hat als zentrales Ziel den Schutz von Menschen und Material in der Umgebung. Der Normbrandversuch soll dazu dienen, sicherzustellen, dass tragbare Feuerlöscher diesen Schutz tatsächlich gewährleisten können. Als erfolgreich bekämpft im Rahmen der Überprüfung des Löschvermögens gilt ein Brand dann, wenn er erstens keine Flammen mehr ausbildet und zweitens innerhalb einer gewissen Zeitspanne (Die DIN EN 3-7 [2] schreibt hierfür 3 min vor.) nicht von Neuem aufflammt [2]. Messgrößen für die Steuerung des Löschvorgangs müssen entsprechend geeignet sein, zu erkennen, wenn die Flammen eines Brandes gelöscht sind. Außerdem ist zu identifizieren, wann davon auszugehen ist, dass keine Rückzündungen mehr stattfinden werden. Diese Messgrößen sollen weiterhin gewährleisten, dass beim Löschen zwischen den verschiedenen Phasen des Löschvorgangs zu günstigen Zeitpunkten ein Übergang stattfindet. Weiterhin ist es erforderlich, die Bestimmung dieser Größen während des Brandversuchs durchführen zu können, damit die Lage fortlaufend aktualisiert wird und dementsprechend Entscheidungen stets auf Basis von aktuellen Informationen getroffen werden können.

3.3 Auswahl geeigneter Merkmale zur Steuerung des Löschvorgangs

Das Vorgehen einer menschlichen Löschperson besteht für den Brandversuch aus zwei verschiedenen Löschphasen, in denen es jeweils verschiedene Löschziele gibt: Zuerst werden die Flammen unter dem kontinuierlichen Einsatz von Löschmittel bekämpft, bis alle Flammen gelöscht sind. Der Fortschritt beim Löschen der Flammen lässt sich mit der Länge der

Flammen beschreiben [10], welche im Falle des Normbrandes aufgrund der kontrollierten Umgebungsbedingungen annähernd der Höhe der Flammen entspricht. Sobald die Länge der Flammen eine Schwelle unterschreitet, kann zum Löschen der Glutnester übergegangen werden. Diese Schwelle sollte entsprechend der Vorgehensweise des menschlichen Löschmeisters bei null liegen. Ein Übergang zur nächsten Phase findet also erst dann statt, wenn keine Flammen mehr detektiert werden können.

Für das Ende des gesamten Löschvorgangs ist hingegen ein anderes Kriterium erforderlich, da beim Löschen der Glutnester keine Flammen vorhanden sind. Im Stand der Forschung wurde die Bewertung des Brandes anhand seiner Oberflächentemperatur vorgeschlagen [15]. Ob eine heiße Stelle sich innerhalb des Holzstapels tatsächlich wieder entzündet, hängt neben ihrer Temperatur besonders von der Menge einfallender Wärmestrahlung ab, die insbesondere von der Größe und Position des jeweiligen Glutnests abhängt. Da die Bewertung der Position anhand der jeweils einfallenden Wärmestrahlung jedoch als zu komplex eingeschätzt wird, sind kritische Glutnester demzufolge anhand ihrer Größe und Temperatur zu identifizieren. Um Rückzündungen zu vermeiden, sollten alle möglichst kritischen Glutnester detektiert und gelöscht werden. Sind keine dieser Glutnester mehr vorhanden, so ist davon auszugehen, dass sich der Brand nicht mehr innerhalb der vorgegebenen drei Minuten entzündet und der Lösversuch folglich beendet werden kann. Der Lösfortschritt in dieser Phase lässt sich entsprechend anhand der verbliebenen Zahl kritischer Glutnester bemessen.

3.4. Modell eines automatisierten Löschvorgangs

Mithilfe der Flammenhöhe und den kritischen Glutnestern kann die Löschung eines Normgerechten Prüfobjekts auch mit einem automatisierten System durchgeführt werden. Es wird dafür angenommen, dass ein automatisiertes System über die erforderlichen Fähigkeiten zur Durchführung des beschriebenen Ablaufs verfügt. Der Lösablauf, wie er von einem geeigneten automatisierten System durchgeführt werden kann, ist in Bild 1 in Form eines Zustandsautomaten abgebildet. Er basiert auf dem bereits beschriebenen Vorgehen eines Menschen, kombiniert mit den zuvor hergeleiteten Größen für die Steuerung der Übergänge zwischen den Phasen des Löschvorgangs.

Die erste Lösphase zur Bekämpfung der Flammen besteht aus insgesamt vier Zuständen, die jeweils die Bekämpfung der Flammen auf einer der Seitenflächen des Prüfobjekts darstellen. So soll sichergestellt werden, dass jede der Seitenfläche zumindest einmal mit Löschmittel besprüht wird. Dies ist aufgrund der begrenzten Eindringtiefe des Löschmittels in das Prüfobjekt erforderlich und ist analog zur Vorgehensweise des Menschen. Werden anschließend keine Flammen mehr detektiert, kann in die zweite Lösphase übergegangen werden. In

dieser Phase werden die vier Seitenflächen des Prüfbjektivs auf kritische Glutnester hin untersucht. Wird eines gefunden, so wird es mithilfe eines gezielten Sprühstoßes Löschmittel abgekühlt. Auch in dieser Phase gibt es insgesamt vier Zustände, je einen für die Inspektion jeder der vier Seitenflächen. Das Ziel ist es, den Löschvorgang erst zu beenden, sobald von keiner Seite aus mehr ein kritisches Glutnest zu detektieren ist. Die Gesamtzahl kritischer Glutnester beträgt nun also null. Abschließend wird in einem Wartezustand die Zeitspanne von drei Minuten abgewartet, innerhalb derer sich der Brand nicht erneut entzünden darf. Dafür wird jede der Seitenflächen für eine bestimmte Zeit betrachtet und nach kritischen Glutnestern oder Flammen abgesucht. Sollte eines der beiden entdeckt werden, geht das System wieder zur zugehörigen Löschtaktik über. Werden diese nicht gefunden, wird das Prüfbjektiv gedreht, sodass die nächste Seitenfläche in Richtung der Kameras zeigt. Der Versuch ist beendet, sobald die Zeit von drei Minuten ohne ein erneutes Löschen und ohne eine erneute Entzündung abgelaufen ist.

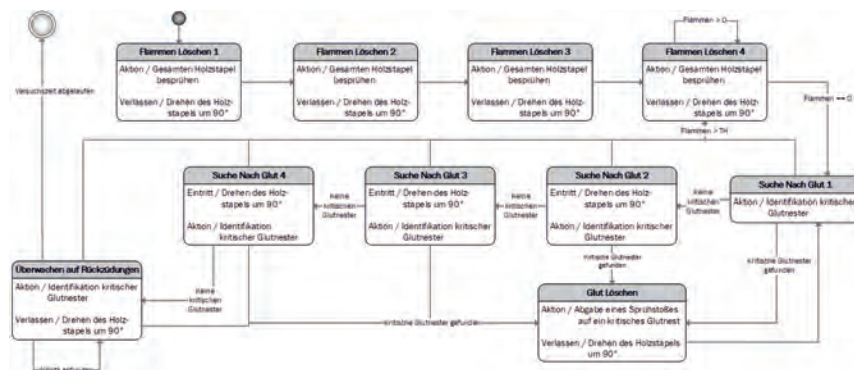


Bild 1: Ablauf des automatisierten Löschvorgangs basierend auf Flammenhöhe und den kritischen Glutnestern.

4. Validierung des Löschvorgangs anhand von Versuchsdaten

Mit dem Ziel zu untersuchen, ob die Durchführung des Löschvorgangs aus Bild 1 mit den beschriebenen Übergangsbedingungen durchführbar ist, sind mehrere Brandversuche durchgeführt worden. Bei diesen Versuchen ist ein normgerechtes Prüfbjektiv mit einer Farb- und einer Infrarotkamera beobachtet worden. Der Aufbau und die Ergebnisse der Versuche werden im Folgenden beschrieben.

4.1 Aufbau und Vorgehen zur Überprüfung der gewählten Methoden

Für die Brandversuche wurde jeweils ein Prüfobjekt 5A nach der DIN EN 3-7 [3] auf einem drehbaren Gestell aufgebaut. Dies ist in **Fehler! Verweisquelle konnte nicht gefunden werden.** gezeigt. In einer Entfernung von 2 m von diesem Aufbau wurden eine Farb- und eine Infrarotkamera auf das Prüfobjekt ausgerichtet, sodass sich jederzeit der gesamte Testbrand im Sichtfeld befand. Das Prüfobjekt wurde mit Heptan entzündet. Dafür wurde das Heptan in eine Wanne unterhalb des Holzstapels eingefüllt, welche vorher bereits zur Hälfte mit Wasser gefüllt worden war. Insgesamt wurde nach der Entzündung des Heptans ein Vorbrennzeit von acht Minuten abgewartet, wie sie auch in der DIN EN 3-7 [3] vorgesehen ist, bevor mit einem Löschvorgang begonnen wurde. Nach Ablauf dieser Vorbrennzeit wurde das Prüfobjekt entsprechend der Vorgehensweise in der ersten Phase des in Bild 1 gezeigten Löschablaufs durch einen Menschen von allen vier Seiten gelöscht, bis keine Flammen mehr vorhanden waren. Es wurde ausschließlich auf der den Kameras zugewandten Seite gelöscht. Nach jedem Löschen wurde das Prüfobjekt um 90° verdreht, sodass die nächste Seitenfläche den Kameras zugewandt war. Diese Prozedur wurde wiederholt, bis keine Flammen mehr zu erkennen waren. Sobald alle Flammen abgelöscht waren, wurde abgewartet, bis sich von neuem Flammen ausbildeten. Wurde eine solche Rückzündung festgestellt, so wurden die entstandenen Flammen erneut nach dem beschriebenen Ablauf gelöscht.

Mit diesem Vorgehen lässt sich der Zeitpunkt untersuchen, zu dem ein Wechsel zwischen den beiden Löschphasen geschehen muss, da dieser in jedem Versuch mehrfach stattfand. Außerdem lassen sich so die Glutnester betrachten, die sich während der Versuche wieder entzündeten. Es lassen sich also die hergeleiteten Größen zur Steuerung des Löschvorgangs gezielt untersuchen.



Bild 2: Aufbau des Prüfobjekts mit einer Vorrichtung zum Drehen während des Versuchs.

4.2 Untersuchung der Flammen

In Bild 3 wird die Höhe der Flammen im zeitlichen Verlauf während des Löschens dargestellt. Es sind in dieser Visualisierung zusätzlich (blau hinterlegt) die Zeitabschnitte, in denen

gelöscht wurde, und (grün hinterlegt) die Zeitabschnitte, in denen der Holzstapel gedreht wurde, markiert. Es zeigt sich in dieser Abbildung der erhebliche Einfluss des Löschens auf die Flammen. Ihre Höhe nimmt während der Auftragung des Löschmittels so stark ab, dass sie nur noch wenige Pixel beträgt. Die Höhe erholt sich jedoch anschließend innerhalb weniger Sekunden, während die nächste Seitenfläche des Brandes zur Kamera gedreht wird. Dies konnte bei allen durchgeführten Versuchen beobachtet werden. Ursächlich dafür ist zum einen, dass in den ersten vier Löschaktionen jeweils eine bisher noch nicht besprühte Seite vor die Kamera gedreht, sodass der Brand an diesen Seiten noch nicht bekämpft werden konnte und die Flammen demzufolge noch größer sind als an den bereits gelöschten Seiten. Zum anderen werden Flammen stark vom durch den Feuerlöscher verursachten Luftstrom beeinflusst, sodass die Flammen in den Zeitabschnitten, in denen gelöscht wird, ihre Größe stark verringern. Eine nachhaltige Löschwirkung entsteht hierdurch allerdings nicht. Eine entscheidende Beobachtung ist hingegen, dass gerade zu Beginn zu erkennen ist, dass die maximale Flammenhöhe mit jedem Löschen geringer wird.

Essenziell für die Bestimmung des Übergangszeitpunktes sowie die frühzeitige Detektion von Rückzündungen ist Fähigkeit der Flammendetektionsalgorithmen, auch sehr kleine Flammen zuverlässig zu detektieren, wie sie beispielhaft in **Fehler! Verweisquelle konnte nicht gefunden werden.**(a) im rot markierten Bereich des Prüfobjekts zu sehen ist. Hierfür wurde der Algorithmus aus [17] verwendet, da er genau dies durch die Verwendung von Bewegung und Farbe als typischer Merkmale von Flammen ermöglicht. Es lässt sich über alle Versuche hinweg beobachten, dass der eingesetzte Algorithmus kleine Flammen mit hoher Sensitivität erkennt. Jedoch sind auch einige Fehldetektionen festzustellen, die besonders dann auftreten, wenn Menschen im Bild sind. Insgesamt lässt sich jedoch festhalten, dass sich durch das Fehlen von Flammen der Übergangszeitpunkt vom Löschen der Flammen zum Löschen der kritischen Glutnester sowie das Auftreten von Rückzündungen bestimmen lassen.

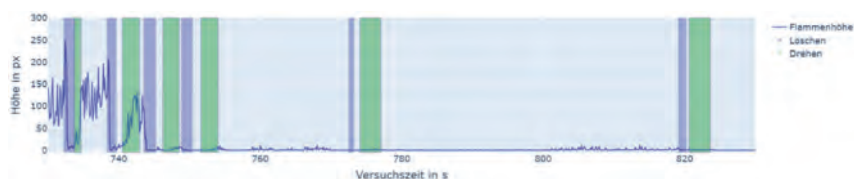


Bild 3: Ausschnitt des Verlaufs der gemessenen Flammenhöhe im Brandversuch.

4.3 Untersuchung der Infrarotaufnahmen

Um zu ermitteln, wie kritische Glutnester im Rahmen eines Normbrandversuchs zu definieren sind, wird, nachdem die Flammen erfolgreich bekämpft sind, der Holzstapel von einer Seite

beobachtet, bis eine Rückzündung stattfindet. Anhand aller stattfindenden Rückzündungen in den Versuchen wurden die Eigenschaften der sich jeweils entzündenden kritischen Glutnester bestimmt. Hierfür wurden die Größe und die Temperatur jedes kritischen Glutnests bestimmt und aufgezeichnet. Aufgrund der Perspektive und des Aufbaus des Prüfbobjekts sind jedoch nicht alle Rückzündungen im Sichtbereich der Kamera geschehen. Dies ist auch in **Fehler! Verweisquelle konnte nicht gefunden werden.**(b) zu sehen. Dort sind die gelben Flächen als Glutnester klar zu erkennen. Sie werden jedoch zum Teil auch davorliegenden Holzstäben des Prüfbobjekts verdeckt. Entsprechend ist es erforderlich, dass das automatisierte Löschsystem beim Warten auf das Ende des Löschversuchs nicht nur eine Seite des Prüfbobjekts beobachtet. Deshalb ist es Vorgesehen das Prüfbobjekt in Wartezustand regelmäßig zu drehen.

Tabelle 1: Eigenschaften einzelner kritischer Glutnester, die während der Versuche Rückzündungen ausbildeten.

	Fläche eines Glutnestes in cm ²	Temperatur eines Glutnestes in K
Minimum	8,13	573,4
Maximum	38,53	713,6
Mittelwert	20,86	655,8

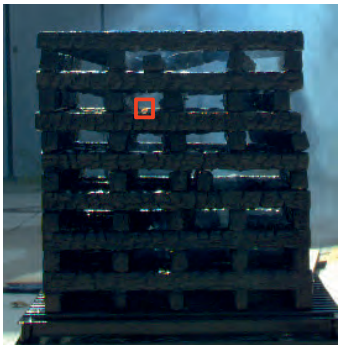


Bild 4(a): Farbbild des Testbrands mit kleiner Flamme.

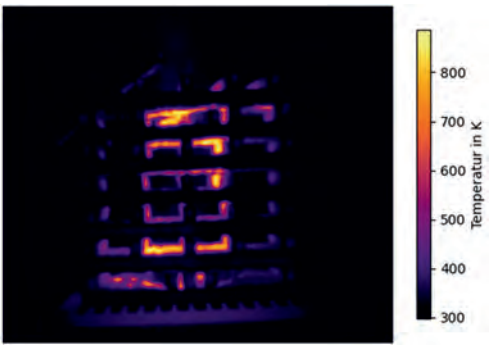


Bild 4(b): Infrarotaufnahme des Testbrandes nach Ablöschen der Flammen.

Insgesamt hat sich in den durchgeführten Brandversuchen die Verteilung der Rückzündungstemperaturen, wie sie in **Fehler! Verweisquelle konnte nicht gefunden werden.** dargestellt ist, ergeben. Es zeigt sich, dass die per Infrarotkamera gemessenen Temperaturen der Glutnester, von denen eine Rückzündung ausgeht, mindestens 573,4 K betragen. Die

Untersuchung der Größe der kritischen Glutnester zeigt erhebliche Unterschiede in der Verteilung zwischen den verschiedenen Versuchen. Die Größen der Glutnester werden anhand der Bilder der Infrarotkamera bestimmt. Die Menge detektierter Pixel wird über die bekannte Geometrie des Aufbaus und der Kamera in cm^2 konvertiert. Insgesamt liegt die Größe der Glutnester, die nach dem Ablöschen wieder eine Flamme ausbilden, im Bereich $8,13 \text{ cm}^2$ und $38,53 \text{ cm}^2$. Somit lässt sich ein kritisches Glutnest als ein solches definieren, dessen Temperatur oberhalb von $573,4 \text{ K}$ liegt oder das eine Größe von mehr $8,13 \text{ cm}^2$ aufweist. Auch wenn diese Werte rein empirisch gewonnen sind, lassen sich durch die kontinuierliche Betrachtung des Prüfobjekts sowie die Betrachtung von allen vier Seiten kritische Glutnester rechtzeitig erkennen, sofern sie sich im Sichtbereich der Kamera befinden.

5. Fazit

Für die Steuerung eines automatisierten Systems zur Durchführung der Überprüfung des Löschvermögens tragbarer Feuerlöscher nach DIN EN 3-7 [2] ist ein Löschvorgang in Form eines Zustandsautomaten hergeleitet worden. Dafür sind zwei essenzielle Merkmale bestimmt und im Versuch untersucht worden, anhand derer Zustandsübergänge ausgelöst werden sollten. Die Höhe der Flammen kann als Indikator für die Intensität der ablaufenden Verbrennungsreaktion genutzt werden. Sobald die Flamme verschwindet, ist in den Löschmodus für Glutnester überzugehen. Für diese zweite Phase werden kritische Glutnester als Glutnester definiert, die die Gefahr bergen, innerhalb der Versuchszeit eine Rückzündung auszubilden. Dafür werden sie anhand ihrer Größe und Temperatur charakterisiert. Grenzwerte für die Größe bzw. Temperatur der kritischen Glutnester wurden auf Basis der Brandversuche empirisch bestimmt. Da das Prüfobjekt jedoch bis zum Ende des Versuchs kontinuierlich weiter beobachtet wird, können Rückzündungen und kritische Glutnester auch dann registriert werden, wenn diese die hier bestimmten Kriterien nicht unmittelbar erfüllen sollten. Eine der größten Herausforderungen sind die Sichtverhältnisse. Bereits durch den Aufbau des Prüfobjekts kann es dazu kommen, dass kleine Flammen oder kritische Glutnester für ein Kamerasystem nicht sichtbar sind.

Auf Basis des hier hergeleiteten Löschvorgangs soll nun ein vollständiges System aufgebaut werden, welches diesen Vorgang automatisch durchführen kann.

Literaturangaben

- [1] Technische Regeln für Arbeitsstätten ASR A2.2:2018-05. *Maßnahmen gegen Brände.*
- [2] Norm 3-7. Tragbare Feuerlöscher. Teil 7: Eigenschaften, Leistungsanforderung und Prüfungen
- [3] DIN EN 2. *Brandklassen*
- [4] Stoller, F., Kümmerlen, F. u. Fay, A.: Vergleichende Bewertung von Algorithmen zur Lokalisierung von Flammen und Glut. *tm - Technisches Messen* 88 (2021) 6, S. 408–419
- [5] Wu, H. B., Li, Z. J., Ye, J. H., Ma, S. C., Li, J. W. u. Yang, X. N.: Firefighting robot with video full-closed loop control. *International Journal of Safety and Security Engineering* 6 (2016) 2, 254–269
- [6] Chen, T., Yuan, H., Su, G. u. Fan, W.: An automatic fire searching and suppression system for large spaces. *Fire Safety Journal* 39 (2004) 4, S. 297–307
- [7] McNeil, J. G. u. Lattimer, B. Y.: Robotic Fire Suppression Through Autonomous Feedback Control. *Fire Technology* 53 (2017) 3, 1171–1199
- [8] Babrauskas, V. u. Peacock, R. D.: Heat Release Rate: The Single Most Important Variable in Fire Hazard. *Fire Safety Journal* 18 (1992) 3, S. 255–272
- [9] Hurley, M. J., Gottuk, D. T., Hall Jr., J. R., Harada, K., Kuligowski, E. D., Puchovsky, M., Torero, J. L., Watts Jr., J. M. u. WIECZOREK, C. J. (Hrsg.): *SFPE Handbook of Fire Protection Engineering*. New York: Springer 2016
- [10] Heskestad, G.: Fire Plumes, Flame Height, and Air Entrainment. In: Hurley, M. J., Gottuk, D. T., Hall Jr., J. R., Harada, K., Kuligowski, E. D., Puchovsky, M., Torero, J. L., Watts Jr., J. M. u. WIECZOREK, C. J. (Hrsg.): *SFPE Handbook of Fire Protection Engineering*. New York: Springer 2016, S. 396–428
- [11] Morandini, F., Toulouse, T., Silvani, X., Pieri, A. u. Rossi, L.: Image-Based Diagnostic System for the Measurement of Flame Properties and Radiation. *Fire Technology* 55 (2019) 6, S. 2443–2463
- [12] Rossi, L., Toulouse, T., Akhloufi, M., Pieri, A. u. Tison, Y.: Estimation of spreading fire geometrical characteristics using near infrared stereovision. *Three-Dimensional Image Processing (3DIP) and Applications 2013. SPIE Proceedings. SPIE 2013*, 86500A
- [13] Toulouse, T., Rossi, L., Akhloufi, M. A., Pieri, A. u. Maldague, X.: A multimodal 3d framework for fire characteristics estimation. *Measurement Science and Technology* 29 (2018) 2, S. 25404
- [14] Li, K., Mao, S. u. Feng, R.: Estimation of Heat Release Rate and Fuel Type of Circular Pool Fires Using Inverse Modelling Based on Image Recognition Technique. *Fire Technology* 55 (2019) 3, S. 667–687

- [15] Entwicklung von Grundlagen für ingenieurtechnische Methoden zur Berechnung der erforderlichen Löschintensitäten für das Löschen von Bränden mit Wasserdurch die Feuerwehr. Brandschutzforschung der Bundesländer Nr. 186, Kohl, K.-J., Jahn, E., Kutz, M. u. Wienecke, F., Heyrothsberge 2016
- [16] Klingsohr, K.: Verbrennen und Löschen. W. Kohlhammer GmbH 2002
- [17] Krooß, J., Kümmerlen, F. u. Fay, A.: Schnelle und präzise Segmentierung von beweglichen und morphologisch variablen Objekten am Beispiel der Deflagrationsdetektion. Forum Bildverarbeitung 2018. Karlsruhe: KIT Scientific Publishing 2018, 241-252

Verwendung Neuronaler Netze zur automatisierten Analyse und Modellierung elasto-plastischer Materialeigenschaften von metallischen Zugproben

Henrik Sebastian Steude, Daniel Vranješ, Jan Lukas Augustin,
Prof. Dr. **Oliver Niggemann,**
Helmut-Schmidt-Universität, Hamburg;
Prof. Dr. **Markus Lange-Hegermann,**
Technische Hochschule Ostwestfalen-Lippe;
Dr. **Daniel Höche,**
Helmholtz-Zentrum Geesthacht Hereon, Geesthacht

Kurzfassung

Die Zugprüfung ist eine etablierte und standardisierte Methode für die werkstofftechnische Überprüfung von Konstruktionsmaterialien. Die Auswertung der Materialkennwerte aus Spannungs-Dehnungs-Kurven ist ein zum Teil aufwändiger und oftmals manueller, jedoch elementarer Prozess, da so wichtige materialspezifische Informationen abgeleitet werden können, die für das Entwerfen mechanischer Komponenten benötigt werden. In dieser Arbeit stellen wir zwei Ansätze vor, die zeigen, wie Neuronale Netze (NN) eingesetzt werden können, um Spannungs-Dehnungs-Kurven automatisiert auszuwerten. Zum einen nutzen wir überwachtes Lernen, um Zugproben hinsichtlich eines geeigneten empirischen Verfestigungsmodells zu klassifizieren und die entsprechenden Modellparameter zu extrahieren. Zum anderen zeigen wir, wie unüberwachtes Lernen genutzt werden kann, um physikalische Konzepte in latenten Variablen zu enkodieren. Die Trainingsdaten wurden mittels numerischer FEM-Simulationen erzeugt. Das langfristige Ziel ist hierbei die Ausgabe von Materialparametern wie Elastizitätsmodul, Streckgrenze und Zugfestigkeit sowie zunächst qualitativer Informationen über das Vorhandensein von (kinematischen) Verfestigungsmechanismen im Bereich plastischer Verformung.

Abstract

Tensile testing is an established and standardized method for material testing. The evaluation of mechanical materials properties from stress-strain curves is a sometimes laborious and very often manual but elementary process, as it can be used to derive materials-specific information needed for the design of mechanical components. In this work, we present two approaches to show how neural networks (NNs) can be used to evaluate stress-strain curves in an automated manner. First, we use supervised learning to classify stress-strain curves with respect to an

appropriate empirical strain hardening model and extract the corresponding model parameters. Second, we show how unsupervised learning can be used to encode physical concepts in latent variables. The long-term goal is to output material parameters such as elastic modulus, yield strength and ultimate tensile strength, as well as qualitative information on the presence of (kinematic) hardening mechanisms during plastic deformation. Respective training data was generated by numerical FEM simulations.

1. Einführung

1.1 Chancen

Die Integrität jeder Konstruktion basiert auf ihren Materialien und deren Eigenschaften. Durch neue Fertigungsverfahren und insbesondere auch durch neuartige metallische Legierungen und Legierungsentwicklung ist auch heute noch die Bestimmung von Materialparametern hin zu neuen Konstruktionskonzepten eine wichtige Aufgabe. Sowohl für neuartige als auch für etablierte Materialien führen Prüflabore standardisierte Materialtests durch, um deren Eigenschaften zu bestimmen. Für metallische Werkstoffe sind Diese in den jeweiligen Normen wie zum Beispiel DIN [EN](#) ISO 6892-1, ISO 6892, [ASTM](#) E 8, ASTM E 21, DIN 50145 festgehalten. Einer der Grundpfeiler der mechanischen Prüfung ist der Zugversuch, bei dem die Spannungs-Dehnungs-Beziehungen in verschiedenen genau spezifizierten Testkonfigurationen (z.B. mit konstanter Dehnrate) gemessen werden. Die Auswertung des Versuchs erfolgt hierbei heutzutage noch primär manuell, oder teilweise mit Hilfe von teilautomatisierter Auswertungssoftware. Unter die zu bestimmenden Größen fallen die klassischen Materialparameter wie E-Modul, Zugfestigkeit und Streckgrenze. Aber auch komplexere Systemdeskriptoren, welche über Parameter zum Beispiel plastische Mechanismen auf der Mikroskala abbilden, und so mittels (kinematischer) Verfestigungsmodelle (bspw. nach Ludwik oder Swift) Verformungskenngrößen enthalten, können aus den Daten extrahiert werden. Die Ermittlung des am besten geeigneten Verfestigungsmodells erweist sich hierbei in der Praxis als eine nicht triviale Aufgabe, da jedes der Modelle nur eine vereinfachte Beschreibung der tatsächlichen atomaren, mikroskopischen als auch mesoskopischen Prozesse darstellt und selbst bei unterschiedlichen Modellen teilweise sehr ähnliche Kurvenverläufe der Spannungs-Dehnungskurven entstehen können. Fortschritte im Bereich des Maschinellen Lernens (ML) motivieren dazu, Deep Learning Ansätze zu verwenden, um datenbasierte Algorithmen zu entwerfen, welche gezielt Eigenschaften und Modelle für neue aber auch etablierte Materialien bestimmen können. Unser Beitrag liefert noch keine vollständige Lösung für die benannten Punkte, stellt jedoch einen ersten Schritt in Richtung einer automatisierten und für Menschen interpretierbaren Auswer-

tung von Zugversuchen dar und eröffnet verschiedene Vertiefungsrichtungen weiterer Forschung. In dieser Arbeit wird ein Ansatz zur automatisierten Analyse von Spannungs-Dehnungs-Kurven von metallischen Rundzugproben mit elasto-plastischem Verformungsverhalten vorgestellt. Das Ziel dieser automatisierten Auswertung ist die Identifikation des geeignetsten Verfestigungsmodells sowie der dazugehörigen Modellparameter zur formalen Beschreibung des metallischen Materialdeformations-verhaltens bei Dehnung mit konstanter Dehnrate. Unterstützt werden soll dieses Vorhaben durch KI-basierte Modelle, die es ermöglichen physikalische Konzepte aus Daten zu erlernen und repräsentativ darzustellen. Das Erreichen solcher Vorhersagefähigkeiten ermöglicht es, Materialwissenschaftlern, (Prüf-)Ingenieuren oder auch Konstrukteuren schneller Parameter für neue zum Teil auch adaptierbare Materialien zu bestimmen und in den Entwicklungs- und Konstruktionsprozess im Sinne „Materials as a Service“ zu integrieren.

1.2 Herausforderungen

Um ein interpretierbares ML Modell zu erhalten, welches auf leicht verständlichen Gleichungen, physikalischen Größen und Formeln basiert, sollten Modellformalismen verwendet werden, auf die auch ein Mensch zurückgreifen würde. Darüber hinaus sollte das Modell eine zuverlässigere Extrapolation z.B. der Verformungseigenschaften oder zum Beispiel auch der Schadensausbreitung ermöglichen. Dies erfordert die erfolgreiche Extraktion physikalischer Gesetze aus Daten (sowohl gemessen als auch mittels physikalischer Simulationen generiert). Um diese Herausforderung zu meistern, werden zunächst folgende zwei Forschungsfragen behandelt:

- a) Identifikation eines geeigneten physikalischen Modells und der dazugehörigen Materialeigenschaften durch überwachtes Lernen:

Auf der Grundlage experimenteller Spannungs-Dehnungs-Kurven soll das richtige Verfestigungsmodell aus einer Menge empirischer Modelle identifiziert werden. Ein Klassifikator wird auf der Grundlage früherer physikalischer Simulationsdaten und dem zugrundeliegenden Verfestigungsmodell als Label angelernet. Nach der Klassifizierung werden die entsprechenden Parameter des physikalischen Modells automatisiert gelernt.

- b) Unüberwachtes Lernen von bedeutungsvollen latenten Variablen aus Zugproben:
Auf Grundlage experimenteller Spannungs-Dehnungs-Kurven sollen mittels generativer Modelle relevante physikalische Konzepte aus den Daten abstrahiert bzw. extrahiert werden, ohne dabei jegliche physikalische Information oder Gleichungen als Vorwissen zu nutzen.

1.3 Maschinelles Lernen und Materialwissenschaft

ML-Methoden bieten ein großes Potenzial zum Erlernen einer Vielzahl von Aufgaben aus großen Datensätzen. Ein Nachteil von ML-Modellen, die ausschließlich auf Daten trainiert werden, ist ihre mangelnde Interpretierbarkeit. NNs werden oft als "Black Box" bezeichnet. Ihre Unfähigkeit, zugrundeliegende physikalischen Gesetze explizit zu erlernen, führt zu einem geringen Generalisierungspotenzial. In den letzten Jahren haben die Bereiche der erklärbaren künstlichen Intelligenz und des Lernens von physikalischen Konzepten oder der maschinengestützten wissenschaftlichen Entdeckung durch NNs stark zugenommen. Sowohl das Lernen von physikalischen Konzepten oder Formeln mit NNs als auch die Anwendung von NN-basierten Algorithmen im Bereich der Materialwissenschaften [1] wurden in mehreren Studien unter verschiedenen Aspekten untersucht, wie die folgenden Abschnitte zeigen.

Die Autoren von [2] trainierten beispielsweise ein NN für das Erlernen von Greenschen Funktionen versteckter partieller Differentialgleichungen für ein interpretierbares Deep Learning Modell auf der Grundlage von Eingabe-Ausgabe-Beispielen eines bekannten Gauß-Prozesses. Zwei Netzwerke wurden auf den generierten Daten trainiert, um die homogene Lösung und die Greensche Funktion zu lernen. Die Verständlichkeit für den Menschen wurde durch Sampling der trainierten NNs erzeugt. In [3] wurden NNs verwendet, um die Hamiltonsche Mechanik eines Masse-Feder-Systems zu lernen. Die Autoren von [4] nutzten die SINDy-Methode (Sparse Identification of Non-linear Dynamics), um ein Modell des Gesetzes der Gravitation aus Falltestdaten mehrerer Objekte lernen zu lassen. Mit dieser Methode gingen die Autoren auf die Notwendigkeit ein, das Training so zu steuern, dass es sich an die Diskrepanz zwischen verrauschten Messdaten und dem gesuchten zugrundeliegenden physikalischen Gesetz anpasst. Die Paper [5] und [6] bauen explizit Vorwissen, gegeben durch Differentialgleichungen, in Gaußprozessmodelle ein. In [7] wurden sparsame NNs genutzt, um die Schmelztemperatur von Materialien anhand grundlegender atomarer und kristalliner Eigenschaften ohne Verwendung von Vorwissen vorherzusagen. Die Autoren von [8] verwendeten latente neuronale gewöhnliche Differentialgleichungen in Kombination mit rekurrenten und MLP-Schichten, um die Quantendynamik eines zweistufigen Quantensystems zu erlernen. Die Autoren zeigten, dass es möglich ist, die erlernte Quantenmechanik unter Beibehaltung der

physikalischen Gesetze zu extrapolieren und die Heisenbergsche Unschärferelation eines Qubits wiederherzustellen, ohne dass Vorkenntnisse erforderlich sind. Die Ergebnisse zeigen, dass NNs in der Lage sind, physikalische Modelle zu erlernen, indem sie z.B. die problembeschreibenden Differentialgleichungen (zum Beispiel generiert durch mathematische Modellierung) lösen bzw. parametrisieren oder indem sie die Entwicklung eines einzelnen, zum Teil auch sehr komplexen physikalischen Systems unter äußeren Einflüssen im Laufe der Zeit modellieren (durch den Einsatz rekurrenter Netze). Durch die Verwendung von Domänenwissen konnten die benannten Autoren die Modelleistung im Vergleich zu einer Basismethode und die Interpretierbarkeit der Ergebnisse in vielen Fällen verbessern. Mit den Studien [7] und [8] existieren bereits auch Ansätze, die sich nicht explizit auf Vorwissen für Modellarchitektur und Trainingsmethode stützen.

Ein prominenter Ansatz für die menschenverständliche Gestaltung von NNs liegt in der Verwendung von beobachtbaren latenten Schichten innerhalb des Netzes, von denen eine Stichprobe genommen werden kann.

Die vorliegende Forschungsarbeit zeigt Methoden zur Erzeugung von für den Menschen verständlichen Deep-Learning-Modellen auf, indem NN-Architekturen entworfen werden, die das spezifische Lernen des gewählten Ziels ermöglichen und Repräsentationen der gewünschten physikalischen Aspekte innerhalb eines Teils des NN aufbauen, die dann direkt und durch Stichproben (Sampling) beobachtet werden können.

Wie Schmidt et al. [9] zeigen, haben Algorithmen des maschinellen Lernens verschiedene Anwendungsbereiche im Bereich der Materialwissenschaften gefunden, darunter auch in die Materialentwicklung, z. B. hinsichtlich der Stabilität eines Materials auf der Grundlage seiner Kristallstruktur und chemischen Zusammensetzung [10], oder auch für die Vorhersage von atomaren Materialeigenschaften, z. B. die Vorhersage und Erklärung von Molekularenergien [11].

Im Bereich des Bauingenieurwesens wurden NNs zur Vorhersage der Spannungs-Dehnungs-Beziehung von Stahlbetonprofilen auf der Grundlage der Eingangsdehnung verwendet. Die Experimente zeigen, dass das Modell in der Lage ist, die Spannungs-Dehnungs-Relation ausreichend genau abzubilden. Die Modellarchitektur umfasst jedoch keine für den Menschen verständlichen Komponenten und versucht auch nicht, ein übergeordnetes interpretierbares physikalisches Konzept zu erlernen.

Eine wichtige mechanische Eigenschaft eines Konstruktionsmaterials ist seine Zugfestigkeit. Beispielsweise in [12] wurden Bayes'sche NNs zur Vorhersage der Zugfestigkeit von Metalllegierungen aus der chemischen Zusammensetzung von Stahl untersucht. Darüber hinaus wurden in [13] Bayes'sche NNs zur Vorhersage der Zugfestigkeit von Metalllegierungen aus der

chemischen Zusammensetzung von Stahl und Walzprozessvariablen oder auch entsprechenden Mikrostrukturdaten eingesetzt. In [14] und [15] wurde ein Feed-Forward-Netzwerk zur Vorhersage des Elastizitätsmoduls, der Streckgrenze, der Zugfestigkeit und der Bruchdehnung verschiedener Legierungen auf der Grundlage der chemischen Zusammensetzung, der Wärmebehandlung und der Konditionierung sowie der plastischen Verformungsgeschichte der Legierungen verwendet. Die Modelle wurden anhand von numerischen und kategorischen Daten aus verwandten Legierungen trainiert. Die Modelle zeigen, dass sie in der Lage sind, charakteristische Eigenschaften bestimmter Legierungen zu erlernen und Vorhersagen über die Eigenschaften ähnlicher Legierungen bzw. Legierungsklassen zu machen, aber es fehlt ihnen nach wie vor an physikalischer Interpretierbarkeit und sie sind nicht mit tatsächlichen Zugprüfungsdaten und der direkten Extraktion von Materialeigenschaften verknüpft.

Der Stand der Technik zeigt verschiedene Forschungsarbeiten zur Anwendbarkeit von NNs im Bereich der Naturwissenschaften hinsichtlich des Lernens physikalischer Konzepte und der Anwendbarkeit von NNs für verschiedene materialtechnische Aufgaben. Obwohl die Forschung zeigt, dass NN-basierte Algorithmen in der Lage sind, physikalische Gesetze direkt aus Daten zu lernen, zeigt der Stand der Technik, dass bisher keine vergleichbaren Studien über das Lernen von Materialkonzepten oder Modellen auf der direkten Grundlage von Spannungs-Dehnungs-Daten durchgeführt wurden. Der aktuelle Stand der Forschung auf dem Gebiet der Werkstofftechnik zeigt zwar erlernbare Beziehungen zwischen ähnlichen Legierungen, aber es mangelt sehr häufig an der physikalischen Interpretierbarkeit dieser Beziehungen. Es gibt zudem keine bekannten Algorithmen für die vollautomatische Analyse von Werkstoffeigenschaften, die ausschließlich auf mechanischen Zugversuchen basieren.

Wie neuere Forschungen zeigen, kann die Identifizierung von Materialparametern durch genetische Algorithmen [16] unterstützt werden, erfordert aber immer noch Expertenwissen.

In der Praxis wird der wichtige Auswertungsprozess der Anpassung eines funktionierenden empirischen Verfestigungsmodells wie Swift, Ludwik oder Voce [17] an ein Material auf der Grundlage von Zugversuchen immer noch meist durch manuelle Anpassung der gemessenen Kurve und der Modelle durchgeführt, gefolgt von einer Bewertung des Rekonstruktionsfehlers durch einen Fachmann. Mit dieser Forschung versuchen wir, diese Lücke zu schließen, indem wir interpretierbare neuronale Netze mittels Daten aus Zugversuchen trainieren und physikalische Materialeigenschaften in die NNs kodieren.

2. Identifikation eines geeigneten physikalischen Modells und der dazugehörigen Materialeigenschaften durch überwachtes Lernen

2.1 Methode

Für unsere Untersuchung notwendige Daten wurden mittels simulierter Zugversuche in der FEM basierten Software Comsol 5.6 erstellt. Basis für die Simulation ist ein vereinfachtes Modell einer rotationssymmetrischen Zugprobe (vgl. Abbildung 1), welche unter Berücksichtigung von Symmetrieeigenschaften auf eine zweidimensionale Struktur reduziert wurde. Simulationsergebnisse beruhen auf der lokalen Auswertung der realen Spannung und Dehnung am im Zentrum der Zugprobe.



Bild 1: Qualitative Darstellung der vereinfachten Zugprobengeometrie (obere Hälfte) mit verwendeter Vernetzung(links) und Belastungen sowie Verformung (rechts)

Für die Simulation der physikalischen Prozesse der Verfestigung, welche durch das plastische Umformen der hier betrachteten Metallbeispiele entstehen, wurden die Modelle nach Swift und Ludwik verwendet, welche wie folgt beschrieben sind. In dem Ludwik Modell für nichtlineare isotrope Verfestigung wird die Fließspannung σ_{ys} durch eine nichtlineare Funktion der effektiven plastischen Verformung definiert (vgl. (1)), mit anfänglicher Fließspannung σ_{ys0} , Verfestigungskoeffizient k_{lud} , Verfestigungsexponent n und plastischer Dehnung ϵ_{pe} .

$$\sigma_{ys}(\epsilon_{pe}) = \sigma_{ys0} + k_{lud} \epsilon_{pe}^n \quad (1)$$

Für nicht-sättigende Materialien verbindet das Swift Modell die anfängliche Fließspannung und isotrope Härtung mit der effektiven plastischen Verformung (vgl. (2)), mit anfänglicher Fließspannung σ_{ys0} , Verfestigungsexponent n , Referenzdehnung ϵ_{ref} und tatsächlicher plastischer Dehnung ϵ_{pe} .

$$\sigma_{ys}(\epsilon_{pe}) = \sigma_{ys0} \left(1 + \frac{\epsilon_{pe}}{\epsilon_{ref}} \right)^n \quad (2)$$

Um einen heterogenen Datensatz zu erzeugen, der sich für Klassifikations- und Regressionsaufgaben eignet wurden die Modellparameter auf Basis realistischer experimenteller Messdaten wie in Tabelle 1 dargestellt abgetastet.

Tabelle 1: Abtastparameter für simulative Generierung von Zugprobenverläufen.

Modell	Parameter	Startwert	Schrittweite	Endwert	Einheit
Ludwik	E	$1,5 \cdot 10^{10}$	$0,5 \cdot 10^{10}$	$4 \cdot 10^{10}$	Pa
Ludwik	σ_{ys0}	$60 \cdot 10^6$	$5 \cdot 10^6$	$80 \cdot 10^6$	Pa
Ludwik	k_{lud}	$100 \cdot 10^6$	$3 \cdot 10^6$	$121 \cdot 10^6$	Pa
Ludwik	n	0,2	0,1	1	-
Swift	E	$1,5 \cdot 10^{10}$	$0,5 \cdot 10^{10}$	$4 \cdot 10^{10}$	Pa
Swift	σ_{ys0}	$60 \cdot 10^6$	$5 \cdot 10^6$	$80 \cdot 10^6$	Pa
Swift	ϵ_{ref}	0,003	0,003	0,024	-
Swift	n	0,2	0,1	1	-

Die Simulationen hatten eine simulierte Zeit von 45045 Sekunden und eine Formänderungsrate von $2,22 \cdot 10^{-6} \frac{1}{s}$. Die Simulationen wurden dabei mit und ohne kinematischem Verfestigungsmodell durchgeführt. Im Falle kinematischer Verfestigung [18] wurde zunächst ein lineares Verfestigungsmodell (Prager [19]) mit $E_k = 50 \text{ Pa}$ verwendet. Der finale Datensatz besteht somit aus 4 Klassen von Kurven (Ludwik mit und ohne kinematische Verfestigung sowie Swift mit und ohne kinematische Verfestigung) mit jeweils 2160 Zugproben pro Klasse. Die Spannungs-Dehnungs-Kurven der simulierten Zugversuche werden als dreidimensionale Sequenzen modelliert, in denen neben der Spannung und Dehnung auch die Zeitpunkte der theoretischen Messungen (entsprechend den Zeitschritten) enthalten sind. Jede Zeitreihe verfügt über Metainformationen (Label) bezüglich deren Klasse und Abtastparameter. Mittels dieser Label werden zwei separate NNs trainiert, wobei das erste NN die Zugproben hinsichtlich der passenden physikalischen Modelle klassifiziert und das zweite NN die physikalischen Modellparameter schätzt.

Beide Architekturen bestehen aus Multi-Layer-Perzeptrons (MLPs). Das NN für die Klassifikationsaufgabe verfügt über vier Ausgabeneuronen, die jeweils eine der Klassen repräsentieren. Als Verlustfunktion wird die Kreuzentropie verwendet. Das NN für die Regressionsaufgabe verfügt über drei Ausgabeneuronen, von denen jeweils ein Neuron für einen Parameter der physikalischen Modelle steht. Als Verlustfunktion dient in diesem Fall die mittlere quadratische Abweichung.

2.2 Ergebnisse

Zur Evaluierung der Güte der trainierten Modelle verwenden wir ausschließlich Daten von Zugproben, die weder zum Modelltraining noch im Rahmen des Hyperparametertunings betrachtet wurden (Test-Datensatz).

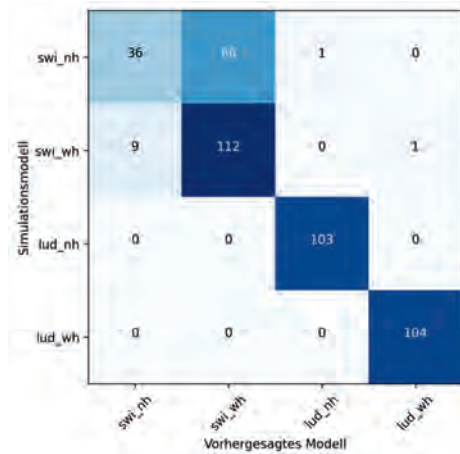


Bild 1: Confusion-Matrix der Klassifizierung der Zugproben im Test-Datensatz

Bild 1 stellt in einer Tabelle die von dem Klassifikationsmodell vorhergesagten physikalischen Modelle denen gegenüber, die tatsächlich zur Simulation der Daten verwendet wurden. Die Abbildung zeigt, dass es dem Modell nahezu perfekt gelingt, die „Ludwik-Zugproben“ von den „Swift-Zugproben“ zu unterscheiden. Darüber hinaus ist das Modell in der Lage, zwischen „Ludwik-Zugproben“ mit und ohne kinematische Verfestigung zu differenzieren. Die Confusion-Matrix zeigt allerdings auch deutlich, dass das Modell es noch nicht vermag, die „Swift-Zugproben“ hinsichtlich der kinematischen Verfestigung voneinander zu unterscheiden. Eine mögliche Erklärung hierfür ist, dass die kinematische Verfestigung im Swift Modell nur einen sehr geringen Einfluss auf das Zugprobenverhalten im Trainings-Datensatz hat, und diese damit nur sehr schlecht von dem NN zu unterscheiden sind. Diese Vermutung festigt sich bei näherer Betrachtung der Daten.

Zur Extraktion der Modellparameter wurden vier Modelle trainiert (eins pro Klasse). Alle Modelle erzielen vergleichbare Ergebnisse und erreichen eine mittlere quadratische Abweichung von deutlich unter 0.001 bei standardisierten Eingabewerten. Bild 2 zeigt exemplarisch die vorhergesagten Parameterwerte von 108 Zugproben aus dem Test-Datensatz aufgetragen

über den tatsächlich zur Simulation verwendeten Sweep-Parametern und verdeutlicht die hohe Qualität der vorhersagen.

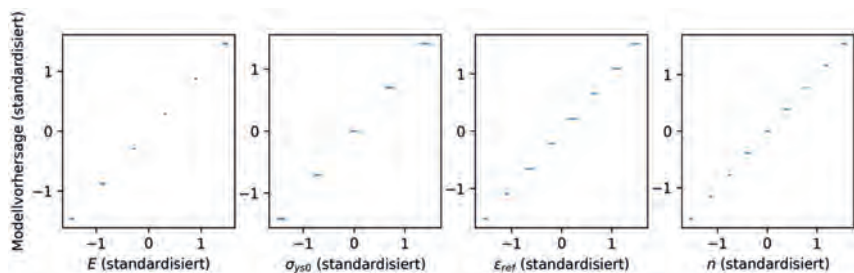


Bild 2: Vorhergesagte Modellparameter aufgetragen über den tatsächlich zur Simulation verwendeten Parametern

3. Unüberwachtes Lernen von bedeutungsvollen latenten Variablen aus Zugproben

3.1 Methode

In diesem Abschnitt beschreiben wir unsere Lösung für das Lernen physikalischer Konzepte auf der Grundlage der oben beschriebenen Spannungs-Dehnungs-Kurven. Dabei ist das Ziel, bedeutungsvolle Latente Variablen aus den Daten allein zu erlernen, d.h. ohne dem Modell Informationen über physikalische Modelle oder deren Parameter zu übergeben. Dies kann z.B. mittels generativer Modelle erreicht werden. Solche Modelle extrahieren die zugrundeliegenden Faktoren der Variation in den Daten (Beobachtungen). Oft wird dabei die gemeinsame Wahrscheinlichkeit $p(\mathbf{x}, \mathbf{z})$ modelliert, wobei $\mathbf{x} \in \mathbb{R}^n$ die Beobachtungen darstellen (in unserem Fall Spannungs-Dehnungskurven) und $\mathbf{z} \in \mathbb{R}^m$ die latenten Variablen. In unserem, wie in den meisten Fällen, ist dabei $n \gg m$. Die Kernannahme dieses Ansatzes ist, dass die zugrundeliegenden Faktoren der Variation in den Daten auch die zugrundeliegenden physikalischen Konzepte beschreiben, die die Beobachtungen verursachen. Im Zugprobenfall wären dies die mechanischen Deskriptoren, die das elast-plastische Verhalten abbilden. In unserem Versuch trainieren wir einen Variational Autoencoder [20] bei dem sowohl der Encoder $q_\phi(\mathbf{z}|\mathbf{x})$ als auch der Decoder $p_\theta(\mathbf{x}|\mathbf{z})$ mit den Parametern ϕ und θ als rekurrente NNs implementiert sind. Jede einzelne Spannungs-Dehnungs-Kurve $\mathbf{x} \in \mathbb{R}^{N \times 3}$ wird als dreidimensionale Sequenz modelliert, wobei die Spannungs- und Dehnungswerte und der relative Zeitstempel die Datenpunkte selbst darstellen und N die Anzahl der Datenpunkte in der Spannungs-/Dehnungskurve beschreibt. Angesichts der kontinuierlichen Natur der Konzepte, an denen wir interessiert sind,

wählen wir einen Gauß-Prior, so dass $z \sim \mathcal{N}(0,1)$ und $\dim(z) = 5$. Auch in diesem Experiment trainieren wir ein Modell pro Klasse.

3.2 Ergebnisse

Die Ergebnisse der VAE-Modelle für die vier Klassen der Zugproben sind vergleichbar. In diesem Abschnitt stellen wir exemplarisch die Ergebnisse des Modells für die Zugproben vor, bei deren Simulation das Swift-Modell ohne kinematische Verfestigung verwendet wurde.

Die Qualität gelernter Repräsentationen zu bewerten ist nicht trivial. In vielen Fällen wird dazu anhand von Downstream-ML Problemen getestet, ob einfache Modelle wie z.B. lineare Regressionen mit den Repräsentationen als Input gute Ergebnisse liefern können. Im konkreten Fall, kann z.B. getestet werden, inwiefern ein einfaches Modell eine Abbildung von den Mittelwerten der Verteilungen im Latenten Raum $q(z|x)$ zu den Parametern des physikalischen Modells erlernen kann, wobei $x \in \mathbb{R}^{N \times 3}$ wie im vorherigen Abschnitt die Zugproben darstellen und $z \in \mathbb{R}^m$ die latenten Variablen. Die lineare Regression erreicht in diesem Fall einen R^2 -Wert von 0.91 und ist damit zwar nicht so akkurat wie das NN aus Kapitel 2.2, gemessen an der Zahl der Parameter im Regressionsmodell aber überraschend performant.

Bei generativen Modellen können alternativ auch die vom Decoder generierten Zugproben analysiert werden, um die Repräsentationen qualitativ zu bewerten. Bild 3 zeigt 25 Beispiele für Zugproben, die von dem Decoder des VAE generiert wurden. Jede Zeile in der Scatterplot-Matrix stellt den Einfluss der Veränderung einer einzigen latenten Variable auf die generierten Spannungs-Dehnungskurven dar. In der unteren Zeile, beispielsweise, liegen allen Plots die Werte $z_1 = z_2 = z_3 = z_4 = 0$ zugrunde, während die Werte für z_5 schrittweise von -3 bis 3 (dem Bereich der größten Dichte in den entsprechenden Verteilungen) erhöht werden. Diese Veränderungen in z_5 scheinen die Steigung im hinteren Bereich der Zugprobe zu beeinflussen. Auffällig ist darüber hinaus, dass die Veränderungen in z_3 keinen Einfluss auf die generierten Samples zu haben scheint. Dass nur vier der fünf latenten Variablen vom VAE zur Codierung der Informationen in den Zugproben genutzt werden war zu erwarten, weil es genau vier unabhängige Sweep-Parameter sind, die der Variation im Datensatz zugrunde liegen.

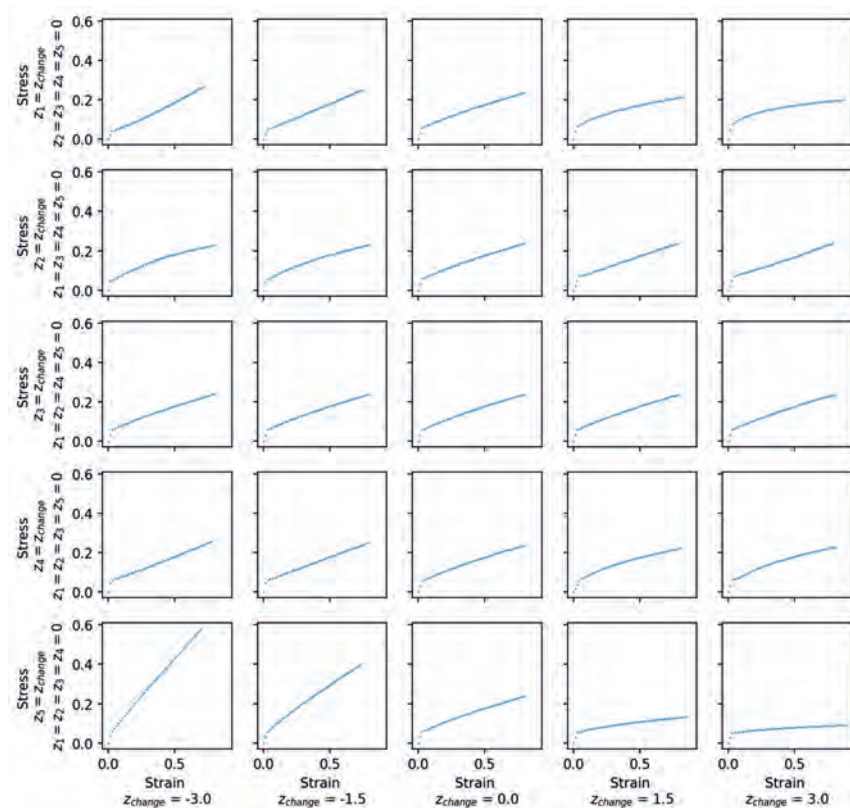


Bild 3: 25 Beispiele für die vom Decoder des VAE generierte Spannungs-Dehnungskurven der Zugproben (normiert)

5. Zusammenfassung

Bisherige Arbeiten haben die Anwendbarkeit von NNs in den Naturwissenschaften bewiesen. Doch während physikalische Gesetze bereits direkt aus Daten gelernt werden konnten, stellt das Lernen von interpretierbaren Materialkonzepten oder Modellen auf der Grundlage von Spannungs-Dehnungs-Daten in den Materialwissenschaften eine bislang zum Teil unzureichend gelöste Herausforderung dar. Dieser Fakt kann auch dadurch begründet werden, dass verschiedene skalenübergreifende Zusammenhänge in hierarchischen Materialstrukturen eine erhebliche Komplexität im Verformungsverhalten bedingen.

Im Rahmen dieser Arbeit wurde aber erfolgreich gezeigt, dass mithilfe einer Kombination von überwacht angelernten Neuronalen Netzen das geeignetste empirische Materialmodell identifiziert und die hierfür passenden Modellparameter ausgewählt werden können. Die Ergebnisse zeigen, dass die Materialmodelle zur elasto-plastischen Verformung der Zugproben zuverlässig klassifiziert werden können, sofern sich die Verläufe der daraus entstehenden Kurven hinreichend unterscheiden. Im Falle des Einflusses der kinematischen Prager Verfestigung der „Swift Zugproben“ ist dies nicht der Fall, sodass die Klassifikation bei einigen Proben fehlschlägt. Die Regression der Modellparameter zeigt für alle Materialmodelle sehr gute Ergebnisse mit minimalen Abweichungen vom Zielwert des Materialdeskriptors.

Des Weiteren wurde unüberwachtes Lernen eingesetzt, um ein generatives Modell zu trainieren. Ziel war ein Modell dessen latente Variablen bedeutungsvoll hinsichtlich des zugrundeliegenden physikalischen Modells sind. Diese latenten Variablen sorgen für eine Interpretierbarkeit des Modells, die mit überwachtem Lernen nicht erreicht werden kann. Die Ergebnisse zeigen, dass einige dieser Variablen in der Tat Eigenschaften des Verlaufes der Spannungs-Dehnungs-Kurven enkodieren, die wiederum mit physikalischen bzw. mechanischen Eigenschaften des Materials verbunden sind. Zwei Beispiele hierfür sind die Steigung der Kurven jenseits der Elastizitätsgrenze sowie die Krümmung der Kurven in ebendiesem Bereich. Die Aussagekraft der gelernten latenten Variablen zeigt sich neben dieser qualitativen Bewertung auch durch das erfolgreiche Training eines linearen Regressionsmodells auf Basis ebendieser Variablen als Eingabewerte.

Die Ergebnisse zeigen, dass NNs nicht nur eingesetzt werden können, um automatisiert geeignete Materialmodelle und deren Parameter auszuwählen, sondern auch physikalische Konzepte zur Beschreibung elasto-plastischer Verformung unüberwacht in latenten Variablen erlernen können. In weiteren Untersuchungen gilt es, die Extrapolationsfähigkeit sowie die Praxistauglichkeit der Klassifikations- und Regressionsmodelle anhand von experimentellen Daten zu überprüfen. Die Beobachtungen legen nahe, dass sich an Stelle einzelner Modelle auch ein einziges generatives Modell trainieren ließe, welches alle Eigenschaften des Materials in latenten Variablen abdeckt und Klassifikation und Regression mit vergleichsweise simplen Modellen ermöglicht.

Um es Materialwissenschaftlern, Prüfern oder auch anderen Ingenieuren zu ermöglichen, mithilfe von KI schneller Parameter für (neue) Materialsysteme zu bestimmen und diese in den Entwicklungsprozess zu integrieren, werden präzise, verlässliche und interpretierbare Modelle benötigt. Diese Arbeit geht einen Teilschritt und zeigt Herausforderungen und Chancen von KI in den Materialwissenschaften und der mechanischen Prüfung.

Literaturangaben

- [1] N. Huber, S. R. Kalidindi, B. Klusemann, and C. J. Cyron, "Editorial: Machine Learning and Data Mining in Materials Science," *Frontiers in Materials*, vol. 7, p. 51, Feb. 2020, doi: 10.3389/FMATS.2020.00051/BIBTEX.
- [2] N. Boullé, C. J. Earls, and A. Townsend, "Data-driven discovery of Green's functions with human-understandable deep learning," May 2021, doi: 10.1038/s41598-022-08745-5.
- [3] S. Greydanus, M. Dzamba, and J. Yosinski, "Hamiltonian Neural Networks," p. 8, Jun. 2019, [Online]. Available: <http://arxiv.org/abs/1906.01563>
- [4] B. M. de Silva, D. M. Higdon, S. L. Brunton, and J. N. Kutz, "Discovery of Physics From Data: Universal Laws and Discrepancies," *Frontiers in Artificial Intelligence*, vol. 3, p. 25, Apr. 2020, doi: 10.3389/frai.2020.00025.
- [5] M. Lange-Hegermann, "Linearly Constrained Gaussian Processes with Boundary Conditions," *AISTATS*, 2021, Accessed: May 03, 2022. [Online]. Available: <http://proceedings.mlr.press/v130/lange-hegermann21a/lange-hegermann21a.pdf>
- [6] M. Lange-Hegermann, "Algorithmic Linearly Constrained Gaussian Processes," *NeurIPS*, 2018, Accessed: May 03, 2022. [Online]. Available: <https://papers.nips.cc/paper/2018/file/68b1fbe7f16e4ae3024973f12f3cb313-Paper.pdf>
- [7] S. Desai and A. Strachan, "Parsimonious neural networks learn interpretable physical laws," *Sci Rep*, vol. 11, no. 1, p. 12761, 2021, doi: 10.1038/s41598-021-92278-w.
- [8] M. Choi, D. Flam-Shepherd, T. H. Kyaw, and A. Aspuru-Guzik, "Learning quantum dynamics with latent neural ODEs," Oct. 2021, [Online]. Available: <http://arxiv.org/abs/2110.10721>
- [9] J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, "Recent advances and applications of machine learning in solid-state materials science," *npj Computational Materials*, vol. 5, no. 1, 2019, doi: 10.1038/s41524-019-0221-0.
- [10] K. Kim, L. Ward, J. He, A. Krishna, A. Agrawal, and C. Wolverton, "Machine-learning-accelerated high-throughput materials screening: Discovery of novel quaternary Heusler compounds," *Physical Review Materials*, vol. 2, no. 12, p. 123801, Dec. 2018, doi: 10.1103/PhysRevMaterials.2.123801.

- [11] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, "Quantum-chemical insights from deep tensor neural networks," *Nat Commun*, vol. 8, p. 13890, 2017, doi: 10.1038/ncomms13890.
- [12] Y. Y. Yang, M. Mahfouf, D. A. Linkens, and Q. Zhang, "Tensile Strength Prediction for Hot Rolled Steels by Bayesian Neural Network Model," *IFAC Proceedings Volumes*, vol. 42, no. 23, pp. 255–260, 2009, doi: 10.3182/20091014-3-CL-4011.00046.
- [13] M. Gayatri Vineela, A. Dave, and P. Kiran Chaganti, "Artificial Neural Network based Prediction of Tensile Strength of Hybrid Composites," *Materials Today: Proceedings*, vol. 5, no. 9, pp. 19908–19915, 2018, doi: 10.1016/j.matpr.2018.06.356.
- [14] D. Merayo, A. Rodríguez-Prieto, and A. M. Camacho, "Prediction of Physical and Mechanical Properties for Metallic Materials Selection Using Big Data and Artificial Neural Networks," *IEEE Access*, vol. 8, pp. 13444–13456, 2020, doi: 10.1109/ACCESS.2020.2965769.
- [15] D. Merayo Fernández, A. Rodríguez-Prieto, and A. M. Camacho, "Prediction of the Bilinear Stress-Strain Curve of Aluminum Alloys Using Artificial Intelligence and Big Data," *Metals (Basel)*, vol. 10, no. 7, p. 904, Jul. 2020, doi: 10.3390/met10070904.
- [16] K. Sedighiani, M. Diehl, K. Traka, F. Roters, J. Sietsma, and D. Raabe, "An efficient and robust approach to determine material parameters of crystal plasticity constitutive laws from macro-scale stress-strain curves," *International Journal of Plasticity*, vol. 134, p. 102779, 2020, doi: 10.1016/j.ijplas.2020.102779.
- [17] H. J. Kleemola and M. A. Nieminen, "On the strain-hardening parameters of metals," *Metallurgical Transactions 1974 5:8*, vol. 5, no. 8, pp. 1863–1866, 1974, doi: 10.1007/BF02644152.
- [18] N. Ohno and J. D. Wang, "Kinematic hardening rules with critical state of dynamic recovery, part I: formulation and basic features for ratchetting behavior," *International Journal of Plasticity*, vol. 9, no. 3, pp. 375–390, Jan. 1993, doi: 10.1016/0749-6419(93)90042-O.
- [19] Dr. W. Prager, "The Theory of Plasticity: A Survey of Recent Achievements:," http://dx.doi.org/10.1243/PIME_PROC_1955_169_015_02, vol. 169, no. 1, pp. 41–57, Feb. 2006, doi: 10.1243/PIME_PROC_1955_169_015_02.

- [20] I. Higgins *et al.*, "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework," *5th International Conference on Learning Representations*. 2017.

Automatisierte Modell- und Testgenerierung aus textuellen Anforderungen

Der Einsatz von Natural Language Processing

Dipl.-Math. **R. Gröpler**, M.Sc. **Viju Sudhi**, B.Sc. **Libin Kutty**,
ifak – Institut für Automation und Kommunikation e.V. Magdeburg

Kurzfassung

Die Grundlage für einen effizienten Softwareentwicklungs- und Testprozess ist die Spezifikation von Anforderungen, die die Funktionalität und Qualitätseigenschaften des zu entwickelnden Softwareprodukts beschreiben soll. Anforderungsspezifikationen sind zumeist in natürlicher Sprache formuliert, was die Kommunikation zwischen den unterschiedlichen Beteiligten vereinfacht, allerdings auch zu Fehleranfälligkeit und großem Aufwand bei der Weiterverarbeitung führt. Durch die wachsende Komplexität und Variantenvielfalt von Softwaresystemen wird damit deren Qualitätssicherung zunehmend aufwendiger. Insbesondere für sicherheitskritische Systeme ist ein umfangreiches Testen auf der Basis von Anforderungen notwendig. Methoden der modellbasierten Testautomatisierung bieten die Möglichkeit, diese Schwierigkeiten zu bewältigen. Allerdings besteht immer noch ein großer Aufwand darin, Anforderungs- und Spezifikationsmodelle aus einer großen Menge von funktionalen Anforderungen in natürlicher Sprache zu erstellen. In diesem Beitrag stellen wir einen Ansatz vor, der auf der Verarbeitung natürlicher Sprache (Natural Language Processing – NLP) und der textuellen Ähnlichkeit von Anforderungen und Produktdesignspezifikationen basiert, um menschen- und maschinenlesbare Modelle zu erzeugen. Dies soll den Formalisierungsprozess von Anforderungen automatisiert unterstützen und damit Kosten und Aufwand reduzieren. Das Verfahren wird an einem industriellen Anwendungsfall aus der Eisenbahnbranche demonstriert. Schließlich werden bestehende Werkzeuge zur automatisierten Testfallgenerierung auf die Modelle angewendet, um ausführbare Testfälle ableiten zu können.

1. Einleitung

Immer komplexere und hoch konfigurierbare Softwaresysteme erhöhen den Aufwand für deren Qualitätssicherung. Die schnelle und gleichzeitig qualitativ hochwertige Entwicklung von industriellen Softwareprodukten erfordert einen immer effektiveren Testprozess.

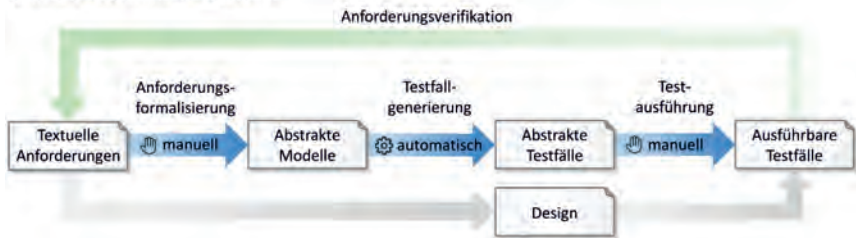
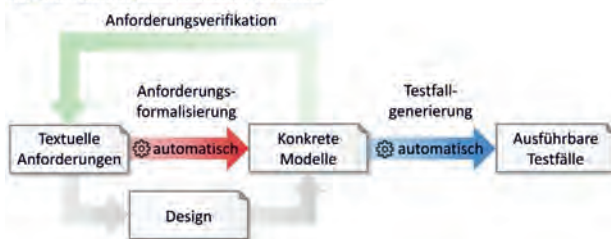
Klassischer modellbasierter Testablauf**NLP-unterstützter modellbasierter Testablauf**

Bild 1: Darstellung des klassischen und eines NLP-unterstützten modellbasierten Testablaufs

Insbesondere für sicherheitskritische Systeme, wie z.B. im Automobil- und Bahnbereich, sind umfangreiche Tests auf Basis von Anforderungen notwendig. Jede manuelle Verarbeitung, wie z.B. Anforderungsverifikation und Testgenerierung, aus textuellen Anforderungen ist jedoch zeitaufwändig, fehleranfällig und erfordert zudem viel Expertenwissen. Methoden zur modellbasierten Testautomatisierung verfolgen das Ziel, diese Schwierigkeiten zu bewältigen. Formale Modelle dienen als Grundlage für die Automatisierung einer Vielzahl weiterer Prozessschritte.

Der klassische modellbasierte Testablauf besteht aus der Erstellung von abstrakten Modellen und daraus wiederum abstrakten Testfällen, die dann in ausführbare Testfälle umgewandelt werden (s. Bild 1). Es wurde bereits eine umfangreiche Toolkette für das modellbasierte Testen entwickelt [1,2]. Testfälle gelten als abstrakt, wenn das beschriebene Verhalten unabhängig von einer bestimmten Realisierung ist, aber detailliert genug, um für diese spezifiziert zu werden. Abstrakten Modellen und Testfällen fehlt die Information über die realen Implementierungsobjekte, wie Komponenten, Signale, und Parameter, so dass es notwendig ist, die abstrakten Objekte auf reale Objekte abzubilden, wenn man von abstrakten Testfällen zur Testausführung übergeht. Es hängt stark vom Design des Systems ab, welche Signale und Parameter existieren und wie sie definiert sind. Insbesondere wenn die interagierenden

Komponenten in der Beschreibung nicht erwähnt werden, kann es sehr mühsam sein, festzustellen, auf welche Komponenten sich das Signal oder die gesamte Anforderung bezieht. So kann es viele Detailfragen geben, die die Umwandlung von abstrakten in konkrete Testfälle zu einem fehleranfälligen und zeitaufwendigen manuellen Prozess machen. In diesem späten Stadium des Testprozesses müssen Probleme gelöst werden, z. B. dass Teile der Anforderungen nicht verifizierbar sind. Dies kann den modellbasierten Ansatz ineffizient machen, da die Anforderungen und die Designspezifikation dann erneut überarbeitet werden müssen.

Aktuelle Fortschritte in der Verarbeitung natürlicher Sprache (NLP) zeigen vielversprechende Ergebnisse für die Unterstützung eines breiten Spektrums von Aufgaben der Anforderungsanalyse. Daher sind NLP-Techniken von wachsendem Interesse für die Automatisierung verschiedener Softwaretestaktivitäten wie der Modell- und Testfallgenerierung. In den letzten Jahren wurde eine beträchtliche Anzahl von NLP-Ansätzen und -Werkzeugen untersucht, die darauf abzielen, Modelle oder Testfälle aus den zugrundeliegenden Anforderungsdokumenten zu generieren [3]. Aufgrund der Unzulänglichkeiten der natürlichen Sprache erzeugen die meisten der bestehenden NL-basierten Ansätze zwar abstrakte Modelle und Testfälle, es fehlen jedoch Informationen über reale Objekte, ihre Beziehungen und ihr Design [4].

In diesem Beitrag wollen wir Produktdesign-Spezifikationen für einen viel präziseren und effizienteren Anforderungsformalisierungsprozess in Betracht ziehen. Unser Ansatz nutzt NLP-Techniken zur automatischen Generierung von Anforderungsmodellen aus natürlichsprachlichen Anforderungen und Designspezifikationen. Wir führen eine textuelle Ähnlichkeitsanalyse zwischen den Anforderungen und den Objektbeschreibungen unter Verwendung verschiedener NLP-Algorithmen durch. Die generierten Modelle werden in einer einfachen, selbst erstellten, maschinen- und menschenlesbaren Sprache dargestellt. Die Integration von Informationen aus der Designspezifikation in diesem frühen Stadium des Testprozesses soll dem Anforderungs- oder Testingenieur eine schnellere Rückmeldung darüber geben, ob die Anforderungen überprüfbar und korrekt entworfen sind. Damit werden insbesondere der stark iterative Prozess der Anforderungsverifikation und die Prozesse der Testgenerierung und -ausführung effizienter gestaltet, s. Bild 1.

2. Anwendungsfall

Für die Demonstration unseres Verfahrens betrachten wir einen Anwendungsfall aus der Eisenbahnindustrie. Wir verwenden die Daten des Antriebssteuerungssystems (PPC) von Alstom (vormals Bombardier Transportation). Das PPC ist Teil eines großen, komplexen,

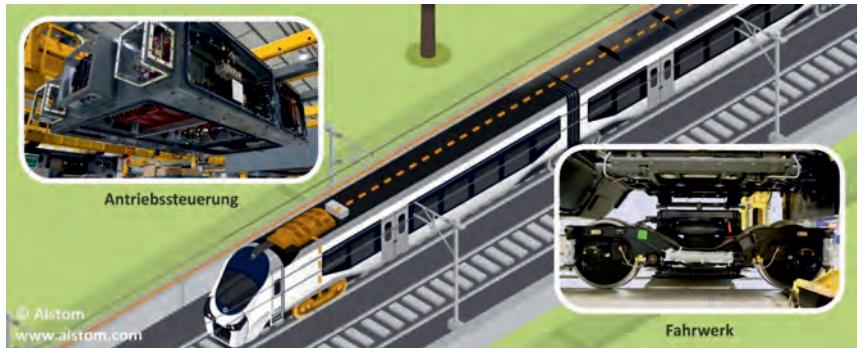


Bild 2: Antriebssteuerungssystem (PPC) eines Schienenfahrzeugs von Alstom

sicherheitskritischen Systems. Es übernimmt die Steuerung des gesamten Antriebssystems, einschließlich der Steuerungssoftware und der elektrischen Funktionen (s. Bild 2).

Die Anforderungen werden in Textform in das Anforderungsmanagement-Tool IBM DOORS geschrieben. Sie sind in englischer Sprache verfasst und können aus mehreren Sätzen bestehen. Sie folgen keinem vorgeschriebenen Format, um sich beim Schreiben nicht auf die Syntax zu konzentrieren. Um die Standards der Sicherheitsintegritätsstufe 2 (SIL2) zu erfüllen, wird ein Designdokument in Textform verfasst und ebenfalls in DOORS bearbeitet. Die Anforderungen und das Designdokument werden von Hand erstellt und können manuell verknüpft werden. Die Systemarchitektur und die Softwaremodule werden als Blöcke in Matlab Simulink modelliert, wobei ein modellbasierter Entwurfsansatz verwendet wird [5]. Wir gehen im Folgenden davon aus, dass für jedes Objekt eine kurze Beschreibung in natürlicher Sprache verfügbar ist. Diese Annahme sollte in den meisten industriellen Produktentwicklungen zutreffen, da Anforderungsingenieure auch Objekte durch eine textuelle Beschreibung identifizieren müssen. Diese Informationen können im implementierten Code, in der unternehmensinternen Standardisierungsdokumentation oder strukturierter in Form einer Tabelle in einem Design-Spezifikationsdokument enthalten sein.

3. Methodik

Im Folgenden stellen wir geeignete NLP-Methoden vor, um den Prozess der Erstellung von Anforderungsmodellen aus den Anforderungen und der Designspezifikation zu automatisieren. Bild 3 veranschaulicht den Verfahrensablauf unseres Ansatzes anhand einer beispielhaften Anforderung mit zugehörigen Signalbeschreibungen. Die Anforderung wurde auf Deutsch und verallgemeinert dargestellt, da die von uns verwendeten Daten vertraulich sind.

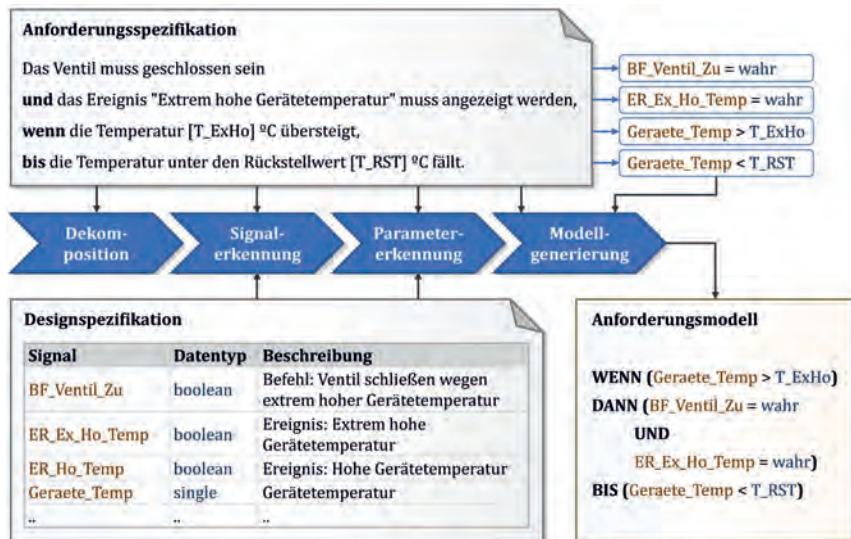


Bild 3: Beispielhafte Anforderung mit einem Überblick über den Verfahrensablauf

Der Prozess kann in die folgenden Schritte unterteilt werden. Die Anforderungen und die Designspezifikation in natürlicher Sprache dienen als Eingangsdaten für das Verfahren. Die Anforderung wird zunächst in Teilsätze zerlegt (Abschnitt 3.1). Jede Teilanforderung wird mit den Signalbeschreibungen in der Designspezifikation verglichen, um die benötigten Signale zu identifizieren (Abschnitt 3.2), dann werden die entsprechenden Parameter ermittelt (Abschnitt 3.3). Für jede Teilanforderung wird eine logische Relation gebildet, die dann in eine logische Struktur und schließlich in ein Anforderungsmodell (Abschnitt 3.4) übersetzt wird.

3.1 Dekomposition

Es ist notwendig, die langen, komplexen industriellen Anforderungen zunächst in kürzere Teilsätze zu zerlegen, die elementare Anweisungen einzeln beschreiben. Dieser Prozess erfordert ein Verständnis sowohl der Struktur (Syntax) als auch der Bedeutung (Semantik) des Anforderungstextes. Dafür haben wir zunächst einen einfachen Ansatz basierend auf Schlüsselwörtern verwendet. Die Konjunktionen zur Einleitung von Nebensätzen (Bedingungssätzen) wie „wenn“, „falls“, „während“ und „bis“ und zur Verbindung von Teilsätzen wie „und“ und „oder“ wurden in der Spanne des Anforderungstextes identifiziert und als Grenze

der Teilsätze betrachtet. In dem betrachteten Anwendungsfall mit gut strukturierten Anforderungen funktionierte dieser Ansatz sehr gut.

In einigen Fällen sollte die Konjunktion nicht als Grenze angesehen werden, beispielsweise bei den Ausdrücken „zwischen Sensor 1 *und* 2“ oder „kleiner *oder* gleich“. Um solche Anforderungen korrekt zu zerlegen, haben wir neben den Schlüsselwörtern auch die syntaktischen Abhängigkeiten des Anforderungstextes berücksichtigt unter Benutzung des frei verfügbaren Tools spaCy [6]. In Bild 3 ist die Zerlegung einer Anforderung durch Zeilenumbrüche veranschaulicht. Unser Anwendungsfall erfordert nur einen einfachen Dekompositionsalgorithmus, wie er hier vorgeschlagen wird, um korrekt zu funktionieren. Dieser könnte jedoch auf weitere linguistische Muster oder auf komplexere Algorithmen erweitert werden.

Für die weiteren Schritte unseres Verfahrens ist die Zerlegung von Anforderungen in Teilsätze am besten geeignet. Ein kürzerer Textabschnitt (nur die Substantivgruppe) führt zu einem Informationsverlust, und ein längerer Abschnitt (ein ganzer Satz) führt zu einer fehlerhaften Erkennung von Signalen und Parametern.

3.2 Signalerkennung

Nachdem die Anforderung zerlegt wurde, wollen wir das entsprechende Signal für jeden Teilsatz ermitteln. Dazu verwenden wir die Signalbeschreibungen in der Designspezifikation (wie in Schritt 2 von Bild 3 gezeigt). Zur Bestimmung der textuellen Ähnlichkeit gibt es unterschiedliche NLP-Methoden:

- klassische Techniken wie Term Frequency (TF) [7] und FuzzyWuzzy [8]
- statische Worteinbettungen wie GloVe [9] und fastText [10] und
- kontextuelle Einbettungen von Sentence-BERT [11].

Während Term Frequency den Text unter Berücksichtigung der Häufigkeit von Wörtern analysiert, sucht FuzzyWuzzy nach teilweisen (oder ungenauen) Übereinstimmungen zwischen Satzpaaren. Bei statischen Einbettungen wird jedem Wort ein Vektor fester Größe zugeordnet und bei kontextuellen Einbettungen wird der gesamte Kontext des Satzes in einem Einbettungsvektor erfasst. Damit kann für jede Signalbeschreibung die Kosinus-Ähnlichkeit

$$\cos(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| |\vec{b}|}$$

zur Teilanforderung berechnet werden, wobei $\vec{a}$ die Vektordarstellung der Teilanforderung ist, deren Signal wir zu erkennen versuchen, und $\vec{b}$ die Vektordarstellung der Signalbeschreibung ist. Das Signal mit dem höchsten Ähnlichkeitswert wird dann als das relevanteste ausgewählt (s. Bild 4).

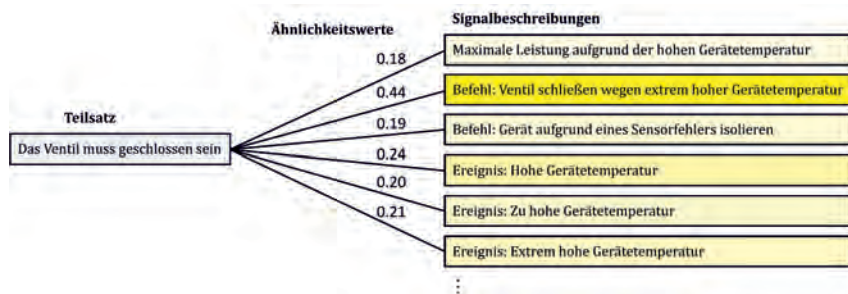


Bild 4: Ähnlichkeitswerte zwischen einem Anforderungs-Teilsatz und Signalbeschreibungen

Nach einer ausführlichen Evaluation [12] stellte sich heraus, dass Sentence-BERT am geeignetsten ist. Obwohl das Sentence-BERT-Modell sehr gute Resultate liefert, gibt es auch einige Fälle, in denen das Modell nicht zwischen "Gerät mit Fehler" und "Gerät ohne Fehler" unterscheiden konnte. Dies kann auf die Trainingsmethode zurückzuführen sein, die darauf abzielt, die Einbettungen ähnlicher Wörter in ihrem Repräsentationsraum so nahe wie möglich zu bringen. Infolgedessen gibt das Modell für solche Paare sehr hohe Kosinus-Ähnlichkeitswerte an und erkennt gelegentlich ein falsches Signal.

3.3 Parametererkennung

Nach der Erkennung der Signale müssen wir die Parameter bestimmen, die den Signalwerten zugeordnet oder mit ihnen verglichen werden. Wenn symbolische oder numerische Parameter in der Anforderung verwendet werden (z. B. T_ExHo), können sie durch einfachen Mustervergleich in der Designspezifikation leicht identifiziert werden. In vielen Fällen sind die Parameterwerte jedoch implizit in der Beschreibung gegeben (vor allem boolesche Parameter) und können nur durch einen semantischen Vergleich zwischen den beiden textuellen Formulierungen identifiziert werden.

In diesen Fällen müssen wir feststellen, ob die Teilanforderung mit der entsprechenden Signalbeschreibung übereinstimmt oder ihr widerspricht, um den Parameter zu erhalten. Wenn beispielsweise der Teilsatz "Das Ventil soll geöffnet werden" und die Signalbeschreibung "Ventil schließen ..." einander widersprechen, wird der Parameter auf „falsch“ gesetzt. Stimmt dagegen die Signalbeschreibung mit der Teilanforderung überein, soll der Parameter als „wahr“ erkannt werden. Die Erkennung von Widersprüchen zählt zu den besonders schwierigen NLP-Aufgaben.

Es gibt unterschiedliche NLP-Methoden, die dafür eingesetzt werden können:

- Antonym-Synonym-Erkennung,
- Paraphrasen-Erkennung und
- Textuelles Schließen (Textual Entailment – TE).

Bei der Antonym-Synonym-Erkennung wird geprüft, ob das Verb in der Teilanforderung ein Antonym (Gegensatzwort) zu dem Verb in der Signalbeschreibung ist. Dazu können klassische Wörterbücher [13] oder Modelle [14] verwendet werden. Während diese Methoden Paare wie „aktiviert“ und „deaktiviert“ recht gut erkennen, versagen sie bei Negationen in Wortgruppen wie „nicht aktiviert“. Zur Erkennung von Paraphrasen (Umschreibungen) können die Ähnlichkeitswerte aus dem Schritt der Signalerkennung verwendet werden, um den Parameter zu identifizieren. Wenn die Werte über einem Schwellenwert liegen, wird der Parameter als „wahr“ gesetzt und andernfalls als „falsch“. Dahinter steht der Gedanke, dass man bei der Umschreibung eines Satzes in einen anderen erkennen könnte, ob er mit dem anderen Satz übereinstimmt oder ihm widerspricht. Beim Textual Entailment wird geprüft, ob sich eine Aussage als Schlussfolgerung einer anderen Aussage ableiten lässt. Wir haben ein vortrainiertes Modell [15] verwendet, dass mit drei Klassifizierungsmarken (Übereinstimmung, Widerspruch, neutral) trainiert wurde. Durch eine ausführliche Evaluation aller genannten Methoden [12] stellte sich heraus, dass Textual Entailment am geeignetsten ist.

3.4 Modellgenerierung

Nachdem die Signale und Parameter für jede Teilanforderung identifiziert wurden, formulieren wir eine logische Relation (in der Form „Signal-Operator-Parameter“) für jeden dieser Teilsätze. Wir stützen uns auf einen wörterbuchbasierten Ansatz [16], um Vergleichsoperatoren (falls vorhanden) in den Teilsätzen zu identifizieren. In Bild 3 sind die Relationen für die Beispielanforderung dargestellt.

Um aus diesen Relationen Anforderungsmodelle zu generieren, führen wir eine domänen-spezifische Sprache (DSL) mit abstrakten logischen Blöcken ein. Sie bildet die Relationen aus einem Bedingungssatz („wenn“/„falls“/„während“) auf einen WENN-Block, die aus Hauptsätzen auf einen DANN-Block und die aus Bedingungssätzen beginnend mit „bis“ auf einen BIS-Block ab. Logische Operatoren (UND/ODER) zwischen Relationen werden ebenfalls in diesen DSL-Blöcken berücksichtigt. Diese Abbildung hat zum Ziel, eine möglichst einfache menschen- und maschinenlesbare Sprache zur Verfügung zu stellen. Außerdem soll sie die Umwandlung einfach und flexibel gestalten, so dass Wege für die Integration mit anderen anspruchsvollen Sprachen offenstehen.

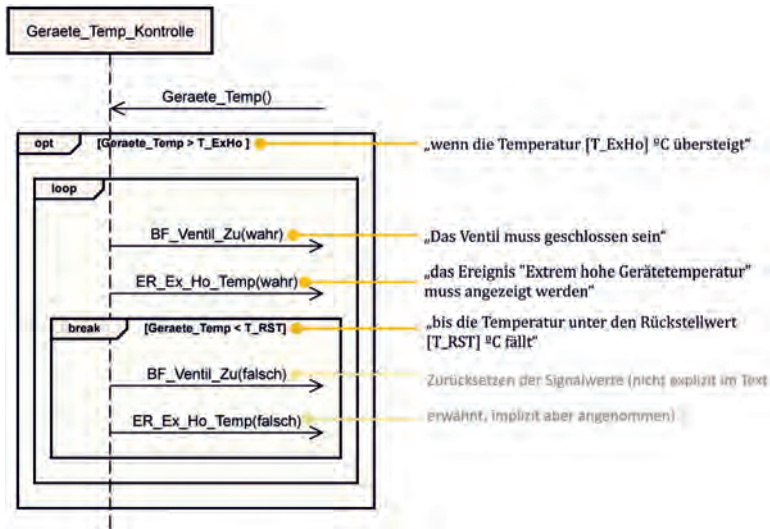


Bild 5: Sequenzdiagramm zur Beispiel-Anforderung

Die resultierenden Modelle können beispielsweise in

- Matlab-Simulink-Modelle oder
- UML-Sequenzdiagramme

umgewandelt werden, je nachdem, was der Endbenutzer wünscht. Ein Anforderungsmodell kann somit in verschiedenen Formaten dargestellt werden. Beispielhaft ist in Bild 5 ein Sequenzdiagramm zur Beispiel-Anforderung aus Bild 3 gezeigt.

3.5 Testgenerierung

Nach der Modellgenerierung können mit einer bestehenden Methode zur modellbasierten Testgenerierung [17] automatisch Testfälle generiert werden. Dazu wird das Anforderungsmodell zunächst in eine Zustandsmaschine umgewandelt. Es ist auch möglich, mehrere Anforderungsmodelle zu einer Zustandsmaschine zusammenzufügen, um Abhängigkeiten und komplexeres Systemverhalten zu modellieren [18]. Der Zustandsautomat wird in eine spezielle Form eines Petrinetzes transformiert, aus dem dann nach Auswahl eines bestimmten graphenbasierten Abdeckungskriteriums wie „alle Pfade“, „alle Entscheidungen“, „alle Stellen“ oder „alle Transitionen“, Testfälle in Form von Sequenzdiagrammen generiert werden. In Bild 6 wird ein Beispiel für einen Testfall für das betrachtete Antriebssteuerungssystem gezeigt.

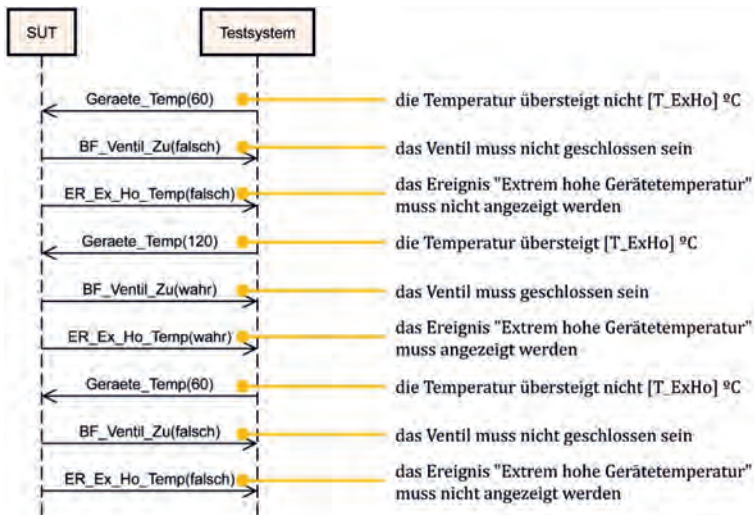


Bild 6: Testfall für die Beispiel-Anforderung des Antriebssteuerungssystems

4. Zusammenfassung und Ausblick

In diesem Beitrag haben wir einen NLP-basierten Ansatz zur automatisierten Anforderungsformalisierung unter Verwendung natürlichsprachlicher Anforderungen und Designspezifikationen zur Testfallgenerierung vorgestellt. Wir haben verschiedene NLP-Methoden für die einzelnen Schritte unseres Verfahrens vorgestellt und sie an einem industriellen Anwendungsfall aus der Eisenbahnbranche demonstriert. Die besten Ergebnisse liefern moderne vortrainierte Modelle wie Sentence-BERT.

Der vorgestellte Ansatz automatisiert den Prozess der Anforderungsformalisierung in hohem Maße, was den Anforderungsingenieur z.B. bei der Anforderungsverifikation und Testfallgenerierung unterstützt. Die Implementierung der entwickelten Methode wurde auf GitHub [19] veröffentlicht. Darüber hinaus wurde eine grafische Oberfläche für den Algorithmus entwickelt, der ein sehr effizientes Erstellen und Anpassen der Anforderungsmodelle ermöglicht.

Für die Zukunft planen wir, diesen Ansatz um mehrere Funktionen zu erweitern, die unsere Methodik verbessern können, z. B. die Behandlung nichtfunktionaler Eigenschaften (wie zeitliches Verhalten, Performance), ein domänenspezifisches Vortrainieren und Fein-Tuning verfügbarer Modelle (wie Sentence-BERT) oder das Lernen aus korrigierten Vorschlägen eines Domänenexperten (Online Machine Learning). Eine weitere interessante Erweiterung wäre die Unterstützung des Anforderungsentwurfsprozesses und der manuellen Erstellung

der Designspezifikation durch die Bereitstellung von Empfehlungen für ähnliche Anforderungen und Komponenten.

Danksagung

Diese Forschung wurde durch das BMBF im Rahmen der ITEA3-Projekte XIVT (Förderkennzeichen 01IS18059E) und SmartDelta (Förderkennzeichen 01IS18059E) gefördert. Wir danken Alstom, insbesondere Daran Smalley, für die Bereitstellung eines industriellen Anwendungsfalls zur Demonstration der vorgestellten Methode.

Literatur

- [1] Gröpler, R., Sudhi, V., Calleja García, E.J., Bergmann, A. (2021): NLP-Based Requirements Formalization for Automatic Test Case Generation.- CS&P 2021: 18–30.
- [2] Reider, M., Magnus, S., Krause, J. (2018): Feature-based testing by using model synthesis, test generation and parameterizable test prioritization.- ICSTW, 130–137.
- [3] Zhao, L., et al. (2020): Natural Language Processing (NLP) for Requirements Engineering (RE): A Systematic Mapping Study.- ACM Computing Surveys: 54/3: 1–41.
- [4] Wang, C., Pastore, F., Goknil, A., Briand, L.C. (2022): Automatic Generation of Acceptance Test Cases From Use Case Specifications: An NLP-Based Approach.- IEEE Trans. Software Eng.: 48/2: 585–616
- [5] Simonson, E. (2018): An MBD Adoption Story from Bombardier Transportation.- MATLAB EXPO 2018 Sweden.
- [6] <https://spacy.io/>
- [7] https://scikit-learn.org/stable/modules/feature_extraction.html#text-feature-extraction
- [8] <https://github.com/seatgeek/fuzzywuzzy>
- [9] <https://nlp.stanford.edu/projects/glove/>
- [10] <https://fasttext.cc/docs/en/english-vectors.html>
- [11] <https://github.com/UKPLab/sentence-transformers>
- [12] Gröpler, R., Sudhi, V., Kutty, L., Smalley, D. (2022): Automated Requirement Formalization using Product Design Specifications.- NLP4RE: 1–10.
- [13] <https://wordnet.princeton.edu/>
- [14] <https://github.com/zengnan1997/moe-asd>
- [15] <https://demo.allennlp.org/textual-entailment/elmo-snli>
- [16] <http://www.roget.org/scripts/hier.php?class=I&division=0&section=III>
- [17] Krause, J.: Testfallgenerierung aus modellbasierten Systemspezifikationen auf der Basis von Petrinetzentransformationen. Dissertation, Universität Magdeburg, 2012.
- [18] Magnus, S., Ruß, T., Krause, J., Diedrich, C. (2017): Modellsynthese für die Testfallgenerierung sowie Testdurchführung unter Nutzung von Methoden zur Netzwerkanalyse.- at – Automatisierungstechnik: 65/1: 73–86.
- [19] https://github.com/ifak-prototypes/nlp_reform

Eine Simulationsumgebung für flexible Cyber-Physische Produktionssysteme zur Generierung realistischer Datensätze für maschinelle Lernverfahren

Niklas Widulle, Jonas Ehrhardt, Maria Krantz, Artur Liebert, Anna Nordhausen, Oliver Niggemann,
Institut für Automatisierungstechnik,
Helmut-Schmidt Universität, Hamburg

Kurzfassung

Um die Autonomie von Cyber-Physischen Produktionssystemen (CPPS) in Bereichen wie Anomaliedetektion, Diagnose und Planung zu verbessern, werden immer häufiger Verfahren aus den Bereichen Künstliche Intelligenz (KI) oder Maschinelles Lernen (ML) eingesetzt. Diese werden auf Datensätzen realer CPPS trainiert. Allerdings weisen vorhandene Datensätze Einschränkungen in der Anzahl ihrer Anwendungsfälle, Fehlerverteilungen und der Komplexität der betrachteten Systeme auf. Dies schränkt die Generalisierbarkeit von ML und KI Modellen, die auf solchen Datensätzen trainiert wurden, ein. In diesem Artikel stellen wir die **Flexible Production Simulation (FlPSi)** vor. FlPSi ist eine Simulationsumgebung für CPPS, die in der Gaming-Engine Unity umgesetzt ist. Das ermöglicht zusätzlich zum CPPS Systemverhalten die Simulation physikalischer und grafischer Aspekte eines CPPS. Die daraus resultierenden Datensätze ermöglichen ein umfangreiches Pre-Training, Training, eine umfangreiche Evaluierung und Vergleichbarkeit von KI oder ML Verfahren für CPPS. Zusätzlich leiten wir einen Anforderungskatalog für Simulationen zur Generierung künstlicher realer Trainingsdatensätze für Verfahren aus KI und ML her.

Abstract

To increase autonomy of Cyber-Physical production systems (CPPS) in domains like anomaly detection, diagnosis and planning, an increasing number of approaches from Artificial Intelligence (AI) and Machine Learning (ML) is developed. While AI and ML approaches heavily rely on diverse datasets to enable generalizability, existing CPPS datasets often have restrictions in their variance of faults, system complexity and use cases applications, restricting the AI and ML approaches. Within this paper we introduce our **Flexible Production Simulation (FlPSi)**. FlPSi is a Simulation environment implemented in the gaming-engine Unity. This enables the simulation of physical and graphical implications in addition to CPPS system behavior. The resulting datasets are suited for extensive pre-training, training, evaluation and comparison of

AI and ML approaches for CPPS. Additionally, we introduce a requirement specification of simulations for generating artificial realistic training datasets for AI and ML approaches.

1. Einleitung

CPPS bestehen aus interagierenden physischen und informationstechnischen Komponenten [7]. Kernaufgabe von CPPS ist die teilweise oder vollständig autonome Herstellung von Produkten [10, 11, 12]. Dabei kommen Verfahren aus den Domänen KI und ML zum Einsatz, um die Systemautonomie in Bereichen wie Optimierungen, Anomalieerkennung, Diagnose oder Planungsaufgaben zu verbessern [8].

Zwar sind KI und ML Ansätze leichter generalisierbar - also auf neue und unbekannte Situationen anwendbar - jedoch werden dazu umfangreiche Trainingsdatensätze benötigt, aus denen Vorhersagen oder Optimierungsstrategien gelernt werden können. Bisherige Trainingsdatensätze realer CPPS für KI und ML Verfahren weisen dabei folgende Defizite auf: (i) *Spezifische Anwendungsfälle*: Um eine generalisierbare Anwendung von KI und ML Verfahren zu gewährleisten müssen diese auf Datensätzen trainiert werden, welche ebenfalls Varianz in den aufgezeichneten Systemen und deren Use Cases aufweisen. Homogene Datensätze erschweren die Generalisierbarkeit von KI oder ML Verfahren. (ii) *Fehlerverteilungen*: In realen Systemen treten Anomalien oder Fehler nur selten und in geringer Varianz auf. Solche engen Fehlerverteilungen erschweren das Erlernen von Fehlereigenschaften. (iii) *Geringe Systemvariabilität*: Die Variabilität von Produkten und Systemkomponenten in bisherigen Datensätzen ist limitiert. Dadurch können keine asymptotischen Analysen von von KI und ML Verfahren bezüglich bspw. ihrer Performanz durchgeführt werden.

Um umfangreiches Pre-Training, Training, Evaluation und Vergleichbarkeit von KI und ML Verfahren im Bereich CPPS zu ermöglichen, sehen wir Bedarf an Datensätzen, welche die oben genannten Defizite ausgleichen. Hierbei wird als Pre-Training das initiale Einlernen des neuronalen Netzes und die damit verbundene Veränderungen der Gewichte für eine bestimmte Aufgabe, wie z.B. Anomaliedetektion von simulierten Produktionsanlagen, bezeichnet. Diese Gewichte können dann für eine weitere Aufgabe, wie z.B. der Anomaliedetektion realer Produktionsanlagen, mit demselben Netz initiiert werden, um das eigentliche Training zu erleichtern. Insbesondere physikalische Simulationen aus Gaming-Engines können hierbei in Betracht kommen. Wir formulieren folgende Forschungsfragen (FF):

FF1: Welche Anforderungen bestehen an eine Simulation von CPPS zur Generierung realistischer künstlicher Trainingsdatensätze für KI und ML Verfahren?

Abgeleitet von den Defiziten (i-iii) erheben wir einen Katalog von minimalen Anforderungen, die eine Simulation erfüllen muss, um realistische, künstliche Datensätze generieren zu können.

FF2: Eignen sich Gaming-Engines, um die Anforderungen aus FF1 einzuhalten und realistische künstliche Trainingsdatensätze für KI und ML Verfahren zu erstellen?

Gaming-Engines, insbesondere Grafik- und Physik-Engines, bieten ein breites Anwendungsspektrum zur Modellierung realistischer Systeme. Die eingebetteten physikalischen und grafischen Zusammenhänge lassen sich auch für die Modellierung von CPPS und die Generierung künstlicher Daten nutzen.

FF3: Welche Modelle und Fehlerzustände lassen sich bereits mit dem hier vorgestellten Tool simulieren, um künstliche realistische Datensätze für das (Pre-)Training zu erstellen?

Die Anforderungen an Datensätze (FF1) und Simulationsumgebungen (FF2) bedingen die Modellierung von CPPS in der Simulation. Die möglichen Modellvariationen sowie die adaptierbaren Fehlerzustände (FF3) werden anschließend abgeleitet.

Als Antwort auf die gestellten Forschungsfragen stellen wir in diesem Artikel FlIPSi vor. FlIPSi ist eine Simulationsumgebung für CPPS, umgesetzt in der Gaming-Engine Unity. In der Simulationsumgebung können modulare CPPS zusammengestellt werden, anhand derer realistische Datensätze zum Pre-Training, Training oder der Evaluation von KI oder ML Verfahren ausgeleitet werden. Die Modularität ermöglicht eine automatische Generierung von CPPS Datensätzen, in den Dimensionen Komplexität, Funktion und Fehlerbehaftung. Dadurch können ML und KI Verfahren auf asymptotisches Verhalten in den verschiedenen Dimensionen getestet werden. Virtuelle Sensorik kann innerhalb der Simulationsmodelle positioniert werden, um zusätzliche Datenströme auszuleiten. FlIPSi kann ebenfalls zum Test und Validierung von Neu- oder Reparametrisierung von CPPS verwendet werden.

Der verbleibende Artikel ist wie folgt strukturiert: In Kapitel 2 werden Probleme und Defizite von bisher existierenden Datensätzen und Simulationsumgebungen dargestellt und ein Anforderungskatalog für Simulationen zur Generierung realistischer künstlicher Datensätze erstellt. Kapitel 3 stellt FlIPSi dar, wie die erarbeiteten Herausforderungen erfolgreich angegangen werden können. In Kapitel 4 werden Vor- und Nachteile unseres Ansatzes diskutiert. Kapitel 5 schließt mit einer Zusammenfassung und einem Ausblick ab.

2. Probleme und Defizite von existierenden Datensätzen

Für die Simulation von CPPS gibt es bislang vornehmlich zwei Richtungen: Die Simulation des funktionalen Systemverhaltens mithilfe von Tools wie Modelica [15] oder Simulink oder alternativ die Nutzung einer spezialisierten, kommerziellen Software wie FlexSim oder Plant Simulation.

Unabhängig ihres Ansatzes muss die Simulation eine Reihe von Anforderungen erfüllen, um für die Nutzung mit ML Verfahren verwendbar zu sein. Balzereit et al stellen in [4] eine Reihe von Anforderungen vor, die eine Benchmark erfüllen muss, um für die Erprobung von ML Algorithmen in der Diagnose von CPPS tauglich zu sein. Diese sind:

- (A1) Hybride Systeme: Die Benchmark muss hybride Systeme abbilden, die kontinuierliche und diskrete Größen enthalten.
- (A2) Verschiedenartige Fehlerarten: Die Benchmark muss eine Vielzahl unterschiedlicher Fehlerarten abbilden können.
- (A3) Variable Komplexität: Die Komplexität des Benchmarks muss von einfachen bis hin zu sehr komplexen Systemen skalierbar sein.

Balzereit et al stellen eine Benchmark vor: Als solches ist die Ausrichtung anders als die von FLiPSi. FLiPSi ist eine Simulation zur allgemeinen Verwendung. Beispielsweise soll FLiPSi auch die Anwendung von Planungsalgorithmen ermöglichen. Die Ausleitung von Daten, wie zur Anomalieerkennung, ist auch möglich. Genauer dazu in Kapitel 3. Die Simulation kann nicht ein konkretes CPPS exakt abbilden, sie muss sich jedoch funktional ähnlich zum CPPS verhalten. Dies dient dazu, dass beispielsweise Pre-Training möglich wird. Spezifisch bilden wir mit FLiPSi jedoch, wiederum abweichend von Balzereit et al., diskrete Fertigungssysteme ab. Hieraus ergibt sich die folgende Anforderung:

- (A4) Anwendbarkeit: Die Simulation muss in der Lage sein, die Funktionalität diskreter Fertigungssysteme in ihrer Essenz abzubilden.

Eine weitere Anforderung leitet sich aus A4 ab:

- (A5) Kausalität: Die Simulation muss in der Lage sein, komplexe Abhängigkeiten zwischen ursächlichen Fehlern und resultierendem Fehlerbild abzubilden.

In realen Systemen ist der genaue Zusammenhang zwischen Ursache und Wirkung, also die kausale Kette, oftmals unbekannt und nicht trivial. Dies muss auch in der Simulation dargestellt werden. Es sollen dementsprechend Effekte simuliert werden, die nachgelagert Einfluss auf vielfältige andere Aspekte der Simulation nehmen können.

Im Folgenden möchten wir eine Reihe von bestehenden Datensätzen und Simulationen mit Blick auf diese Anforderungen betrachten. Simulationen werden hierbei hinsichtlich ihrer Eignung für die Generierung von Daten zur weiteren Verwendung betrachtet. Dies ermöglicht die

Vergleichbarkeit mit realen Datensätzen. Die Ergebnisse dieser Betrachtungen sind in Tabelle 1 abgebildet.

Datensätze, die hingegen direkt an realen CPPS aufgenommen wurden, sind oftmals rar. Dies geht vor allem darauf zurück, dass es sich dabei oftmals um Firmengeheimnisse handelt. Selbst wenn diese Daten jedoch bereitstehen - für eine ausgiebige Ansammlungen dieser verweisen wir auf [17] - eignen sich solche Systeme nur begrenzt für die Anwendungen mit ML. So sind Fehlerfälle selten, was sie für das Anlernen von Modellen zur Anomalieerkennung weniger geeignet macht. Auch ist der betrachtete Zeithorizont oftmals nicht ausreichend, die Datenmenge ist begrenzt. Des Weiteren lassen sich Probleme der Planung und Rekonfiguration an realen Anlagen schwerlich testen: Die Kosten für das Abändern des realen CPPS sind viel zu hoch. Ein Übersicht bestehender Datensätze und ihrer Anwendungsgebiete wurden vom Fraunhofer IPT erstellt [17].

Einige bestehende Datensätze basieren auf der funktionalen Modellierung von bestehenden Cyber-Physischen Systemen (CPS). Ein gutes Beispiel hierfür ist die Windturbinen Benchmark von Simani et al. [5]. Hier wird ein konkretes Problem zur Fehlererkennung in Windturbinen gelöst. Es wird die Fähigkeit der Algorithmen aufgezeigt. Da es sich jedoch um ein konkretes System handelt bleibt die Verallgemeinerungsfähigkeit beschränkt. Ein weiteres klassisches Benchmarkproblem ist der Tennessee Eastman Process (TEP) [14]. Er beschreibt einen Herstellungsprozess aus der Verfahrenstechnik. Die vorab beschriebene Benchmark von Balzereit et al. basiert auf dem TEP, erweitert diesen jedoch um mehr Konfigurationen und hybride Signale. Damit ist er jedoch für unseren Fall eines CPPS für die diskrete Fertigung nicht anwendbar. Außerdem handelt es sich um eine reine Benchmark, keine allgemein verwendbare Simulation.

Tabelle 1: Bewertung verschiedener Datenquellen

● = erfüllt ○ = teilweise erfüllt ○ = nicht erfüllt

	A1 Hybrid	A2 Fehler	A3 Komplexität	A4 Anwendbarkeit	A5 Kausalität
Echte Datensätze	●	○	○	○	●
Funktionale Simulationen	●	●	●	○	●
Kommerzielle Software	●	●	●	●	○

Alternativen stellen kommerzielle Softwareprodukte zum Abbilden von Produktionssystemen dar. Beispiele hierfür sind FlexSim und Plant Simulation [13]. Plant Simulation basiert auf einer klassischen diskreten Eventsimulation (DES). Derartige Softwareprodukte haben aus Sicht der Forschung das Problem, dass ihre Funktionsweise nicht offengelegt wird, sondern Geschäftsgeheimnis ist. Das macht ihre offene Verwendung in einem kollektiv erweiterten Rahmen nur eingeschränkt möglich. Der Fokus ihrer Anwendung liegt außerdem klar auf der Abbildung von Logistikketten: Die Funktionalität der einzelnen Maschinen bleibt abstrakt.

In jüngster Zeit gibt es erste Ansätze - wie auch bei unserer Lösung - Game Engines zur Erstellung von Simulationen für CPPS zu verwenden. Eine solche Simulation wird von Müller et al als Proof of Concept vorgestellt [6]. Andere Simulationen wie die von Boccaccio et al [16] haben primär die Darstellung im Fokus. Eine gute Übersicht der bisherigen Arbeiten liefern Ottogalli et al. in ihrer Vorstellung eines Frameworks zur Einbindung von Simulatoren [8]. Da die Eignung dieser Ansätze als Datenquelle jedoch noch unklar ist, bleiben sie in Tabelle 1 unberücksichtigt.

3. Lösungsansätze durch FiPSi

Die Simulationsumgebung FiPSi kann verwendet werden, um Datensätze für CPPS zu erstellen. FiPSi wurde innerhalb der Game Engine Unity erstellt und nutzt die eingebaute Physik-Engine, um physikalisch realistische Szenarien zu simulieren. Durch die Physik-Engine ist das Verhalten von Körpern zueinander bereits implementiert und kann nach Belieben angepasst werden. So ist z.B. die Gravitation bereits fester Bestandteil der Simulationsumgebung.

Innerhalb von FiPSi können Module von CPPS erstellt werden. Die Erstellung von Modulen besteht aus zwei Komponenten - der dreidimensionalen Visualisierung und der programmierbaren Funktionalität. Es besteht die Möglichkeit vorgefertigte Visualisierungen aus externen Quellen einzubinden oder diese von Grund auf eigens mit Unity zu erstellen. Die Funktionalität wird in der Skriptsprache von Unity programmiert und muss letztlich noch mit der Visualisierung verknüpft werden, um ein vollwertiges Modul zu ergeben.

Auch ist es möglich, CAD-Daten zu importieren. Das Verhalten der Module kann frei angepasst werden, um die Schritte in einem realistischen Produktionsprozess darzustellen. Die Module können anschließend miteinander kombiniert werden, um eine Produktionskette zu simulieren. Die Interaktion der Module mit dem Produkt, beziehungsweise die Interaktionen der Module und Produkte untereinander, werden dabei von Unity gehandhabt.

Der Vorteil der Game Engine liegt, neben der eingebauten Physik-Engine, auch in der Adaptierbarkeit der Simulation. Durch die Verknüpfung mit einer Programmiersprache (standardmäßig C#) können Module, Produkte und deren Verhalten beliebig angepasst werden. Ferner

findet nicht nur die Simulation in einer 3D Umgebung statt, sondern auch die Gestaltung der Module sowie das Erstellen der Anlagen wird in 3D vorgenommen. So erlaubt FliPSi auch die Verknüpfung zwischen Design und funktionaler Simulation.

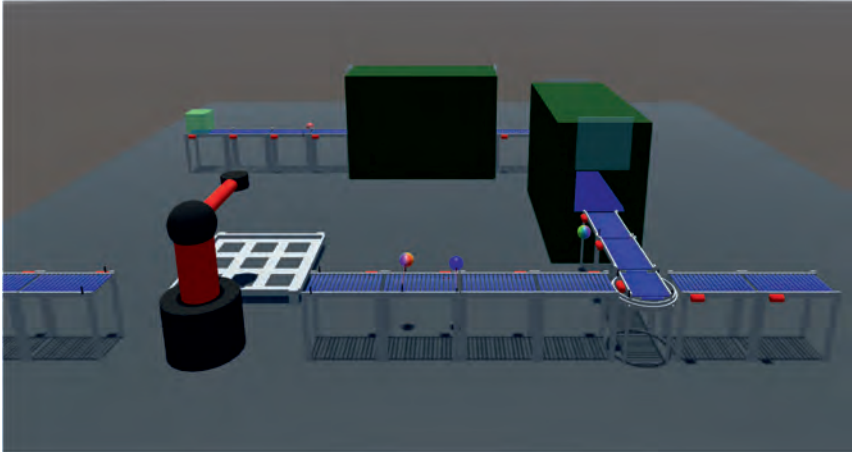


Bild 1: Eine Produktionsanlage in FliPSi. Zu sehen ist eine Kombination aus verschiedenen Produktionsmodulen, verbunden durch Förderbänder. Des Weiteren wurden der Anlage ein Lager und Sensoren angefügt.

Durch die oben beschriebenen Eigenschaften ist FliPSi in der Lage, Lösungsansätze für die in (2) beschriebenen Probleme zu liefern. Durch die Kombinierbarkeit der Module, wie in etwa Bohr-, Fräs-, oder Lackiermodule, Rollbahnen, Roboter Greifarme und Lager, können modulare CPPS manuell zusammengestellt oder automatisch erzeugt werden. Hierdurch wird die Varianz der Anwendungsfälle erhöht. Diese Modularität bietet die Möglichkeit, große, umfangreiche und vielfältige Datensätze, beispielsweise durch die Ausstattung von CPPS mit verschiedenster Sensorik, anzufertigen. Durch die hier beschriebene Anpassbarkeit der Simulation erfüllen Datensätze, welche mit FliPSi erstellt wurden, die Anforderung (A1), da es möglich ist sowohl kontinuierliche als auch diskrete Größen zu simulieren und aufzuzeichnen. Ebenso wird die Anforderung (A3) erfüllt, da Systeme von verschiedener Komplexität erstellt werden können, sowie die Anforderung (A4), da FliPSi in der Lage ist die Funktionalität verschiedener Fertigungsanlagen modellhaft darzustellen.

Durch die einfache Adaptierbarkeit der Simulation können gezielt Fehlerquellen, wie das Aussetzen einer Rollbahn, Fehlbohrungen im Bohrmodul oder gar ein gesamter Stromausfall, eingebaut werden, sodass eine Simulation von anormalen Betriebszuständen in beliebiger Häufigkeit möglich wird. Auch können verschiedene Arten von Fehlern oder Anomalien kombiniert werden. So ist es z.B. möglich, einen Ausfall eines elektrischen Motors und das Verkeilen eines Produktes auf einem Rollenband gleichzeitig zu simulieren und so realistische Datensätze auch für kombinierte Fehlerzustände zu erstellen. Dies erfüllt die Anforderung (A2).

Die Adaptierbarkeit der Simulation erlaubt nicht nur das Erstellen von Datensätzen mit Fehlerzuständen, sondern auch von Datensätzen im Falle von Anpassungen oder Veränderungen an einem CPPS, z.B. wenn eine Maschine ausgetauscht wird oder ein Produkt verändert werden soll. FliPSi kann verwendet werden, um die Auswirkungen solcher Veränderungen zu simulieren. Dies erlaubt auch die Erstellung von Datensätzen mit einer hohen Variabilität in Modulen und Produkten. Die Echtzeit-Visualisierung der Umgebung ermöglicht das Erkennen von direkten Veränderungen im Produktionssystem, sodass mögliche Auswirkungen in 'erwünscht' oder 'unerwünscht' charakterisiert und behandelt werden können. Dadurch wird auch die Anforderung (A5) erfüllt, da kausale Zusammenhänge zwischen Fehlern und deren Ursachen innerhalb der Simulation nachvollzogen werden können und somit auch in den Daten abgebildet werden können. Zusätzlich können durch die Visualisierung mögliche Fehler im Skript effizient nachvollzogen und behoben werden.

Durch die hier aufgeführten Eigenschaften ermöglicht FliPSi das systematische Generieren von Datensätzen mit veränderlichen Features für das (Pre-)Training von ML Modellen im Bereich von CPPS.

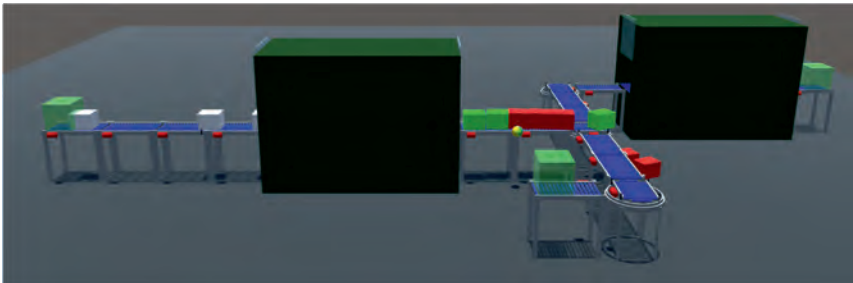


Bild 2: Ein Fehlerfall in einer Produktionsanlage in FliPSi. Es ist schnell ersichtlich, dass der Fehler durch die unterschiedlichen Geschwindigkeiten der Rollenbahnen verursacht wurde.

4. Diskussion

FliPSi fungiert als die kanonische Erweiterung der synthetischen Datengenerierung für CPPS. Nachteile gängiger CPPS Datensätze, wie in etwa die geringe Anzahl von Anomalien und Fehlern in den Daten, können hier durch gezielte Fehlergenerierung ausgemerzt werden. Die deutliche Erhöhung der Varianz in den generierten Datensätzen birgt implizit einen weiteren Vorteil dieses Ansatzes in sich. FliPSi erhöht die Vielfalt der Anwendungsfälle durch die Möglichkeit, Module der CPPS im Simulationsframework dynamisch zusammenzustellen.

Die gesteigerte Realitätsnähe des physikalischen Simulationsframeworks bietet zusätzlich eine Möglichkeit, unvorhersehbare Umstände zu generieren. Diese werden in der Gaming Engine präzise visualisiert, sodass Problemlösungen und das Verständnis modularer Zusammenhänge deutlich erleichtert werden. Der direkte Eingriff in die Simulation, als flexible Interaktivität, hat enorme Auswirkung auf die Deutung der generierten Daten.

Die Kehrseite der Generierung facettenreicher Datensätze stellt die Erstellung und Verknüpfung aufwändiger Module dar. Besonders in komplexen Systemen können sich Fehlerquellen einschleichen, die jedoch dank obig angeführter Visualisierung erleichtert behoben werden können. Deshalb stellt dies noch keine Grenze sondern lediglich eine Hürde für FliPSi dar.

Der aktuelle Stand von FliPSi basiert auf einer begrenzten Anzahl von Modulen. Die simulierten Anlagen stützen sich vorerst nicht auf reale Fabriken, sodass der Realitätsgrad noch angenähert werden kann. Ebenfalls werden Bearbeitungsvorgänge in der Simulation nicht konkret ausgeführt, sondern durch das Ersetzen des Produktes durch seine verarbeitete Version simuliert. Hier ist eine weitere Annäherung an die Realität möglich.

Die aufgeführten Möglichkeiten von FliPSi weisen auf deren enormes Potential hin und bilden eine richtungsweisende Basis für die Zukunft von CPPS.

5. Schlussfolgerung und Ausblick

In diesem Artikel wurde die Simulationsumgebung FliPSi vorgestellt, mit der realitätsnahe CPPS-Daten erzeugt werden können. Vor allem die vorherrschenden Herausforderungen, wie z.B. fehlende Variabilität von CPPS, können somit umgangen werden. Zusammenfassend kann dokumentiert werden, dass den evaluierten Defiziten (FF1) mit Hilfe von künstlich aus DES generierten Datensätzen entgegengewirkt werden kann. FliPSi ermöglicht dynamische Variationen (iii), die die Möglichkeit aufweisen, systeminternes Fehlverhalten zu erzeugen (ii) und den Algorithmus somit generalisiert (i) trainieren zu können. Die für FliPSi genutzte Gaming-Engine Unity zeigt, dass sich Gaming-Engines, aufgrund der physikalischen Eigenschaften, der Freiheiten in Bezug auf die grafische Modulerstellung und die Datengenerierung gut eignen (FF2). Die Möglichkeit, künstliche realistische Datensätze für das (Pre-)Training zu

erstellen, leitet sich aus den modular erstellbaren physikalischen Simulationsmodellen ab. Diese können variabel aus den vorhandenen CPPS-Modulen zusammengesetzt werden. Hierbei gibt es die Möglichkeit, das Modulverhalten flexibel einzustellen und beliebiges Fehlverhalten, wie Stromausfälle o.ä., zu adaptieren, was die Generierung realitätsnaher Daten ermöglicht (FF3).

Zukünftig eröffnet FliPSi eine Möglichkeit für jegliche Projekte im Bereich von CPPS, hohe Mengen variabler Systemdaten zu simulieren, um Forschungslücken in Bereichen wie KI und Automatisierung schließen zu können. Beispielhaft ist hier der Forschungsbereich der Diagnose, in der Fehlerverteilungen, die in Realität selten auftreten, von hoher Wichtigkeit sind. Weiterhin soll FliPSi zukünftig für Interessierte auf der Website des Instituts zur Verfügung gestellt werden.

Referenzen

[1] J. J. Downs and E. F. Vogel, "A plant-wide industrial process control problem," *Computers & Chemical Engineering*, vol. 17, no. 3, pp. 245–255, Mar. 1993, doi: 10.1016/0098-1354(93)80018-i.

[2] B. Heiming and J. Lunze, "Definition of the three-tank benchmark problem for controller reconfiguration," Aug. 1999. Accessed: Mar. 13, 2022. [Online]. Available: <http://dx.doi.org/10.23919/ecc.1999.7099963>

[3] M. Bartyś, R. Patton, M. Syfert, S. de las Heras, and J. Quevedo, "Introduction to the DAMADICS actuator FDI benchmark study," *Control Engineering Practice*, vol. 14, no. 6, pp. 577–596, Jun. 2006, doi: 10.1016/j.conengprac.2005.06.015.

[4] K. Balzereit, A. Diedrich, J. Ginster, S. Windmann, and O. Niggemann, "An ensemble of benchmarks for the evaluation of AI methods for fault handling in CPPS," Jul. 2021. Accessed: Mar. 13, 2022. [Online]. Available: <http://dx.doi.org/10.1109/in-din45523.2021.9557516>

[5] S. Simani, S. Farsoni, and P. Castaldi, "Fault diagnosis of a wind turbine benchmark via identified fuzzy models," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 6, pp. 3775–3782, Jun. 2015, doi: 10.1109/tie.2014.2364548.

- [6] T. Müller, N. Jazdi, J.-P. Schmidt, and M. Weyrich, "Cyber-physical production systems: Enhancement with a self-organized reconfiguration management," *Procedia CIRP*, vol. 99, pp. 549–554, 2021, doi: 10.1016/j.procir.2021.03.075.
- [7] A. W. Colombo, S. Karnouskos, O. Kaynak, Y. Shi, and S. Yin, "Industrial cyberphysical systems: A backbone of the fourth industrial revolution," *IEEE Industrial Electronics Magazine*, vol. 11, no. 1, pp. 6–16, Mar. 2017, doi: 10.1109/mie.2017.2648857.
- [8] Ottogalli, Rosquete, Amundarain, Aguinaga, and Borro, "Flexible framework to model industry 4.0 Processes for virtual simulators," *Applied Sciences*, vol. 9, no. 23, p. 4983, Nov. 2019, doi: 10.3390/app9234983.
- [9] S. Zeiml, K. Altendorfer, T. Felberbauer, and J. Nurgazina, "Simulation based forecast data generation and evaluation of forecast error measures," Dec. 2019. Accessed: Mar. 13, 2022. [Online]. Available: <http://dx.doi.org/10.1109/wsc40007.2019.9004876>
- [10] A. Villalonga et al., "A decision-making framework for dynamic scheduling of cyber-physical production systems based on digital twins," *Annual Reviews in Control*, vol. 51, pp. 357–373, 2021, doi: 10.1016/j.arcontrol.2021.04.008.
- [11] T. Tomiyama and F. Moyen, "Resilient architecture for cyber-physical production systems," *CIRP Annals*, vol. 67, no. 1, pp. 161–164, 2018, doi: 10.1016/j.cirp.2018.04.021.
- [12] E. A. Lee, "Cyber physical systems: Design challenges," May 2008. Accessed: Apr. 22, 2022. [Online]. Available: <http://dx.doi.org/10.1109/isorc.2008.25>
- [13] S. Bangsow, *Tecnomatix plant simulation: Modeling and programming by means of examples*. Springer Nature, 2020.
- [14] A. Bathelt, N. L. Ricker, and M. Jelali, "Revision of the Tennessee eastman process model," *IFAC-PapersOnLine*, vol. 48, no. 8, pp. 309–314, 2015, doi: 10.1016/j.ifa-col.2015.08.199.

[15] Hilding Elmqvist, "Modelica — A unified object-oriented language for physical systems modeling," *Simulation Practice and Theory*, vol. 5, no. 6, p. p32, Aug. 1997, doi: 10.1016/S0928-4869(97)84257-7.

[16] A. Boccaccio *et al.*, "Exploiting Augmented Reality to Display Technical Information on Industry 4.0 P&ID," in *Advances on Mechanics, Design Engineering and Manufacturing II*, Cham: Springer International Publishing, 2019, pp. 282–291. Available: http://dx.doi.org/10.1007/978-3-030-12346-8_28

Anwendungsoptimierte Modelladaption des Digitalen Zwillings eines modularen Produktionssystems während der Betriebsphase

D. Dittler, M.Sc., **T. Müller**, M.Sc., **V. Stegmaier**, M.Sc.,
Dr.-Ing. **N. Jazdi**, Univ.-Prof. Dr.-Ing. **M. Weyrich**,
Universität Stuttgart, Institut für Automatisierungstechnik und
Softwaresysteme, Stuttgart

Kurzfassung

Während der Engineering- und Betriebsphase modularer Produktionssysteme entstehen zunehmend Modelle. Der Digitale Zwilling kann diese Modelle für Anwendungen wie die virtuelle Inbetriebnahme, Optimierung, Prognose oder betriebsparallele Simulation zur Verfügung stellen. Durch unvorhersehbare Veränderungen in der Betriebsphase verschlechtert sich die Realitätsnähe der Modelle des Digitalen Zwillings, bspw. aufgrund von Anforderungsänderungen, neuen Anwendungen oder sich ändernden Umwelteinflüssen wie etwa Temperatur- oder Druckschwankungen und eine Modelladaption wird notwendig.

Die Adaption der Modelle des Digitalen Zwillings weist einige Herausforderungen auf, die in diesem Beitrag aufgezeigt werden. Um diesen Herausforderungen zu begegnen und dem Adaptionsaufwand während der Betriebsphase entgegenzuwirken, wird ein Konzept präsentiert, welches es dem Digitalen Zwilling ermöglicht, die Abweichung zur Realität zu erkennen und sich durch eine anschließende Modelladaption realitätsnah zu halten.

1. Einleitung

Globalisierung, volatile Märkte und kürzere Innovations- und Produktlebenszyklen fordern immer flexiblere und wandlungsfähigere Anlagen [1, 2]. Mit dem Einzug der Industrie 4.0 und der damit verbundenen erhöhten Digitalisierung und Vernetzung können Daten, Dienste und Funktionen dort gehalten und genutzt werden, wo diese den meisten Nutzen bringen [3]. Dies resultiert in dem Aufbrechen der klassischen Automatisierungspyramide, welche durch modulare, dezentrale, organisierte Systeme ersetzt wird [3], um die geforderte Flexibilität und Wandlungsfähigkeit zu ermöglichen. Dies bietet einerseits viele Potentiale für die zukünftige Automatisierungstechnik, auf der anderen Seite ergeben sich aber auch neue Herausforderungen. Die zunehmende Vielfalt der verteilten Daten, Dienste und Funktionalität in denen Abhängigkeiten und Wechselwirkungen existieren, führt zur ansteigenden Komplexität [4]. Ein wesentliches Ziel der Anlagenbetreiber ist es, deren Produktionssysteme

mit möglichst hoher Verfügbarkeit zu betreiben [5]. Insbesondere Anwendungen wie prädiktive Wartung, Optimierung, Qualitätsvorhersagen und Rekonfiguration bekommen in der Betriebsphase einen immer höheren Stellenwert [6–8]. Zur Realisierung derartiger Anwendungen werden Algorithmen aus den Bereichen künstliche Intelligenz oder mathematische Optimierung genutzt. Diese benötigen Informationen, welche im Zuge der oben genannten Digitalisierung zunehmend in Form von Modellen abgebildet werden [9]. Das Konzept des Digitalen Zwillings realisiert eine solche modell-basierte virtuelle Abbildung der realen Welt in die Informationswelt [10] und stellt dadurch einen bedeutenden Fortschritt für die Modellierungs-, Simulations- und Optimierungstechnologien dar [11]. Daher bietet der Digitale Zwilling die Möglichkeit unter anderem die oben genannten Anwendungen, insbesondere basierend auf der Durchführung simulativer Experimente, zu adressieren. Dabei kann der Digitale Zwilling prinzipiell für unterschiedliche Anwendungen genutzt werden, ohne dabei das reale Produktionssystem zu beeinflussen.

In der Entwicklungsphase einer Anlage (siehe Abbildung 1) werden hierzu verschiedene Modelle (z.B. Verhaltens-, Funktions-, Strukturmodelle, etc.) für die vorhergesehenen Anwendungs- und Betriebsszenarien entwickelt. Diese unterschiedlichen Modelle können durch einen Digitalen Zwilling in Relation gesetzt und bspw. für eine durchgängige virtuelle Inbetriebnahme genutzt werden.

Zu beachten ist, dass die Modelle für spezifische Anwendungen entwickelt werden und die dahingehenden geforderten Anforderungen erfüllen [12]. Um dabei ein rentables Aufwand-Nutzen-Verhältnis zu gewährleisten, wird die Realität anwendungsspezifisch abstrahiert, sodass die Realitätsnähe sinkt und sich eine sogenannte „reality gap“ [13] ergibt.

Diese „reality gap“ vergrößert sich aufgrund des Alterungsprozesses in der Betriebsphase (siehe Abbildung 1, rot), weshalb sich die Realitätsnähe des Digitalen Zwillings zur physischen Anlage verschlechtert. Dieser Alterungsprozess ist auf eine Vielzahl von möglichen Veränderungen zurückzuführen. Zum einen können während der Betriebsphase neue Anforderungen an bestehende Anwendungen gestellt werden, die in der Entwicklungsphase noch unbekannt waren. Aber auch neue Anwendungen mit den dazugehörigen Anforderungen können hinzukommen. Außerdem können Veränderungen am Asset (z.B. unberücksichtigter Umbau, Austausch, Verschleiß) und in der Umgebung des Assets (unberücksichtigte Einflussgrößen bspw. Temperatur-, Druckschwankungen) die „reality gap“ vergrößern [14]. Um diesem Alterungsprozess entgegenzuwirken, ist es notwendig, den Digitalen Zwilling in der Betriebsphase zu adaptieren, um eine anwendungsgerechte Realitätsnähe zu halten.

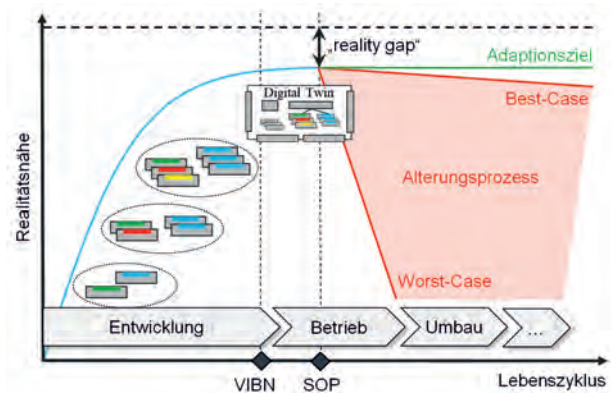


Bild 1 Die Realitätsnähe des Digitalen Zwillings

2. Grundlagen und Herausforderungen

In diesem Kapitel wird auf die heterogene Modelllandschaft sowie den Digitalen Zwilling eingegangen und die damit verbundenen Herausforderungen, die bei einer Adaption des Digitalen Zwillings während der Betriebsphase einhergehen, dargestellt.

2.1 Heterogene Modelle über den Anlagenlebenszyklus

Wie in der Einleitung beschrieben, werden zunehmend Modelle zur Abbildung von Informationen verwendet. Modelle sind dabei als „vereinfachte Nachbildung eines geplanten oder existierenden Systems mit seinen Prozessen in einem anderen begrifflichen oder gegenständlichen System“ [15] definiert. Dabei kann die Art der Nachbildung sehr heterogen sein. In [16] werden unterschiedliche Arten von Modellen (z.B. Struktur-, Funktions-, Verhaltensmodelle) aufgeführt. Unter anderem werden hier Verhaltensmodelle für Simulationen genutzt. Die Simulation, als virtuelles Experiment, ermöglicht es ein System besser zu verstehen und nimmt einen zunehmenden Stellenwert über alle Lebenszyklusphasen hinweg ein [17, 18]. Erste Simulationen umfassen z.B. die Erstellung von stationären Prozesssimulationen, um ein Anlagenprozesslayout zu entwickeln [12]. Anschließend wird die Automatisierung dieser Anlage geplant und modelliertes Verhalten von Steuerungskomponenten kann für den Test der korrekten Funktion (z.B. durch Software-in-the-Loop) herangezogen werden [12]. Des Weiteren kann die reale Steuerung, gekoppelt mit den Verhaltensmodellen der Anlage, durch Hardware-in-the-Loop getestet werden [12]. Aber auch während der Betriebsphase können Simulationen, z.B. für die Prognose oder Optimierung von Betriebsparametern eingesetzt werden [14]. Im Folgenden wird auf einige

Aspekte der heterogenen Modelllandschaft eingegangen, welche Herausforderungen für die Verwendung und Adaption der Modelle während der Betriebsphase darstellen.

Anlagen bestehen zunehmend aus einer Vielzahl unterschiedlicher Komponenten, welche von unterschiedlichen Komponentenlieferanten zur Verfügung gestellt werden [19]. Diese Komponenten umfassen durch die zunehmende Vernetzung mehrere mechatronische Disziplinen (Mechanik, Elektrotechnik und Softwaretechnik) [20]. Für die Erstellung der Modelle in den unterschiedlichen Domänen hat sich dabei eine Vielfalt an Simulations-Software auf dem Markt etabliert [12]. Aber auch die Modellierungsart kann zwischen White-, Grey- und Blackbox-Ansätzen variieren, um die jeweiligen Fragestellungen einer Anwendung zu beantworten [21–23]. Dabei sind die Blackbox-Ansätze nicht mit Black-Box-Modellen zu verwechseln, welche bspw. Komponentenlieferanten bei der Modellübergabe nutzen, um hier Know-How zu schützen [24]. Neben Verhaltensmodellen entstehen außerdem zunehmend Informationsmodelle. Diese können bspw. für den Austausch von Informationen zwischen den einzelnen Lebenszyklusphasen, für die nahtlose Integration von Geräten, für das Monitoring oder die Steuerung in übergeordneten Systemen verwendet werden. Dabei werden die Informationsmodelle für unterschiedliche Lebenszyklusabschnitte, verschiedene Anwendungsfälle oder für spezifische Ebenen der Automatisierungspyramide entwickelt [25]. Aufgrund der beschriebenen Heterogenität gehen die Autoren von einer erschwerten Wiederverwendbarkeit der Modelle in der Betriebsphase aus. Dies wird mit Zitaten wie etwa *„Von einer durchgängigen Nutzung von Simulation ist die industrielle Praxis allerdings noch weit entfernt“* [12] und *„Hier besteht die Herausforderung darin, die für den jeweiligen Anwendungsfall benötigten Informationen zu finden“* [25] in der Literatur untermauert. Diesen Herausforderungen kann das Konzept des intelligenten Digitalen Zwillings begegnen. Die Grundlagen sowie die Herausforderungen zu diesem werden im Folgenden beschrieben.

2.2 Der Digitale Zwilling

Der Digitale Zwilling findet in den unterschiedlichsten Bereichen, wie bspw. in der Produktion, im Gesundheitswesen oder bei Smart Cities, Anwendung [26, 27]. Das Konzept des Digitalen Zwillings, welches sich auf Ursprünge in den 2000er-Jahren zurückverfolgen lässt, hat ab 2016 einen deutlichen Anstieg an Veröffentlichungen gezeigt [28]. Dabei sind für den Begriff und das Konzept vielfältige Definitionen entstanden [29]. Die Autoren von [30] haben dazu eine Literaturrecherche durchgeführt und anschließend eine Definition für einen Digitalen Zwilling vorgeschlagen, welche die wesentlichen Aspekte der anderen Definitionen vereint. Diese Definition wird für diesen Beitrag herangezogen.

Den Grundbaustein eines solchen Digitalen Zwillings bilden Modelle und deren Relation zueinander, welche ein virtuelles Abbild eines physischen Assets darstellen. Die Modelle und ihre Relationen sind mit der Realität synchronisiert, sodass zu jeder Zeit das statische Abbild der Realität repliziert wird. Es muss eine aktive Datenerfassung von der Realität zum Digitalen Zwilling erfolgen, um auch dynamischen Vorgänge wahrnehmen zu können. Das Vorhandensein eines ausführbaren Modells ist notwendig, damit das statische Abbild um die Replikation des dynamischen Verhaltens der Realität bereichert wird. Um Dienste wie z.B. die Optimierung des Prozessablaufs, die automatische Generierung von Steuercodes für neu hinzukommende Maschinen im Rahmen von „Plug and Produce“ oder die prädiktive Wartung auf Basis der im Digitalen Zwilling gespeicherten Betriebsdaten während der Betriebsphase zu ermöglichen, ohne eine erhöhte und damit fehleranfällige Modellpflege durchführen zu müssen, muss der Digitale Zwilling mit Intelligenz angereichert werden.[30, 31]

Der Digitale Zwilling wird meist anwendungsspezifisch realisiert. „Zudem sind die heutigen Anwendungsfälle häufig isoliert voneinander, das heißt, eine Wiederverwendung von Modellen und Erfahrungen findet nur selten statt“ [12]. Bei einer notwendigen Modelladaption aufgrund der zuvor genannten unvorhergesehenen Veränderungen, wie etwa eine Veränderung am Asset oder aus Anwendungssicht, wird Systemwissen benötigt. Also das Wissen über Abhängigkeiten und Wechselwirkungen innerhalb der heterogenen Modelllandschaft sowie das Wissen über die einzelnen Modelle, um geeignete Modelle für die jeweilige Veränderung zu identifizieren. Das Modellwissen liegt hier meist beim Modellersteller wie z.B. dem Komponentenlieferanten. Daraus folgt, dass die manuelle Pflege, insbesondere bei Simulationsmodellen, zeit- und kostenintensiv sowie fehleranfällig ist [32]. Hier kann der intelligente Digitale Zwilling unterstützen, indem das Wissen und der Adaptionaufwand über intelligente Algorithmen zur Verfügung gestellt werden. Dieses Konzept ist jedoch Teil aktueller Forschungsaktivitäten und muss dahingehend noch ausgeprägt und konkretisiert werden. Daher wird die folgende Forschungsfrage gestellt: Wie kann der Digitale Zwilling um eine anwendungsoptimierte Modelladaption erweitert werden, um in einer heterogenen Modelllandschaft auf unvorhergesehene Veränderungen reagieren zu können, sodass die Realitätsnähe des Digitalen Zwillings während der Betriebsphase gehalten werden kann?

3. Konzept für die anwendungsoptimierte Modelladaption

Für die Beantwortung der Forschungsfrage wird in diesem Kapitel das Konzept vorgestellt. Zunächst wird in Abbildung 2 ein Überblick über notwendige Teilbereiche (grau) und den Fokus (blau) für diesen Beitrag gegeben.

Wie oben beschrieben kann die „reality gap“ sowohl durch veränderte Anforderungen einer Anwendung, einer neuen Anwendung als auch durch Veränderungen am Asset oder dessen Umgebung abdriften. Die Identifikation einer Abweichung z.B. durch die Veränderung der Umgebung (etwa eine neue Einflussgröße) stößt die *Modelladaption* an, welche eine optimierte Modellkonfiguration hinsichtlich des „magischen Dreiecks“ Zeit, Kosten und Qualität zur Verfügung stellt. Dadurch soll im Sinne eines „Model-as-a-Service“ Ansatzes die anwendungsoptimierte (anwendungsgerechte und optimierte) Modellkonfiguration automatisch zur Verfügung gestellt werden.

Eine bestehende *Anwendung* (z.B. Prognose, Optimierung oder betriebsparallele Simulation) wird durch eine *Aktive Modellkonfiguration* bedient. Über den Abgleich, der von einer Anwendung gelieferten Werte mit den Ist-Werten aus der *Aktiven Datenerfassung*, ist die *Erkennung der Abweichung* möglich. Sowohl Informationen über die Erkennung der Abweichung, als auch die Aktive Modellkonfiguration sowie die Prozessdaten der aktiven Datenerfassung werden der *Modelladaption* zur Verfügung gestellt. Damit die Modelladaption anwendungsoptimierte Modelle liefern kann, greift diese auf den vorhandenen *Modellpool* zu. Dabei besteht dieser Modellpool wie in Kapitel 2.1 beschrieben aus einer heterogenen Modelllandschaft. Für eine automatische Auswahl der Modelle, bedarf es daher einer Strukturierung dieser heterogenen Modelllandschaft. In der Literatur gibt es eine Vielzahl von Vorschlägen für diese Strukturierung. In [34] haben die Autoren eine Studie durchgeführt und anschließend eine vielversprechende vierdimensionale Strukturierung von Verhaltensmodellen beschrieben. Diese Struktur ermöglicht es Modelle über die Modellbreite, den Modellumfang, die Modelltiefe und die Verhaltensart anwendungsgerecht auszuwählen. Der Schwerpunkt der Autoren liegt auf der Klassifizierung der Modelltiefen, welche in fünf Ebenen, vom diskreten Verhalten (Modelltiefe 1) bis hin zum physikalisch-räumlichen Verhalten (Modelltiefe 5) eingeordnet werden.

Um dem Alterungsprozess zu begegnen, hat sich der Plan-Do-Check-Act-Zyklus, eine systematische Vorgehensweise für die kontinuierliche Verbesserung aus dem Lean-Management, in der Industrie etabliert. Dieser wird als Basis für die Ausprägung einer Methodik zur automatischen Modelladaption herangezogen und im Folgenden näher beschrieben. Dabei kann diese Methodik prinzipiell auch manuell angewendet werden, dies kann jedoch aufgrund der heterogenen Modelllandschaft, sowie der Anzahl an möglichen Lösungskonfigurationen kostenintensiv, nicht objektiv und fehleranfällig sein.

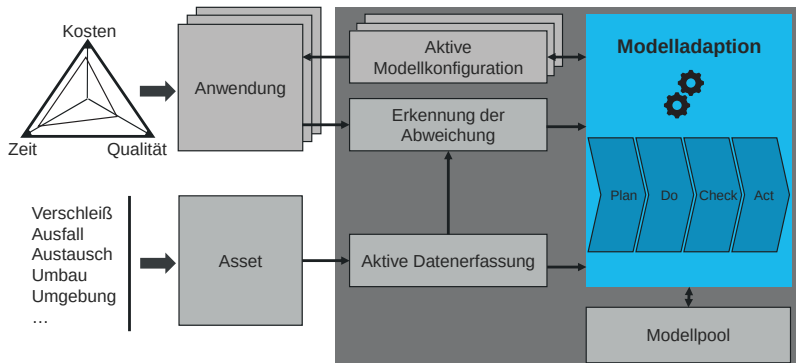


Bild 2: Überblick der anwendungsoptimierten Modelladaption (vereinfachte Darstellung)

Die vier Schritte des Plan-Do-Check-Act-Zyklus sind in Abbildung 3 dargestellt.

(1) Plan: Um eine automatisierte, anwendungsspezifische Modelladaption durchzuführen muss der Modelladaptionsansatz zum einen die Informationen, welche aus der Erkennung der Abweichung übergeben werden verarbeiten. Dabei werden die drei unterschiedlichen Trigger: Veränderung der Asset-Werte, Anforderungsänderung einer Anwendung oder neue Anwendung in der *Analyse Ist-Zustand* geprüft. Die Anforderungen der Anwendungen werden mit den Meta-Informationen der aktiven Modellkonfigurationen hinsichtlich der funktionalen Eignung abgeglichen, um einen *Adaptionsbedarf* zu ermitteln. Tritt eine Abweichung trotz erfolgreichem Mapping der Anwendungsanforderungen und der aktiven Modellkonfiguration auf, so handelt es sich stets um eine Veränderung am Asset oder dessen Umgebung. Kann keine der aktiven Modellkonfiguration, die identifizierte Abweichung auflösen, wird das *Adaptionsziel* festgelegt und der Do-Schritt, der im Folgenden beschrieben wird, angestoßen. Das Adaptionsziel setzt sich dabei aus der funktionalen Eignung sowie der Gewichtung des „magischen Dreiecks“ zusammen.

(2) Do: Im Do-Schritt werden geeignete Modellkonfigurationen generiert. Die Generierung neuer geeigneter Modellkonfigurationen basiert dabei auf einem zweistufigen Modelladaptionsansatz. Die erste Stufe untersucht nach einer minimal-invasiven Vorgehensweise, im Sinne der Aufwandsreduktion, ob eine aktuell aktive Modellkonfiguration über eine *Parameteranpassung* (Parametrierung) in eine geeignete Modellkonfiguration überführt werden kann, welche das Adaptionsziel erfüllt. Sollte dies nicht möglich sein, aktiviert sich die zweite Stufe, in dieser werden neue Modellkonfigurationen aus dem Modellpool in Form von ausführbaren Simulationsmodellen orchestriert. Dabei werden zunächst alle

funktional geeigneten Modellkonfigurationen im Sinne einer *Strukturanpassung* gebildet. Dies beinhaltet bspw. die Auswahl eines Modells in einer erhöhten Modelltiefe, welche eine Detaillierung des Modells durch eine fein-granularere Parametrisierung darstellen kann.

(3) Check: Im Check-Schritt werden die gefundenen Modellkonfigurationen hinsichtlich des „magischen Dreiecks“ verglichen und anschließend die geeignetste Modellkonfiguration ausgewählt. Dazu werden die einzelnen Modellkonfigurationen automatisiert simuliert. Basierend auf den Informationen über die Modelltiefe, Modellumfang und Modellbreite aus dem Modellpool kann eine simulationsbasierte Evaluierung hinsichtlich Zeit, Kosten und Qualität für eine entsprechende Anwendung durchgeführt werden. Der Vergleich der von den Modellkonfigurationen gelieferten Werte mit den Prozessdaten aus der aktuellen Datenerfassung ermöglicht dabei die Bewertung der Qualität hinsichtlich der Realitätsnähe. Wird keine Modellkonfiguration gefunden wird der Do-Schritt erneut angestoßen. Ist eine geeignete Modellkonfiguration identifiziert, wird der Act-Schritt angestoßen.

(4) Act: Im Act-Schritt wird die geeignetste Modellkonfiguration aktiv geschaltet und auf den nächsten Trigger für einen erneuten Durchlauf des Plan-Do-Check-Act-Zyklus gewartet. Durch die iterative Vorgehensweise mit den Feedbackschleifen wird sichergestellt, dass die Modellkonfigurationen so genau wie nötig und nicht so genau wie möglich zur Verfügung gestellt werden, was dem Grundgedanken des Plan-Do-Check-Act-Zyklus, also der kontinuierlichen Verbesserung entspricht.

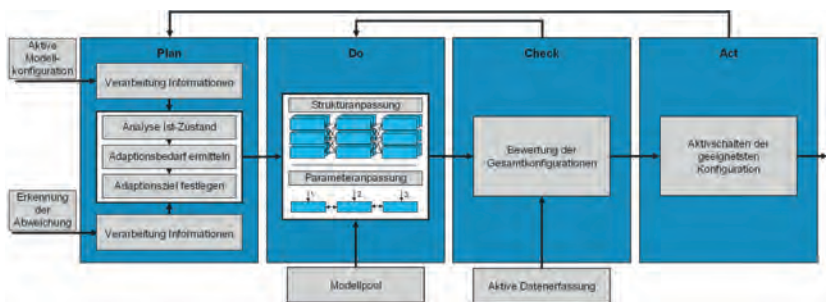


Bild 3: Konzept für die Modelladaptation

4. Anwendungsszenario

Das Konzept soll an dem modularen Produktionssystem (Abbildung 4, links) am Institut für Automatisierungstechnik und Softwaresysteme (IAS) evaluiert werden. Dieses besteht aus einem Lager, fünf Bearbeitungsmodulen und drei fahrerlosen Transportsystemen (FTS). Die FTS übernehmen, je nach Arbeitsplan der Produkte, den Transport zwischen den relevanten

Bearbeitungsmodulen. Diese sind wiederum mit Förderbändern ausgestattet, um die von den FTS angelieferten Produkte, an die jeweilige Bearbeitungsstation der Bearbeitungsmodule zu befördern.

Für das Anwendungsszenario wird ein Bearbeitungsmodul um eine Bearbeitungsstation in Form eines physischen Greifsystems zur Handhabung erweitert. Die Anwendung, die den Digitalen Zwilling nutzt, ist eine betriebsparallele Simulation, um das reale Anlagenverhalten abzubilden und dem Bediener eine Visualisierung mit Hilfe eines HMI zur Verfügung zu stellen. Damit kann bspw. die Produktivität der Anlage überwacht werden. Konkret muss für diesen Anwendungsfall eine Kopplung der Modelle der Anlagenkomponenten (FTS, Bearbeitungsmodule und deren Bearbeitungsstationen) erfolgen, sodass eine aktive Modellkonfiguration zur Abbildung des taktgebunden Materialflusses vorliegt. Für jede der Anlagenkomponenten gibt es unterschiedliche Modelle, die im Modellpool vorliegen. Im Anwendungsszenario ist die Ausgangssituation derart spezifiziert, dass für die aktive Modellkonfiguration, Anlagenkomponenten in der Modelltiefe 2 gewählt wurden, was nach [34] einer zeit-diskreten Modellierung entspricht und mit Totzeitgliedern realisiert wurde. In Abbildung 4 (rechts) ist zur Veranschaulichung das Komponentenmodell Greifsystem in der Modelltiefe 2 dargestellt.

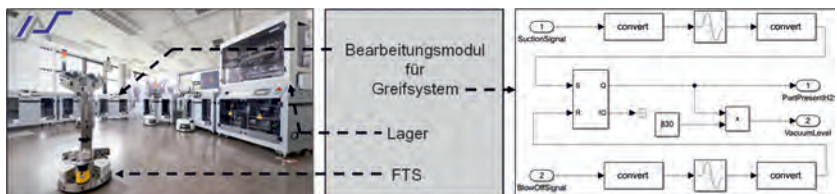


Bild 4: Modulares Produktionssystem (links), modelliertes Greifsystem in Modelltiefe 2 (rechts)

In diesem Szenario wird die Veränderung am Asset, im konkreten am Greifsystem, und dessen Umgebung als Trigger für die Modelladaption gewählt. Das Greifsystem kann vereinfacht mit zwei Eingängen (Ansaug- und Abblas-Signal) und zwei Ausgängen (Teilekontrolle, Unterdruck) dargestellt werden (siehe Abbildung 5). Bereits in Bezug auf diese einzelne Komponente der Gesamtanlage existiert eine Vielzahl von möglichen Veränderungen, die eine Abweichung vom Sollwert verursachen. Diese können durch die Verwendung von Modellen einer höheren Modelltiefe adressiert werden. Im vorhandenen Modellpool liegen hierbei für die Anlagenkomponente Greifsystem Komponenten-Modelle vor, welche sich hinsichtlich ihrer Modelltiefe wie folgt beschreiben lassen. Modelltiefe 2 und 3 der Komponenten-Modelle

wurden in separaten *Simulink*-Modellen realisiert. Die Dauer zwischen Ansaugen und Teilekontrolle t wird in Modelltiefe 2 in Form eines Totzeitglieds modelliert. Die Parameter der Modelltiefe 3 werden mit Hilfe einer approximierten Gleichung mit mathematischen Bausteinen realisiert. Dabei werden Modelltiefe 2 und 3 durch unidirektionale Informationsflüsse modelliert. Modelltiefe 4 wurde in *Simscape* (bidirektionale Informationsflüsse) umgesetzt. Die Komponenten des Greifsystems (Schlauch, Ejektor und Sauger) und deren physikalisches Verhalten wurden mit Hilfe von Differentialgleichungen abgebildet.

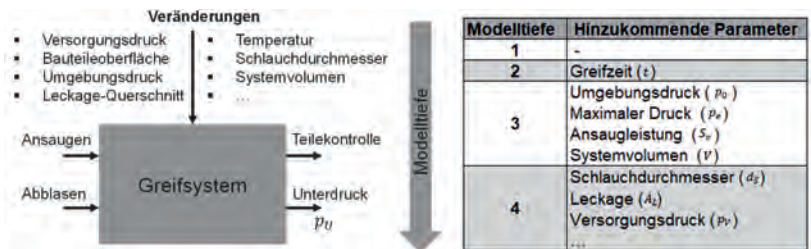


Bild 5: Vereinfachte Darstellung Greifsystem, Veränderungen und unterschiedliche Modelltiefen

Um die Vorgehensweise des Konzeptes zu evaluieren wurde ein Prototyp (siehe Abbildung 6) aufgebaut. Dazu wurde ein Greifsystem der Firma J. Schmalz GmbH bestehend aus vier Sauggreifern, Schlauchsystem, Ejektor und Drucksensor herangezogen. Die Ansteuerung der Eingänge (Ansaugen, Abblasen) und die Messung der Ausgänge (Teilekontrolle, Unterdruck) wurde mittels einer *Beckhoff CX9020* und mit dem Engineering-Tool *TwinCAT* realisiert. Der Vergleich der Messung (durchgängige Linien) mit den Simulationen in den unterschiedlichen Modelltiefen (gestrichelte und gepunktete Linien) zeigt, dass das virtuelle Verhalten der unterschiedlichen Modelltiefen prinzipiell mit dem realen Verhalten im Idealbetrieb übereinstimmt (siehe Abbildung 6). In Modelltiefe 2 zeigt sich vor allem bei der Betrachtung des Vakuumwerts eine Vereinfachung durch die zeit-diskrete Modellierung.

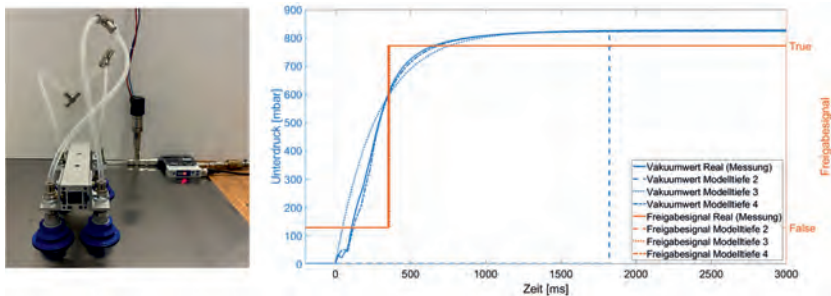


Bild 6: Versuchsaufbau (links), Modellwerte entsprechen realen Werten (rechts)

Anschließend wurden gezielte Veränderungen (p_V , A_L , S_V , siehe Abbildung 5) am Versuchsaufbau eingebracht, um „reality gaps“ zu erzeugen und damit die Modelladaption anzustoßen. Zur Übersichtlichkeit wird im Folgenden auf die Veränderung durch den schwankenden Versorgungsdruck p_V eingegangen. Diese Veränderung tritt in der Praxis bspw. durch erhöhte Verbräuche anderer Anlagenkomponenten oder Defekte im Druckluftsystem auf und kann zu einer erhöhten Greifzeit oder einem erfolglosen Greifen des Teils aufgrund des zu niedrigen Unterdrucks führen. In Abbildung 7 (links) sind die „reality gaps“ deutlich zu erkennen. Dabei entsteht eine „reality gap“ (1) zwischen dem realen (durchgängige Linie) und simulierten (gestrichelte Linien) Freigabesignal (orange) und eine „reality gap“ (2) zwischen dem gemessenen (durchgängige Linie) und simulierten (gestrichelte Linien) Unterdruck (blau).

Nach Erkennung der Abweichung wird der Plan-Schritt angestoßen. Da es sich bei diesem Experiment um eine gleichbleibende Anwendung (Betriebsparallele Simulation) handelt und keine neue Anforderung an die Anwendung durch den Nutzer vorgegeben wurde, ergibt die Ist-Analyse des Plan-Schritts, dass eine Veränderung am Greifsystem oder dessen Umgebung aufgetreten ist. Als Adaptionziel wird die Steigerung der Realitätsnähe (Qualität) vorgegeben. Eine solche Steigerung ist durch die Erhöhung der Modelltiefe möglich. Die Strukturierung der Modelle im Modellpool ermöglicht eine iterative Vorgehensweise für das Finden des geeigneten Modells im Do-Schritt. Ausgehend von der Modelltiefe 2 wird in der Strukturangepassung die Anwendung der Modelltiefe 3 geprüft. Modelltiefe 3 ist nach [34] als „kontinuierliches vereinfachtes physikalisches Verhalten“ definiert. In [35] wurde hierfür eine approximierende Gleichung mit dem Gültigkeitsbereich zwischen 700-740 mbar hergeleitet. Da die reale Messgröße von p_U jedoch 620 mbar (vgl. Abbildung 7) beträgt, verliert die Modelltiefe 3 in diesem Szenario die Modellgültigkeit und es wird auf die Modellstufe 4

umgeschaltet. In dieser Modelltiefe werden die Parameter p_V, A_L, d_S in der Parameteranpassung iterativ in sinnvollen Wertegrenzen variiert. Im Check-Schritt wird die Modelltiefe 4 mit den unterschiedlichen Parametersätzen ausgeführt und die Ergebnisse mit den Prozessdaten verglichen. Durch die Parameteranpassung von A_L oder d_S kann zwar die „reality gap“ (1) zum Freigabesignal für die Teilekontrolle geschlossen werden, jedoch lässt sich nur durch die Variation des Versorgungsdrucks p_V ein realistischer Wert finden, um beide „reality gaps“ zu schließen (siehe Abbildung 7, rechts). Somit wurde eine geeignete Modellkonfiguration (inkl. Parametersatz) gefunden. Diese wird im Act-Schritt aktiv geschaltet.

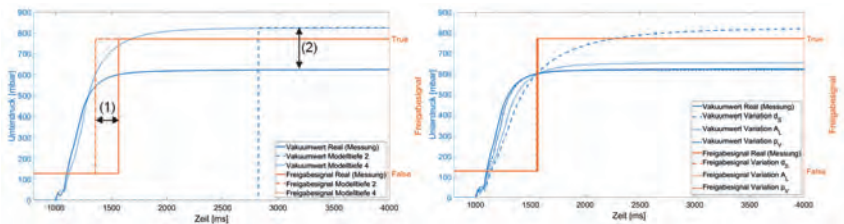


Bild 7: Erkennung der Abweichung (links), Parameteranpassung in Modelltiefe 4 (rechts)

Neben der Demonstration der prinzipiellen Machbarkeit konnten im Rahmen der Versuchsdurchführung folgende Erkenntnisse erschlossen werden. Die vorgeschlagene Verwendung des Modellpools stellt eine vielversprechende Möglichkeit dar, Verhaltensmodelle unterschiedlicher Komponenten zu strukturieren. Diese Strukturierung bildet eine gute Ausgangsbasis für eine automatisierte Modelladaption, jedoch wurde festgestellt, dass zusätzliche Informationen benötigt werden um dies vollumfänglich zu realisieren. So ist es notwendig auch die Anforderungen einer Anwendung zu kennen, um abgleichen zu können, ob eine zu untersuchende Modellkonfiguration die notwendigen Aspekte abbilden kann. Diese notwendigen Aspekte müssen dazu ebenfalls geeignet beschrieben werden. Hierzu müssen ebenso die Gültigkeitsbereiche der Modelle bekannt sein. Des Weiteren müssen Informationen über die einstellbaren Parameterbereiche für die jeweiligen Komponenten-Modelle vorhanden sein, um subsequent gültige Parametersätze einstellen und simulationsbasiert untersuchen zu können. Das für die automatisierte Modelladaption benötigte Wissen, vollumfänglich und in geeigneter Form, zur Verfügung zu stellen, ist eine wesentlich größere Herausforderung als den zugehörigen Algorithmus umzusetzen. In diesem Zusammenhang wurden Knowledge Graphen unter anderem in [33] als vielversprechende Lösung identifiziert, da sich diese durch eine hohe semantische Ausdrucksstärke auszeichnen und mit Heterogenität umgehen können.

5. Zusammenfassung und Ausblick

Der Beitrag stellt die Notwendigkeit einer Modelladaption der Modelle des Digitalen Zwillings während der Betriebsphase dar. Die heterogene Modelllandschaft sowie der Digitale Zwilling und die im Kontext der Modelladaption entstehenden Herausforderungen werden präsentiert. Die Modelladaption ist demnach kostenintensiv, nicht objektiv sowie fehleranfällig. Es bedarf daher einer Systematik, welche in einer automatisierten Modelladaption resultiert. Das Konzept des intelligenten Digitalen Zwillings weist dahingehend die notwendige Intelligenz auf, jedoch bedarf es eines Konzepts, das diese Intelligenz konkretisiert und ausprägt.

Daher wird ein neuartiges Konzept präsentiert, welches den Digitalen Zwilling um eine anwendungsoptimierte Modelladaption erweitert, um in einer heterogenen Modelllandschaft auf unvorhergesehene Veränderungen reagieren zu können. Dazu wird eine Methodik basierend auf dem Plan-Do-Check-Act-Zyklus vorgestellt. In dieser werden zunächst die Informationen der Erkennung der Abweichung verarbeitet sowie die Anforderungen der Anwendung mit den aktiven Modellkonfigurationen auf die funktionale Eignung geprüft, um einen Adaptionbedarf zu ermitteln. Der zweistufige Modelladaptionsansatz im Do-Schritt ist in die Parameter- und Strukturanpassung gegliedert, um auf Basis aktiver Modellkonfigurationen oder des strukturierten Modellpools anwendungsgerechte Modellkonfigurationen zu generieren. Diese werden im Check-Schritt nach den Kriterien Zeit, Kosten und Qualität bewertet. Die für die Anwendung geeignetste Modellkonfiguration wird dann im Act-Schritt aktiv geschaltet. Anschließend wird auf einen Trigger für einen erneuten Zyklusdurchlauf gewartet. Dadurch wird die Realitätsnähe des Digitalen Zwillings in der Betriebsphase gehalten. Abschließend wird die Methodik anhand eines prototypischen Versuchsaufbaus veranschaulicht.

Zukünftig werden das vorgestellte modulare Produktionssystem und das Greifsystem für die weitere Evaluierung eingesetzt. Hierzu wird das Konzept verfeinert, um die automatische Parameter- und Strukturanpassung zu ermöglichen. Vor allem die Modellierung des notwendigen Wissens und der funktionalen Anforderungen der Modelle nimmt einen wichtigen Stellenwert für den Model-as-a-Service Aspekt ein.

Literaturverzeichnis

- [1] E. Järvenpää, N. Siltala, and M. Lanz, "Formal resource and capability descriptions supporting rapid reconfiguration of assembly systems," in *2016 IEEE International symposium on assembly and manufacturing (ISAM)*, 2016, pp. 120–125.
- [2] T. Müller, N. Jazdi, and M. Weyrich, "Intelligentes Rekonfigurationsmanagement selbstorganisierter Produktionssysteme in der diskreten Fertigung," 2020.
- [3] J. Schlick, P. Stephan, and T. Greiner, "Kontext, Dienste und Cloud Computing," *atp*, vol. 55, no. 04, pp. 32–41, 2013.
- [4] C. Müller-Schloer, H. Schmeck, and T. Ungerer, "Organic Computing," *Inform. Spektrum*, vol. 35, no. 2, pp. 71–73, 2012.
- [5] T. Müller, S. Walth, N. Jazdi, and M. Weyrich, "Identification of Reconfiguration Demand and Generation of Alternative Configurations for Cyber-Physical Production Systems," in *Advances in Automotive Production Technology-Theory and Application*: Springer, 2021, pp. 63–70.
- [6] T. Müller, B. Lindemann, T. Jung, N. Jazdi, and M. Weyrich, "Enhancing an Intelligent Digital Twin with a Self-organized Reconfiguration Management based on Adaptive Process Models," *Procedia CIRP*, vol. 104, pp. 786–791, 2021.
- [7] B. Lindemann, N. Jazdi, and M. Weyrich, "Adaptive Quality Control for discrete large-scale Manufacturing Systems subjected to Disturbances," *25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, Vienna, Austria, pp. 379–386, 2020.
- [8] P. Aivaliotis, K. Georgoulas, and G. Chrysosouris, "The use of Digital Twin for predictive maintenance in manufacturing," *International Journal of Computer Integrated Manufacturing*, vol. 32, no. 11, pp. 1067–1080, 2019.
- [9] R. Rosen, G. von Wichert, G. Lo, and K. D. Bettenhausen, "About The Importance of Autonomy and Digital Twins for the Future of Manufacturing," *IFAC (International Federation of Automatic Control)*, 2015.
- [10] M. Ehrlich, L. Wisniewski, H. Trsek, and J. Jasperneite, "Modelling and automatic mapping of cyber security requirements for industrial applications: Survey, problem exposition, and research focus," *14th ieee international workshop on factory communication systems (wfcs)*, pp. 1–9, 2018.
- [11] S. Weyer, T. Meyer, M. Ohmer, D. Gorecky, and D. Zühlke, "Future modeling and simulation of CPS-based factories: an example from the automotive industry," *Ifac-Papersonline*, vol. 49, no. 31, pp. 97–102, 2016.

- [12] D. Labisch, C. Leingang, O. Lorenz, M. Oppelt, B. Pfeiffer, F. Pohmer, "Evolution eines Digital Twin am Beispiel einer Ethylen-Anlage," *atp*, 6-7, pp. 70–84, 2019.
- [13] A. Meier, S. Carroccio, R. Dornberger, and T. Hanne, "Discussing the Reality Gap by Comparing Physics Engines in Kilobot Simulations," *Journal of Robotics and Control (JRC)*, vol. 2, no. 5, pp. 441–447, 2021.
- [14] C. Härle, M. Barth, and A. Fay, "Betriebsparallele Adaption einer Co-Simulation für diskrete Fertigungsanlagen," in *Automation 2021*: VDI Verlag, pp. 517–534, 2021.
- [15] *Virtuelle Inbetriebnahme: Modellarten und Glossar*, 3693 Blatt 1, Verein Deutscher Ingenieure e.V.; Verband der Elektrotechnik, Elektronik und Informationstechnik, Berlin, Aug. 2016.
- [16] F. Oestersötebier, "Modellbasierter Entwurf intelligenter mechatronischer Systeme mithilfe semantischer Technologien," *Doktorarbeit. Paderborn, Universität Paderborn, 2018*.
- [17] M. Oppelt, M. Barth, and L. Urbas, "The Role of Simulation within the Life-Cycle of a Process Plant - Results of a global online survey," 2015.
- [18] R. Rosen, et al., "VDI - Statusreport Simulation und digitaler Zwilling im Anlagenlebenszyklus," 2020. [Online]. Available: https://www.researchgate.net/publication/339238064_Simulation_und_digitaler_Zwilling_im_Anlagenlebenszyklus
- [19] C. Scheifele, "Plattform zur Echtzeit-Co-Simulation für die virtuelle Inbetriebnahme," Stuttgart: Fraunhofer Verlag, 2019.
- [20] *Entwicklung mechatronischer und cyber-physischer Systeme*, 2206, Verein Deutscher Ingenieure e.V.; Verband der Elektrotechnik, Elektronik und Informationstechnik, Berlin, Nov. 2021.
- [21] D. Schröder, and M. Buss, *Intelligente Verfahren*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2017.
- [22] N. Dervilis, *Special Topics in Structural Dynamics, Volume 6*. Cham: Springer International Publishing, 2017.
- [23] B. Lindemann, T. Müller, H. Vietz, N. Jazdi, and M. Weyrich, "A survey on long short-term memory networks for time series prediction," *Procedia CIRP*, vol. 99, pp. 650–655, 2021.
- [24] C. Härle, M. Barth, and A. Fay, "Simulationsmodellgenerierung im modularen Maschinen- und Anlagenbau," *atp*, vol. 62, no. 9, pp. 72–79, 2020.
- [25] A. Schüller, D. Großmann, M. Birkenkamp, "Informationsmodelle im Anlagenlebenszyklus," *atp*, no. 2, pp. 50–59, 2021.

- [26] A. Fuller, Z. Fan, C. Day, and C. Barlow, "Digital Twin: Enabling Technologies, Challenges and Open Research," *IEEE Access*, vol. 8, pp. 108952–108971, 2020.
- [27] F. Tao, H. Zhang, A. Liu, and A. Y. C. Nee, "Digital Twin in Industry: State-of-the-Art," *IEEE Trans. Ind. Inf.*, vol. 15, no. 4, pp. 2405–2415, 2019.
- [28] M. Sjarov, T. Lechler, J. Fuchs, M. Brossog, A. Selmaier, F. Faltus, T. Donhauser, J. Franke, *The Digital Twin Concept in Industry – A Review and Systematization*. Piscataway, NJ: 25th IEEE International Conference, 2020.
- [29] E. Negri, L. Fumagalli, and M. Macchi, "A Review of the Roles of Digital Twin in CPS-based Production Systems," *Procedia Manufacturing*, vol. 11, pp. 939–948, 2017.
- [30] B. Ashtari Talkhestani *et al.*, "An architecture of an Intelligent Digital Twin in a Cyber-Physical Production System," *at - Automatisierungstechnik*, vol. 67, no. 9, pp. 762–782, 2019.
- [31] N. Jazdi, B. Ashtari Talkhestani, B. Maschler, and M. Weyrich, "Realization of AI-enhanced industrial automation systems using intelligent Digital Twins," *Procedia CIRP*, vol. 97, pp. 396–400, 2021.
- [32] L. März, W. Krug, O. Rose, and G. Weigert, *Simulation und Optimierung in Produktion und Logistik*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.
- [33] N. Sahlab, S. Kamm, T. Müller, N. Jazdi, and M. Weyrich, "Knowledge graphs as enhancers of intelligent digital twins," *4th IEEE International Conference on Industrial Cyber-Physical Systems (ICPS)*, pp. 19–24, 2021.
- [34] V. Stegmaier, D. Dittler, N. Jazdi, and M. Weyrich, "A structure of modelling depths in behavior models for Digital Twins", submitted, 2022
- [35] V. Stegmaier, W. Schaaf, N. Jazdi, and M. Weyrich, "Simulation model for digital twins of pneumatic vacuum ejectors", submitted, 2022

Validierung von sicherheitsrelevanten KI-Systemen in der Pharmabranche

M.Sc. **A. Charizanis**,

Universität Stuttgart und Exyte GmbH, Stuttgart;

M.Sc. **M. Müller**, Prof. Dr.-Ing. Dr. hc **M. Weyrich**,

Institut für Automatisierungstechnik und Softwaresysteme, Universität Stuttgart;

H. Mettler, Exyte GmbH, Stuttgart

Kurzfassung

Moderne Automatisierungssysteme arbeiten zunehmend in komplexen und dynamischen Umgebungen und verwenden Methoden der Künstlichen Intelligenz (KI). Um den daraus resultierenden Anforderungen an die Anpassungsfähigkeit und Generalisierungsfähigkeit gerecht zu werden, kommen daher KI-Systeme immer mehr zum Einsatz.

Da vor allem Deep-Learning-Modelle Black-Box-Eigenschaften aufweisen, sind diese für Menschen schwer zu interpretieren. Herkömmliche Methoden des Software-Engineerings oder spezifische Lifecycle-Validierungsmodelle für automatisierte Systeme im Life Science Bereich (z.B. V-Modell, ISPE GAMP 5) zeigen zudem nur eine bedingte Eignung zur Validierung von KI-Systemen in der regulierten Industrie.

Der regulatorische Rahmen, insbesondere für den KI-Einsatz im pharmazeutischen, durch Good-Manufacturing-Practice (GMP) regulierten Bereich, ist bisher praktisch ungeklärt und greift die besonderen Anforderungen von KI nicht auf. Demnach bleibt die Validierung von KI-Systemen in dieser Domäne eine offene Frage.

Ziel des Beitrages ist es, eine Validierungsstrategie vorzustellen, welche die besonderen Anforderungen von KI für die Domäne der Pharmabranche im Fokus hat und den Aufwand der Validierung mithilfe automatisierter Teststrategien begrenzt.

Dazu wird in diesem Beitrag eine Methode vorgestellt, die im Einzelnen folgende Punkte umfasst:

- Die Einteilung der KI-Anwendung entsprechend ihres Einsatzzweckes anhand eines Regelwerks und Zuordnung der relevanten Normen.
- Die Ableitung von Qualitätskriterien des KI-spezifischen Teils der Anforderungsspezifikation basierend auf dem Stand der Technik.

- Die Überprüfung des Datenmanagements und Modell-Trainings anhand von neuronaler Auslastung, statistischer Repräsentativität der Testdatensätze im Vergleich zu Trainings- und Referenzdaten und einer Laufzeit-Überwachung von KI-Systemen auf Basis der o.g. Qualitätskriterien zur Erstvalidierung.

Anhand einer Demonstrations-Applikation wird anschließend die Umsetzung der Methodik aus Nutzersicht präsentiert. Als Fallstudie dient eine einsatzfähige KI im Bereich der pharmazeutischen visuellen Inspektion. Anhand von verschiedenen Testszenarien wird die Validierungsmethodik veranschaulicht.

1. Einleitung

Moderne Automatisierungssysteme des pharmazeutischen Bereichs arbeiten zunehmend in komplexen und dynamischen Umgebungen. Um den daraus resultierenden Anforderungen an die Anpassungsfähigkeit und die Generalisierungsfähigkeit gerecht zu werden, kommen daher zunehmend Methoden der Künstlichen Intelligenz (KI), insbesondere des Deep Learnings (DL), zum Einsatz. Auch im pharmazeutischen Bereich wird KI von der Mehrheit der Experten als eine der Schlüsseltechnologien der kommenden Jahre gesehen [1].

Die Sicherstellung einer kontinuierlichen hohen Qualität und Sicherheit von pharmazeutischen Produkten sind letztendlich Voraussetzung für die Sicherstellung der Patientensicherheit. Zudem steht die pharmazeutische Industrie vor der Herausforderung, bei gleichbleibender Qualität immer höhere Stückzahlen zu produzieren.

Der Einsatz von KI-Anwendungen in der Qualitätskontrolle der pharmazeutischen Produktion ist daher ein interessantes Szenario, da beispielsweise Prozesse zur visuellen Inspektion von Arzneimittelprodukten meist Tätigkeiten mit ermüdendem Charakter für das Personal sind und so Potential für Fehler darbieten. Der Einsatz von KI kann dabei Abhilfe schaffen, indem die Expertise des Personals durch die KI ersetzt wird. Im pharmazeutischen Bereich müssen solche computergestützten Systeme allerdings validiert werden, um sicherzustellen, dass diese ihre Aufgabenstellung ordnungsgemäß erfüllen. Besondere Eigenschaften von KI, wie beispielsweise das nicht-deterministische Verhalten, dynamische Lernverhalten oder deren Black-Box Charakter, stellen Hürden für die Validierung dar. Aus regulatorischer Sicht ist darüber hinaus die Lage bezüglich des Einsatzes von KI-Systemen bisher nicht eindeutig geklärt. Es fehlen harmonisierte Normen, welche diese spezifischen Charakteristika von KI-Systemen berücksichtigen und so die Validierung des sicherheitserfüllenden Charakters standardisieren. Die Konsequenz aus einem nicht vollends validierten System in der visuellen Inspektion ist eine Fehlklassifikation und somit ein negativer Einfluss auf die Produkt- und Patientensicherheit.

Um eine Lösung für das Problem der Validierung von KI-Systemen in der Pharma-Domäne zu erarbeiten, werden im Rahmen dieser Arbeit folgende Forschungsfragen gestellt. Diese gilt es zu beantworten:

- Gibt es die Möglichkeit, durch den Einsatz von einer assistierten Validierungsstrategie Risiken zu identifizieren?
- Wie können Risiken beim Einsatz von KI-Applikationen in der regulierten pharmazeutischen Herstellung minimiert werden?

Um dies bewerkstelligen zu können werden erst die regulatorische Lage für KI und mögliche Validierungsansätze aufgezeigt (Kapitel 2). Anschließend wird ein Validierungskonzept für pharma-spezifische KI-Applikationen entwickelt, in eine Demonstrationsapplikation überführt

und vorgestellt (Kapitel 3). Dieses Konzept wird dann anhand einer Fallstudie zur visuellen Inspektion von Paracetamol Blistern evaluiert (Kapitel 4). Abschließend werden die Ergebnisse zusammengefasst und ein Ausblick präsentiert (Kapitel 5).

2. Regulatorische Lage und mögliche Ansätze

Die Norm zur funktionalen Sicherheit sicherheitsbezogener elektronischer Systeme (DIN EN 61508) sieht künstliche Intelligenz als Maßnahme zur Fehlerkorrektur und empfiehlt deren Verwendung für die Sicherheitsintegritätslevel (SIL) 2 - 4 nicht. Falls diese dennoch verwendet wird, dann sollte der Grund ausführlich dargelegt werden [2]. Dies erweist sich als schwierig, da keine europäische, harmonisierte Norm, welche auf den pharmazeutischen Bereich zugeschnitten ist, vorliegt.

In der Forschung liegen allerdings einige Konzepte und Vorschläge zur Regulierung von KI-Systemen vor. Die wohl bekanntesten und relevantesten sind der Vorschlag des europäischen Parlaments [3] und das von der „Food and Drug Administration“ (FDA) veröffentlichte Diskussionspapier zur Regulierung von Modifikationen von KI-basierten Medizinprodukten [4]. Ersterer könnte in Zukunft rechtlich bindend werden und beschreibt allgemeingültige Anforderungen an KI-Systeme. Die FDA befasst sich in ihrem Vorschlag vor allem mit der Problematik von Modifikationen und der damit einhergehenden Re-Zertifizierung. Für die Entwicklung der KI verweist die FDA auf die „Good Machine Learning Practices“ (GMLP), welche bisher aus 10 Leitprinzipien bestehen [5].

Neben diesen Vorschlägen liegen noch einige Leitfäden und Guidelines zur Entwicklung von KI-Systemen vor. Die DIN SPEC 13266 und 13288 bieten einen Leitfaden zur Entwicklung von Bilderkennungssystemen dar und die DIN SPEC 9200 – Reihe stellt ein Lebenszyklus-Modell für KI-Systeme vor [6–8]. Eine Gruppe aus über 30 Industrievertretern der US-amerikanischen Universität Xavier stellt außerdem Guidelines zur Sicherstellung einer ausreichenden Datenqualität zur Verfügung [9–11]. Sie behandelt dabei auch weiterlernende Systeme. Das Deutsche Institut für Normung veröffentlichte eine „Normungs-Roadmap“, welche Anforderungen und Herausforderungen von KI-Systemen beschreibt und den Standardisierungsbedarf klarstellt [12]. Die Interessengemeinschaft der Benannten Stellen für Medizinprodukte in Deutschland (IG-NB) hat in Kooperation mit dem Johner Institut einen Fragenkatalog zur Zertifizierung von KI-Systemen erstellt [13, 14]. Dieser wird auch vom TÜV Süd als Hilfestellung benutzt, um KI-Systeme zu zertifizieren [15]. Die deutsche Akademie der Technikwissenschaften „acatech“ hat darüber hinaus ein Whitepaper veröffentlicht, welches Kriterien zur Zertifizierung von KI aufzeigt [16]. Weitere Leitlinien und Standards sind in Entstehung, beispielsweise eine Reihe an Projekten der ISO/IEC JTC1 SC42, der IEEE oder der British Standards Institution(BSI)

[17, 18]. Im pharmazeutischen Bereich liegen solche konkreten Projekte und Leitlinien nicht vor. Allerdings wurden einige Interessengruppen von der ISPE, PDA oder Biophorum gegründet, um das Thema zu diskutieren.

Für die Konzeptionierung wurde eine Recherche durchgeführt, um mögliche Ansätze zur Validierung von KI-Systemen ausfindig zu machen. Ansätze, um explizit eine KI-spezifische Validierung im pharmazeutischen Bereich zu ermöglichen liegen dabei nur in geringer Anzahl vor [19, 20]. Allerdings können einige Ansätze, wie beispielsweise der Vorschlag der FDA für die Konzeptionierung einer Validierungsstrategie herangezogen werden. Es zeigt sich, dass die gefundenen Ansätze entweder einen prozessorientierten [4, 6, 7, 13, 19, 20] oder einen risikobasierten Ansatz [21, 22] verfolgen. Als äußerst umfangreich und geeignet für die Identifizierung und Minimierung von Risiken hat sich der Prüfkatalog des Fraunhofer IAIS herausgestellt [22]. Für die Konzeptionierung wurden anschließend diese Ansätze kombiniert und ergänzt. Das Ergebnis wird in Kapitel 3 vorgestellt.

3. Konzept zur Validierung von KI-Systemen

Das in Bild 1 vorgestellte Schema beschreibt den Ablauf des Validierungskonzeptes. Ausgangspunkt bildet eine Zweckbestimmung, welche anhand eines Regelwerkes abgeleitet wird und den Anwendungsbereich der KI näher konkretisiert. Darauf basierend werden Anforderungen, relevante Normen und Regularien zugewiesen. Diese werden anschließend vom Prüfer untersucht, näher formalisiert und bekannten Risiken zugewiesen, welche vom Prüfer bewertet und priorisiert werden. Basierend auf dieser Priorisierung ergeben sich Maßnahmen für die KI-spezifischen Punkte des Daten- und Modellmanagements. Diese Maßnahmen können Dokumentationen, die Etablierung von Prozessen oder die Durchführung von Tests sein, wobei letztere das einhergehende Erfolgskriterium erfüllen müssen. Ist dies nicht der Fall, wird vorerst untersucht, ob eine Modifikation der KI-Anwendung notwendig ist oder ob ein Spezifikationsfehler vorliegt und das Kriterium zu hoch gesetzt wurde. In der darauffolgenden Laufzeitüberwachung wird anschließend nach vordefinierten Zeitabständen geprüft, ob sich die KI-Anwendung im validen Zustand befindet. Hierfür werden Test-Verfahren und deren dazugehörigen Überwachungsmetriken herangezogen. Sollte das Erfolgskriterium einer Überwachungsmetrik nicht erfüllt sein, wird untersucht, ob eine Modifikation nötig ist und eine Re-Validierung erfolgen muss.

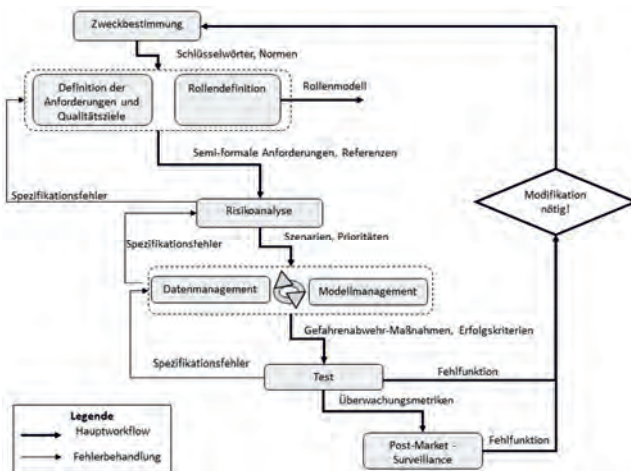


Bild 1: Ablauf des Validierungskonzeptes

Um die Zweckbestimmung des KI-Systems näher zu konkretisieren, erfolgt wie in Bild 2 beschrieben eine Einteilung der KI-Anwendung anhand eines Regelwerkes. Dieses stellt dem Prüfer verschiedene Optionen zur Auswahl. Hierbei wird erst die pharmazeutische Aufgabenstellung näher konkretisiert. Anschließend folgt eine Einteilung anhand der KI-spezifischen Aufgabe & Funktionalität und des Einsatzkontextes.

Die eingegebenen Auswahloptionen werden verschlagwortet. Dabei werden den Auswahloptionen kurze und prägnante Schlüsselwörter zugewiesen, welche zum einen für die Suche nach relevanten Normen und Regularien herangezogen und zum anderen für die spätere Umfeldanalyse verwendet werden. Normen und Regularien werden dabei im Vorfeld in einer Datenbank gesammelt und mit passenden Schlüsselwörtern annotiert, um diese mit denen der Auswahloptionen vergleichen zu können.

Auf Basis der Schlüsselwörter und Normen werden relevante Vorgängerprojekte gesucht. Dies kann beispielsweise eine ähnliche visuelle Inspektion anhand eines anderen Arzneimittels sein. Aus diesen Vorgängerprojekten werden anschließend Anforderungen abgeleitet und auf KI-spezifische Anforderungen gefiltert. Dadurch wird der Aufwand zur Erfassung der relevanten Anforderungen reduziert. Um diese näher zu konkretisieren und Qualitätskriterien abzuleiten, wird eine Umfeldanalyse anhand der beschriebenen Schlüsselwörter durchgeführt. Hierfür wird aus den vorliegenden Schlüsselwörtern ein Suchstring generiert, welcher anschließend für eine Literaturrecherche verwendet wird. Liegt die Anzahl der Suchergebnisse über einer

vorgegebenen Threshold, werden diese gespeichert und mithilfe des Open Source Programms „VosViewer“ [23] einer bibliometrischen Analyse unterzogen. „VosViewer“ ist in der Lage, eine Reihe zugeführter Literatur auf Basis eines „Natural language processing“-Algorithmus auf häufig vorkommende Begriffe zu untersuchen. Die für die KI-Anwendung relevanten Begriffe werden ausgewählt. Anschließend werden die Literaturergebnisse nach der Anzahl der auftretenden Schlüsselwörter sortiert. Die Literatur mit den meisten Schlüsselwort-Vorkommnissen wird dann als relevant bezeichnet und als Basis für die Ableitung von Benchmarks herangezogen, um die Anforderungen so näher zu formalisieren. Dieser Prozess hilft dabei, den Stand der Technik schneller und zuverlässiger abzudecken und so einen Vergleich mit bestehenden Anwendungen darzubieten.

Simultan werden die Rollen für die KI-relevanten Tätigkeiten vergeben. Diese werden auf Basis der Vorgängerprojekte vorgeschlagen und fordern die Rolle eines KI-Fachexperten.

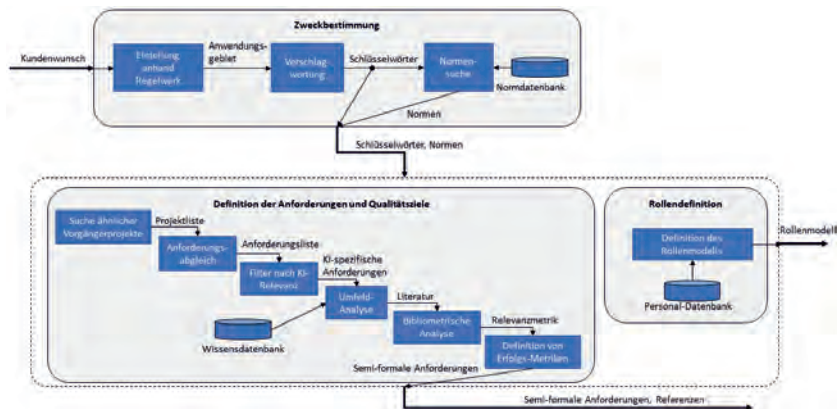


Bild 2: Konzeptschema zur Zweckbestimmung und Definition der Anforderungen und Qualitätsziele

Die eben beschriebenen Anforderungen werden herangezogen, um mögliche Risikoszenarien zu ermitteln. Hierzu werden die Anforderungen mit einem Risikoszenario-Katalog abgeglichen und entsprechende Risikoszenarien vorgeschlagen (siehe Bild 3). Diese gliedern sich in Szenarien in Bezug auf Daten, Robustheit, Transparenz, Post-Market-Surveillance (PMS) und dem Automatisierungsgrad auf. Durch die Zuweisung der Risikoszenarien wird der Aufwand für den Prüfer in Grenzen gehalten und sichergestellt, dass kein mögliches Risikoszenario unberücksichtigt bleibt. Diese Risikoszenarien können anschließend durch den Prüfer bear-

beitet und gespeichert werden. Angelehnt an die Risikoanalyse nach dem Leitfaden für Computersystemvalidierung ISPE GAMP 5 [21] werden diese Risikoszenarien anhand ihres Schweregrades, ihrer Auftretenswahrscheinlichkeit und ihrer Erkennungswahrscheinlichkeit bewertet. Dabei beschreibt der Schweregrad den Einfluss auf Produktqualität, Patientensicherheit oder Datenqualität. Die Auftretenswahrscheinlichkeit kann entweder qualitativ oder durch Fehlerraten beschrieben werden. Weiterhin beschreibt die Erkennungswahrscheinlichkeit die Geschwindigkeit, mit der ein Fehler erkannt wird und ist somit auch abhängig vom Automatisierungsgrad. In einem vollautomatisierten System werden Fehler beispielsweise später entdeckt als in einem System mit Berater-Funktion. Diese Kategorien werden in drei mögliche Stufen eingeteilt: gering (1), mittel (2), und hoch (3) für Schweregrad und Auftretenswahrscheinlichkeit und sofort (1), vor dem Vorfall (2) und nach dem Vorfall (3) für die Erkennungswahrscheinlichkeit. Die Multiplikation dieser Einstufungen ergibt dann die Risikoprioritätszahl (RPZ), welche für die Priorisierung herangezogen wird. Dadurch soll gezeigt werden, auf welche Risikoszenarien besonders Acht gegeben werden muss. Basierend auf diesen priorisierten Risiken ist nun eine Zuweisung an entsprechenden Maßnahmen zur Minimierung des Risikos möglich. Diese werden ebenfalls aus einem Maßnahmen-Katalog entnommen und dem Risiko zugewiesen. Grundsätzlich lassen sich die getroffenen Maßnahmen in die Kategorien **Datenmanagement** und **Modellmanagement** einteilen und können eine Dokumentation, die Etablierung eines Prozesses oder die Durchführung eines Tests fordern.

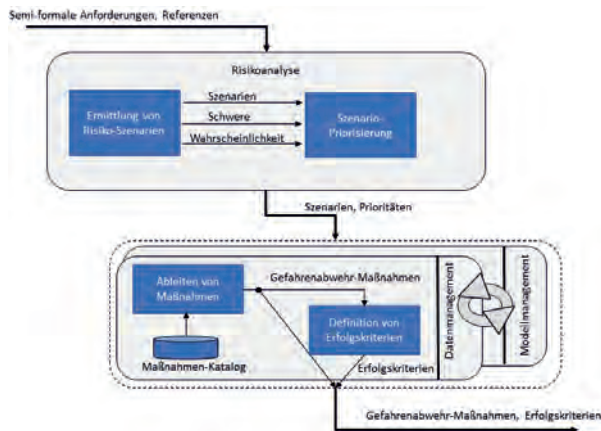


Bild 3: Konzeptschema zur Ermittlung der Risikoszenarien und der Ableitung von Maßnahmen

Für den Fall einer folgenden Testdurchführung folgt eine Testauswahl anhand einer zugeordneten Testfall-Datenbank (siehe Bild 4). Dabei werden den vorgegebenen Maßnahmen mögliche Tests zugewiesen, welche die Minimierung des zugrundeliegenden Risikos als Ziel haben. Im Rahmen dieser Arbeit wurden 4 Tests implementiert, um die Repräsentation der Test-Daten, die erlernten Entscheidungskriterien und die Leistung auf augmentierten Daten zu überprüfen.

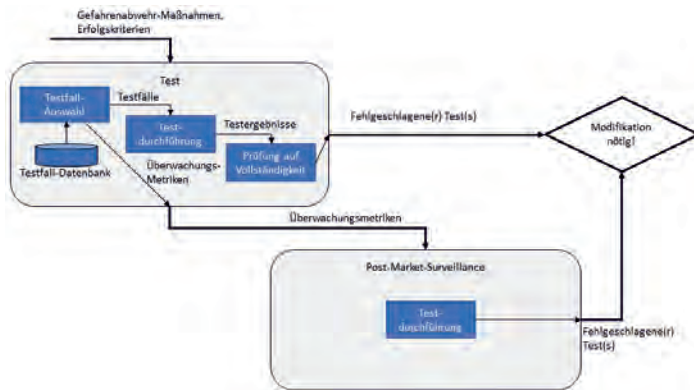


Bild 4: Konzeptschema zu Testen und Laufzeitüberwachung

Für die Überprüfung der **Test-Daten-Repräsentation** wird untersucht, ob die verwendeten Testdaten und Trainingsdaten aus derselben Daten-Verteilung stammen. Dies wird mithilfe eines antrainierten GAN-Diskriminators („Generative adversarial network“ – GAN) bewerkstelligt. GANs sind bekannt dafür Daten anhand eines neuronalen Netzes künstlich zu generieren und zeichnen sich durch eine Kombination aus zwei neuronalen Netzen aus. Eines übernimmt die Aufgabe des Generators und generiert künstliche Bilder auf Basis der zugeführten Trainingsdaten. Das zweite neuronale Netz dient als Diskriminator und versucht die künstlich generierten Bilder von originalen Bildern zu unterscheiden. Der dabei trainierte Diskriminator des GANs wird dann verwendet, um zugeführte Daten auf Echtheit zu überprüfen (siehe Bild 5). Um die Threshold für die Beurteilung der Verteilungszugehörigkeit zu berechnen, stellt der Prüfer sowohl einen validen Datensatz als auch einen invaliden Datensatz mit einer überschaubaren Anzahl an Test-Daten bereit. Für jedes einzelne Bild wird anschließend die Wahrscheinlichkeit zur Verteilungszugehörigkeit berechnet. Diese Sätze an Wahrscheinlichkeiten werden mithilfe des Boot-Strap-Verfahrens künstlich vermehrt, um so eine erhöhte statistische

Aussagekraft zu erhalten. Durch das Aufstellen eines Konfidenzintervalls werden zwei Grenzwerte festgelegt. Die untere Grenze des validen Datensatzes beschreibt, welchen Wert valide Daten mit einem bestimmten Wahrscheinlichkeitsgrad mindestens besitzen. Die obere Grenze des invaliden Datensatzes hingegen beschreibt welchen Wert invalide Daten mit einer bestimmten Wahrscheinlichkeit besitzen. Der Mittelwert dieser zwei Werte bildet den Schwellwert zur Beurteilung von validen und invaliden Daten. Liegt der Diskriminator-Wert eines Test-Daten-Beispiels über diesem Schwellwert, kann dieses als repräsentativ angesehen werden.

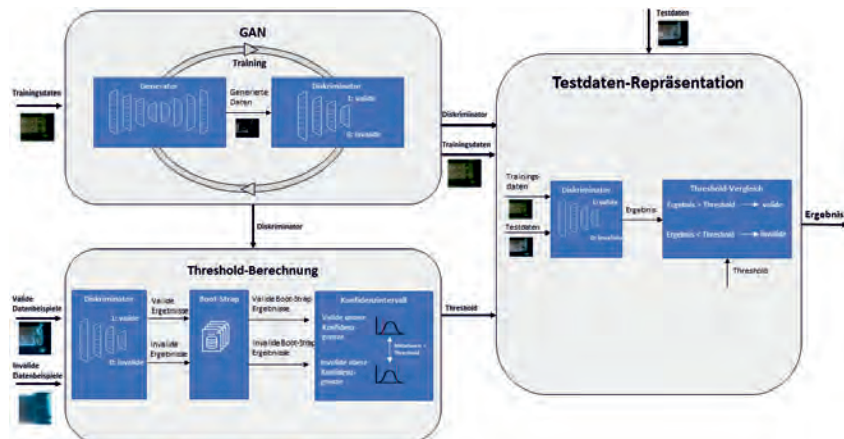


Bild 5: Überprüfung auf Echtheit mittels GAN-Diskriminator

Die **neuronale Auslastung** eines neuronalen Netzwerkes wird durch das Verhältnis aktivierter Neuronen zur Gesamtanzahl an Neuronen definiert [24]. Ein Datenbeispiel, welches eine hohe neuronale Auslastung erzielt, aktiviert das neuronale Netz besonders und weist möglicherweise selten auftretende Merkmale auf. Um solche Datenbeispiele zu identifizieren, werden mithilfe einer Datenaugmentierung verschiedene Testszenarien augmentiert und anschließend getestet, ob deren erzielte neuronale Auslastung über der durchschnittlichen Auslastung des verwendeten Test-Datensatzes liegt (siehe Bild 6). Dieser Datensatz wird dann dem KI-Modell zugeführt und ausgewertet. Liegt die Leistung des Datensatzes über einem vordefinierten Kriterium, gilt der Test als bestanden. Bild 6 veranschaulicht diesen Prozess. Sollten Datenbeispiele nicht korrekt klassifiziert werden, so werden diese durch den Prüfer näher untersucht.



Bild 6: Testgenerierung und Prüfung anhand neuronaler Auslastung

Beim **metamorphischen Testen** werden Testszenarios generiert, welche in Bezug auf ein korrektes Testszenario leicht modifiziert werden [25]. Die vordefinierten metamorphischen Beziehungen bestimmen dabei den Erfolg der Tests. So werden mithilfe eines Datengenerators Testbilder innerhalb bestimmter Qualitätskriterien augmentiert. Diese augmentierten Bilder sollten anschließend keinen Einfluss auf die Aussage der KI besitzen. Eine metamorphische Relation kann beispielsweise eine vertikale Spiegelung des Test-Beispiels sein. Wenn die relevanten Merkmale des Test-Beispiels unabhängig von dieser Relation sind, so sollte die Entscheidung der KI nicht dadurch beeinflusst werden. Um dies zu prüfen, wird der Test-Datensatz anhand einer metamorphischen Relation augmentiert und anschließend die Aussage der KI vor bzw. nach der Augmentierung verglichen. Dadurch soll sichergestellt werden, dass die Aussagen der KI konsistent im Rahmen dieser Relationen sind und nicht durch leichte Modifikationen beeinflusst werden. Ändert sich die Aussage innerhalb einer Relation, so gilt der Test als nicht bestanden.

Es gibt verschiedene Methoden, um die Aussage der KI erklärbar zu gestalten. Diese Methoden werden auch **Explainability-Methoden** genannt. Eine davon ist das GradCAM++-Verfahren, welches eine Heatmap der Aussage generiert [26]. Diese markiert die Merkmale, welche den größten Anteil für die Entscheidung der KI tragen. Auch das LIME-Verfahren [27] bietet eine solche Interpretation an und wird als Zusatz zu einem Heatmap-Verfahren verwendet. Es unterteilt das originale Bild in kleinere Teile und prüft deren Vorhersagen, um so die Bildpartien mit der höchsten Vorhersage-Genauigkeit herauszufiltern. Die Ausgaben dieser Methoden werden dann anhand der vorgegebenen Klasse von dem Prüfer untersucht. Es liegt am Prüfer, zu entscheiden, ob das KI-Modell die Aussagen auf sinnvoller Basis tätigt. Wenn dies nicht der Fall ist, gilt der Test als nicht bestanden. Dadurch soll sichergestellt werden, dass die KI relevante Entscheidungskriterien erlernt hat und die definierte Zweckbestimmung auch ordnungsgemäß erfüllt.

Für die nachfolgende **Laufzeitüberwachung** werden Tests mit entsprechenden Überwachungsmetriken zugewiesen. Diese können quantitativ, beispielsweise anhand der Genauigkeit, aber auch qualitativ, zum Beispiel durch Explainability-Methoden, sein. Durch die Tests sollen folgende Punkte gegeben sein.

Eine **Performanzüberwachung** der KI während des Betriebs soll sicherstellen, dass sich die Leistung der KI innerhalb der vorgegebenen Qualitätskriterien befindet. Hierfür werden die zur Evaluierung des KI-Modells verwendeten Metriken übernommen (z.B. die Genauigkeit). Die beschriebenen Tests anhand der neuronalen Auslastung, metamorphischen Relationen und Explainability-Methoden werden ebenfalls für die Überwachung der Leistung verwendet. Hierfür werden die Tests nach vordefinierten zeitlichen Abständen durchgeführt und ausgewertet. Weiterhin muss untersucht werden, ob ein möglicher „Concept Drift“ besteht. Hierfür wird geprüft, ob die eingehenden Daten und die für die KI-Anwendung vorgesehene Aufgabe weiterhin mit den definierten Eingabedaten-Schema bzw. Aufgabenstellung übereinstimmen und ob die definierten Betriebsumgebungen immer noch zutreffen. Tritt ein solcher Ausbruch aus den vordefinierten Grenzen ein, muss eine Modifikation am KI-System vorgenommen und eine erneute Validierung durchgeführt werden (siehe Bild 4). Der Validierungsaufwand hält sich allerdings durch die Ableitung der Anforderungen, Risiken und Maßnahmen anhand von Referenzprojekten in Grenzen, da diese nicht neu erarbeitet, sondern nur angepasst werden müssen. Um mögliche Fehlklassifikationen während des Betriebs zu verhindern, werden die einkommenden Daten zusätzlich auf Plausibilität geprüft. Dies kann mithilfe des in Bild 5 beschriebenen Diskriminators vollzogen werden. Analog zur Prüfung der Test-Daten-Repräsentation sollen dabei die eingehenden Daten durch den Diskriminator bewertet und bei repräsentativen Daten der KI zugeführt bzw. bei nicht repräsentativen Daten dem Nutzer zur Kontrolle vorgelegt werden.

Ein Prozess zur Überwachung und **Kontrolle der KI durch den Menschen** muss zudem für einen gewissen Zeitraum gewährleistet sein. Die entsprechende Zeitspanne korreliert dabei mit dem einhergehenden Risiko der KI-Anwendung. Auffälligkeiten, wie Fehlklassifikationen oder unsichere Aussagen müssen dabei aufgezeichnet und bewertet werden. Nach dieser Kontrollphase sollte auch der eigentliche Nutzer (im Falle einer Berater-Funktion der KI) in der Lage sein, solche Auffälligkeiten und Beschwerden einzureichen.

Für alle Vorgänge der Performanzüberwachung und der im Betrieb vollzogenen Entscheidungen, muss ein **Protokollierungs-** bzw. Speicherkonzept angefertigt werden, welches sicherstellt, dass alle Auffälligkeiten während des Betriebs der KI gesichert werden und nachverfolgbar sind. Dabei ist darauf zu achten, vor allem bei Unfällen die Art und den Verlauf zu

protokollieren und festzuhalten. Es kann beispielsweise ein erneutes Training der KI erfolgen oder eine zusätzliche Absicherungsmaßnahme implementiert werden.

4. Evaluierung anhand einer Fallstudie

Im Rahmen der Evaluierung wird ein Anwendungsfall für eine visuelle Kontrolle von defekten Blister-Packungen nachgestellt. Hierfür wurden im Vorfeld eine Reihe an Arzneimittel des Wirkstoffs Paracetamol erworben und über ein Förderband bewegt. Ein Stopper hält dabei den Träger samt Blister beim Vorbeifahren an, wodurch ein Schnappschuss des Blisters durch ein Kamerasystem erzeugt wird. Bild 7 zeigt den Versuchsaufbau. Diese Bilder werden anschließend in 70% Trainings- 15% Referenz- und 15% Testdatensätze unterteilt und einem neuronalen Netz für ein Training zugeführt. Dieses neuronale Netz erreicht anschließend auf allen Datensätzen eine Genauigkeit von nahezu 100%.



Bild 7: Versuchsaufbau Förderband für Blister-Transport und Beispielaufnahmen

Nun wird eine Validierung anhand des beschriebenen Konzeptes und durch die Demonstrationsapplikation vollzogen. Die Zweckbestimmung ermöglicht dabei die Zuweisung bekannter Anforderungen aus einem Referenzprojekt. Durch die Umfeldanalyse und bibliometrische Analyse werden anschließend Studien vorgeschlagen, welche zur Konkretisierung betrachtet werden können. Dieser Prozess ist allerdings nicht implementiert, weshalb er manuell durchgeführt wird. Diese konkretisierten Anforderungen führen dann zu einer Zuweisung von Risikoszenarien. Ein beispielhaftes Risikoszenario ist eine Fehlklassifikation des neuronalen Netzes. Dieses könnte durch das Auftreten eines Datenbias entstehen. Als Maßnahmen, um dieses Risiko zu minimieren, werden Tests anhand von Explainability-Methoden und die Modellprüfung anhand der neuronalen Auslastung vorgeschlagen. Der Test zur Prüfung der Kompatibilität zwischen Test- und Trainingsdatensatz ist für die weiteren Verfahren grundlegend wichtig, da dieser prüft, ob der vorliegende Test-Datensatz überhaupt repräsentativ für den Anwendungsfall ist und für die weiteren Tests auch verwendet werden kann. Aus diesem Grund wird

dieser vor den anderen Tests durchgeführt. Hierfür werden jeweils 15 valide und invalide Beispiel-Daten als Referenz hochgeladen und anschließend für die Bestimmung der Threshold gemäß Bild 5 herangezogen. Für die Datenverteilung zeigt sich, dass 98,26% der Test-Daten über der Threshold liegen und als repräsentativ eingeschätzt werden können. Test-Daten, welche als nicht repräsentativ eingestuft wurden, werden genauer betrachtet. Bei den Beispielen handelt es sich allerdings um Ausreißer, welche nach näherer Betrachtung als repräsentativ eingeschätzt werden können.

Explainability Methoden: Um zu prüfen, ob die KI die für die Aufgabenstellung relevante Entscheidungskriterien erlernt hat, werden in der Applikation unter dem Bereich **Explainability** entsprechende Methoden angezeigt. Es kann dann zwischen dem GradCAM++ und dem LIME-Verfahren ausgewählt werden. Bild 8 zeigt einen Auszug der Ergebnisse für das GradCAM++-Verfahren auf. Der rot markierte Bereich zeigt dabei relevante Merkmale auf, welche am stärksten für die aktuelle Aussage der KI sprechen. Ein Vergleich zwischen den Bildern zeigt, dass vor allem helle Bildpixel für die Entscheidung des Netzes eine relevante Rolle spielen (dieses Phänomen hat sich auch über den restlichen Test-Daten bestätigt). Für das LIME-Verfahren ergeben sich ähnliche Ergebnisse (siehe Bild 8). Der gelb markierte Bereich zeigt dabei die Bildpartie mit der höchsten Vorhersage-Übereinstimmung an. Auch hier werden Bildpartien markiert, welche keinen Einfluss für die eigentliche Defektvorhersage haben sollten. Aus den Ergebnissen dieser Methoden können die Entscheidungen der KI somit nicht nachvollzogen werden, weshalb dieser Test als nicht bestanden gilt.

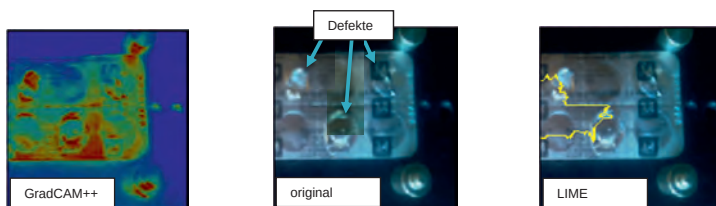


Bild 8: Ergebnisse des GradCAM++-Verfahrens und LIME-Verfahrens

Modellprüfung anhand von neuronaler Auslastung: Für diesen Test werden auf 88,75% der generierten Test-Daten die Ergebnisse korrekt vorhergesagt. Fehlklassifikationen, wie beispielsweise in Bild 9 zu sehen, zeigen allerdings auf, dass teilweise repräsentative Bilder nicht korrekt klassifiziert werden. Dabei werden defekte Blister als unbeschädigt markiert (links) und unbeschädigte Blister als defekt (rechts). Der Annahme aus den Explainability-Methoden fol-

gend, wäre es möglich, dass vor allem die Helligkeit eines Bildes einen Einfluss auf die Vorhersage des KI-Modells hat. Helle Pixel scheinen dabei mit dem Vorhandensein eines Defektes in Verbindung zu stehen. In einer realen Anwendung des Systems könnte man demnach mit einer einfachen Lichtquelle die Aussage der KI beeinflussen. Aus diesem Grund gilt dieser Test ebenfalls als nicht bestanden.

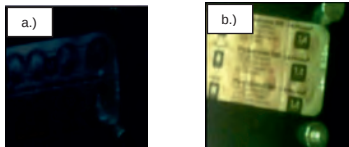


Bild 9: Fehlklassifikationen bei augmentierten Testdaten: a.) ist defekt aber klassifiziert als unbeschädigt; b.) ist unbeschädigt aber klassifiziert als defekt

Die Rückschlüsse, welche aus diesen Tests gezogen werden, zeigen auf, dass das vorliegende Risiko einer Fehlklassifikation höher ist als zu erwarten war und somit unannehmbar ist. Grund dafür ist die Belichtung bei der Aufnahme von defekten und unbeschädigten Blistern. Defekte Blister-Aufnahmen wurden über Nacht mit künstlicher Beleuchtung und unbeschädigte Blister bei Tageslicht aufgenommen. Diese systematische Verzerrung führt dazu, dass die KI helle Bildpixel mit einem defekten Blister in Verbindung setzt. Das KI-System hat somit die Validierung nicht bestanden und muss modifiziert werden. Da diese Fallstudie nur beispielhaft für einen realen Anwendungsfall ist, und die Lichtverhältnisse nicht standardisiert werden können, muss eine Anpassung der Trainingsdaten herangezogen werden. Hierfür müssen mehr Trainingsbilder sowohl zu defekten als auch unbeschädigten Blistern bei unterschiedlichen Lichtbedingungen generiert werden. Dadurch wird die Varianz in beiden Klassen erhöht und verhindert das Auftreten einer solchen Korrelation.

Im Rahmen dieser Arbeit konnten diese Modifikationen aus zeitlichen Gründen nicht umgesetzt werden.

5. Zusammenfassung und Ausblick

In dieser Arbeit wurde ein Konzept zur Validierung von sicherheitsrelevanten KI-Systemen in der Pharma-Domäne entwickelt und vorgestellt.

Dieses zeichnet sich durch das Aufstellen und nähere Formalisieren KI-spezifischer Anforderungen basierend auf einer initialen Zweckbestimmung und dem Vergleich mit relevanten Vorgängerprojekten aus, um anschließend mögliche Risikoszenarien zuweisen zu können. Die Analyse dieser Risikoszenarien resultiert in einer dem Risiko entsprechenden Priorisierung.

Dadurch wird die Identifizierung der Risiken abgedeckt und transparent gestaltet. Das anschließende Vorschlagen von Maßnahmen in Bezug auf die KI-spezifischen Punkte des Datenmanagements und Modellmanagements weist Testverfahren zu den Punkten Test-Daten-Repräsentation, Explainability, neuronaler Auslastung und metamorphischen Relationen zu. Eine nachfolgende Laufzeitüberwachung stellt den validen Zustand des KI-Systems anhand von vordefinierten Performanzkriterien und einer temporären Kontrollphase durch den Menschen sicher. Durch diese Vorhaben können die identifizierten Risiken zum einen näher untersucht bzw. bewertet und in Bezug auf den nachfolgenden Einsatz minimiert werden.

Eine Evaluierung anhand einer Fallstudie zur visuellen Inspektion von beschädigten Paracetamol-Blistern zeigt auf, dass gerade das Testen anhand von Explainability-Methoden wichtige Rückschlüsse für die Beurteilung der Tauglichkeit des KI-Systems liefert und bei der Anforderung an Transparenz und Nachvollziehbarkeit Abhilfe schaffen kann.

Als Ausblick für die weitere Forschung wird die Erweiterung der Applikation auf weitere KI-spezifische Anwendungsfelder wie z.B. der Sequenzanalyse und die Ergänzung der Testverfahren durch beispielsweise dem Testen anhand von adversen Beispielen empfohlen.

Zudem muss im pharmazeutischen Bereich ein Branchenstandard mit entsprechenden KI-spezifischen Guidelines erstellt werden, um regulatorische Klarheit zu schaffen.

6. Literaturverzeichnis

- [1] Bitkom Research Group ; Tata Consultancy Services: Deutschland lernt KI : Wie Unternehmen digitale Technologien einsetzen. 25.11.2020
- [2] DIN EN 61508 - Funktionale Sicherheit sicherheitsbezogener elektrischer/elektronischer/programmierbarer elektronischer Systeme
- [3] Europäische Kommission: Vorschlag für eine Verordnung des europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz und zur Änderung bestimmter Rechtsakte der Union. URL <https://eur-lex.europa.eu/legal-content/DE/TXT/HTML/?uri=CELEX:52021PC0206&from=EN> – Überprüfungsdatum 2022-01-20
- [4] FDA - U.S. Food & Drug Administration: Proposed Regulatory Framework for Modifications to Artificial Intelligences / Machine Learning (AI/ML) - Based Software as a Medical Device (SaMD) : Discussion Paper and Request for Feedback. URL <https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf> – Überprüfungsdatum 2022-01-20
- [5] U.S. Food & Drug Administration: Good Machine Learning Practice for Medical Device Development: Guiding Principles
- [6] 13266. April 2020. DIN SPEC 13266 - Leitfaden für die Entwicklung von Deep-Learning-Bildererkennungssystemen
- [7] 13288. März 2021. DIN SPEC 13288 - Leitfaden für die Entwicklung von Deep-Learning-Bildererkennungssystemen in der Medizin
- [8] DIN SPEC 92001. April 2019. Artificial Intelligence - Life Cycle Processes and Quality Requirements
- [9] Baird, Pat ; Nayak, Rohit ; Karceski, Jackie: Data Quality for AI in Healthcare. URL https://static1.squarespace.com/static/58d0113a3e00bef537b02b70/t/6058aec5ce9a277129f04def/1616424647537/AI_Whitepaper_GMLP_Data+Quality+for+AI_final3.pdf – Überprüfungsdatum 2022-01-29
- [10] Baird, Pat ; Nayak, Rohit ; Nienburg, Kelly: Building Explainability and Trust for AI in Healthcare
- [11] Baird, Pat ; Nayak, Rohit: Perspectives and Good Practices for AI and Continuously Learning Systems in Healthcare
- [12] DIN Deutsches Institut für Normung e.V. ; DKE Deutsche Kommission elektrotechnik Elektronik Informationstechnik: Deutsche Normungsroadmap Künstliche Intelligenz. URL <https://www.din.de/resource/blob/772438/6b5ac6680543eff9fe372603514be3e6/normungs-roadmap-ki-data.pdf>

- [13] IG-NB - Interessengemeinschaft der Benannten Stellen für Medizinprodukte in Deutschland: Fragenkatalog "Künstliche Intelligenz bei Medizinprodukten". URL https://www.ig-nb.de/dok_view?oid=795601 – Überprüfungsdatum 2022-01-20
- [14] Johner Institut: Leitfaden zu KI bei Medizinprodukten. URL https://github.com/johner-institut/ai-guideline/blob/master/Guideline-AI-Medical-Devices_DE.md – Überprüfungsdatum 2022-01-31
- [15] Rad, Abtin: Artificial Intelligence in Medical Devices : Verifying and validating AI-based medical devices. URL <https://www.tuvsud.com/en/resource-centre/white-papers/artificial-intelligence-in-medical-devices> – Überprüfungsdatum 2021-08-19
- [16] Heesen, Jessica ; Müller-Quade, Jörn ; Wrobel, Stefan: Zertifizierung von KI-Systemen : Kompass für die Entwicklung und Anwendung vertrauenswürdiger KI-Systeme. November 2020
- [17] Petersen, Eike ; Potdevin, Yannik ; Mohammadi, Esfandiar ; Zidowitz, Stephan ; Breyer, Sabrina ; Nowotka, Dirk ; Henn, Sandra ; Pechmann, Ludwig ; Leucker, Martin ; Rostalski, Philipp ; Herzog, Christian: Responsible and Regulatory Conform Machine Learning for Medicine: A Survey of Technical Challenges and Solutions. 20.07.2021
- [18] COCIR - the European Coordination Committee of the Radiological, Electromedical and Healthcare IT Industry: Artificial Intelligence in EU Medical Device Legislation. September 2020
- [19] Tom Maes et al: Strategy for the GxP-validation of AI/Machine Learning solutions. URL https://www.invite-research.com/fileadmin/user_upload/Research/Publications/2019a/White_Paper_Strategy_for_the_GxP_validation_of_AI_Machine_Learning_solutions.pdf – Überprüfungsdatum 2022-01-20
- [20] Huysentruyt, Kristof ; Kjoersvik, Oeystein ; Dobracki, Pawel ; Savage, Elizabeth ; Mishalov, Ellen ; Cherry, Mark ; Leonard, Eileen ; Taylor, Robert ; Patel, Bhavin ; Abatemarco, Danielle: Validating Intelligent Automation Systems in Pharmacovigilance: Insights from Good Manufacturing Practices. In: Drug safety 44 (2021), Nr. 3, S. 261–272
- [21] Wingate, Guy (Hrsg.); Wyn, Sion (Hrsg.): GAMP 5 : A risk-based approach to compliant GxP computerized systems. Tampa, Florida : ISPE, 2008
- [22] Poreschkin, Maximilian ; Schmitz, Anna ; Akila, Maram ; Adilova, Linara ; Becker, Daniel ; Cremers, Armin ; Hecker, Dirk ; Houben, Sebastian ; Mock, Michael ; Rosenzweig, Julia ; Sicking, Joachim ; Schulz, Elena ; Voss, Angelika ; Wrobel, Stefan: Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz. URL https://www.iais.fraunhofer.de/content/dam/iais/fb/Kuenstliche_intelligenz/ki-pruefkatalog/202107_KI-Pruefkatalog.pdf

- [23] VOSviewer: VOSviewer - Visualizing scientific landscapes. URL <https://www.vosviewer.com/>. – Aktualisierungsdatum: 2022-05-16 – Überprüfungsdatum 2022-05-16
- [24] Fabrice Harel-Canada ; Lingxiao Wang ; Muhammad Ali Gulzar ; Quanquan Gu ; Miryung Kim: Is Neuron Coverage a Meaningful Measure for Testing Deep Neural Networks? : Zenodo, 2020
- [25] Xie, Xiaoyuan ; Ho, Joshua W. K. ; Murphy, Christian ; Kaiser, Gail ; Xu, Baowen ; Chen, Tsong Yueh: Testing and Validating Machine Learning Classifiers by Metamorphic Testing. In: The Journal of systems and software 84 (2011), Nr. 4, S. 544–558
- [26] Chattopadhyay, Aditya ; Sarkar, Anirban ; Howlader, Prantik ; Balasubramanian, Vineeth N.: Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks (2018), S. 839–847. URL <http://arxiv.org/pdf/1710.11063v3>
- [27] Ribeiro, Marco Tulio ; Singh, Sameer ; Guestrin, Carlos: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. 16.02.2016

Strategisches Drohnenverkehrsmanagement: Konzept für eine automatisierte Fluggenehmigung und -freigabe von Drohnenflügen

T. Grebner, A. Fay, Helmut-Schmidt-Universität, Hamburg

Kurzfassung

Die sichere Integration von Unbemannten Flugsystemen (Unmanned Aircraft Systems - UAS) in den bestehenden Luftverkehr ist eine große Herausforderung bei der Einrichtung eines Drohnenverkehrsmanagements in einem U-Space-Luftraum. Die sogenannte U-Space-Verordnung beschreibt verpflichtende U-Space-Dienste, die UAS-Betreibern helfen sollen, UAS sicher in den urbanen Luftraum zu integrieren. Einer dieser Dienste ist der UAS-Fluggenehmigungsdienst (U-FAS), welcher Fluggenehmigungsanfragen in der strategischen Planungsphase von UAS Flügen koordinieren und verwalten soll. Bisher existieren keine detaillierten Spezifikationen dieser Dienste. Dieser Beitrag stellt ein Konzept für einen automatisierten und U-Space-konformen Fluggenehmigungs- und Aktivierungsprozess vor, welcher die Mindestanforderungen an einen UAS-Fluggenehmigungsdienst aktueller U-Space-Rahmenwerke erfüllt.

1. Einleitung und Motivation

Unbemannte Luftfahrtsysteme (Unmanned aerial systems - UAS), besser bekannt als Drohnen, bergen ein großes wirtschaftliches Potenzial, indem sie die multimodale Mobilität um ein zusätzliches Transportmittel bereichern [1]. Mit der zunehmenden Zahl komplexer UAS-Operationen wie Infrastrukturinspektionen, Überwachungs- sowie Such- und Rettungseinsätzen hat sich der Einsatzbereich von UAS erweitert, und es wird erwartet, dass der UAS-Luftverkehr in urbanen Lufträumen in den kommenden Jahren erheblich zunehmen wird [2]. Um das wirtschaftliche Potenzial der UAS zu nutzen und gleichzeitig eine sichere Integration in den urbanen Luftraum zu ermöglichen, werden praxisnahe Konzepte und Lösungen für ein effizientes Drohnenverkehrsmanagement (UAS Traffic Management - UTM) benötigt. Das Drohnenverkehrsmanagement kann dabei grundsätzlich in zwei Phasen, eine strategische (vor dem Flug) und eine taktische (während des Fluges), eingeteilt werden [3].

In Europa wurden erste Ansätze zum Aufbau eines UTM in mehreren „SESAR-Horizon 2020“ Projekten wie USIS [4] und CORUS [5] entwickelt. In diesen Projekten wurde der Begriff U-Space für den europäischen Raum geprägt, er umfasst eine Reihe von möglichen Diensten

und Verfahren. Diese müssen im nächsten Schritt spezifiziert und auf Basis der rechtlichen Anforderungen im U-Space [6] entwickelt werden, um ein rechtlich-technisches Gesamtsystem zu realisieren.

Ein zentraler Dienst im U-Space ist der UAS-Fluggenehmigungsdienst (U-FAS), welcher primär Funktionen in der strategischen Phase übernimmt. Die Einordnung des Dienstes im gesamtheitlichen Drohnenmanagement erfolgte bereits in [7]. Zum Aufgabenumfang gehören die Prüfung, Genehmigung und Aktivierung von UAS-Flügen sowie die Verwaltung und Koordination bereits genehmigter Flugpläne. Zudem müssen auftretende Konflikte zwischen Flugplänen detektiert und unter Anwendung von Prioritätsregeln gelöst sowie andere Luftraumnutzer informiert werden [6].

Die Ausführung dieser Aufgaben wird dadurch erschwert, dass der Zeitpunkt des Eingangs einer UAS-Fluggenehmigungsanfrage Einfluss auf die Handhabung und Lösung hat. Zum einen liegen bei langfristig geplanten Flügen weniger Informationen zugrunde, wodurch beispielsweise die Wetter- oder Verkehrslage zum Zeitpunkt der Fluggenehmigungsanfrage nicht berücksichtigt werden kann. Dies erfordert eine erneute Prüfung des Flugplans zu einem Zeitpunkt, an dem diese Informationen zur Verfügung stehen. Zum anderen ist bei kurzfristig geplanten Flügen weniger Zeit, um die Prüfungs-, Genehmigungs-, und Aktivierungsprozesse zu realisieren. Kurzfristig geplante Flüge können demnach nur realisiert werden, wenn die Prozesszeiten möglichst kurz sind. Dies erfordert eine hohe Automatisierung der Prozesse, da die komplexen und zeitaufwändigen Aufgaben bei steigender Zahl von UAS nicht durch manuelle Handhabung realisiert werden können.



Bild 1: Strategische Planungsphase von Drohnenflügen

Um diese Probleme zu adressieren, wurde von den Autoren ein Konzept für einen möglichst automatisierten UAS-Fluggenehmigungsdienst entwickelt. Da für diesen Dienst bisher keine vollständigen Workflows oder Ablaufschemata existieren, welche die rechtlich-technischen Mindestanforderungen erfüllen, wurde im ersten Schritt ein robustes und adaptives Prozessmodell für einen effizienten und U-Space-konformen UAS-Fluggenehmigungsdienst entwickelt. Im zweiten Schritt sollen die erarbeiteten Workflow-Prozesse in ein ausführbares Gesamtprozessmodell überführt werden, welches die gesamte strategische Planungsphase adressiert. Erste Teilmodule der Gesamtsimulation konnten bereits softwaretechnisch umgesetzt werden, um erste Simulationsstudien durchzuführen. Der Fokus dieser Veröffentlichung liegt dabei zunächst auf der Vorstellung des Gesamtkonzepts für einen automatisierten Fluggenehmigungs- und Aktivierungsprozess.

2. Stand der Forschung – Hintergrund und bisherige Arbeiten

Bestehende Beiträge zum Thema U-FAS im U-Space können in rechtliche und technische Arbeiten unterteilt werden. Mit der Verabschiedung der U-Space-Verordnung [6] hat die EU die gesetzliche Grundlage für die Einrichtung von U-Space-Lufträumen geschaffen. In diesen geographischen Gebieten soll es UAS-Betreibern durch Inanspruchnahme von U-Space-Services ermöglicht werden, UAS sicher und effizient im U-Space-Luftraum zu nutzen. Die Umsetzung der Durchführungsverordnung in nationales Recht sowie die konkrete Ausgestaltung der Prozesse, Akteure und Verantwortlichkeiten obliegt den einzelnen Mitgliedsstaaten.

Bisherige Arbeiten zu den prozesstechnischen Abläufen eines U-FAS wie beispielsweise das im CORUS-Projekt beschriebene strategische Konfliktmanagement-Modell [8], beschreiben diese zumeist nur abstrakt, sodass derzeit existierende Prozessmodelle bisheriger Arbeiten nicht als Grundlage für die oben beschriebene Ausgestaltung der Prozesse, Akteure und Verantwortlichkeiten genutzt werden können. Zudem können die bisherigen Prozessmodelle die aus der U-Space-Verordnung ableitbaren Anforderungen nicht erfüllen.

Die Konzeption und Automatisierung der einzelnen Dienste im U-Space ist ein wichtiges Forschungsthema im Bereich des UTM-Forschungsfelds [7] zur Erfüllung der Forderungen der EU, diese Prozesse zu gestalten. Hierbei lassen sich zentrale Forschungsfragen ableiten. Eine zentrale Forschungsfrage, die sich im Zusammenhang mit dem U-FAS und der stetig wachsenden Anzahl an Drohnenoperationen ergibt, zielt darauf ab, wie ein U-FAS auch hohe Verkehrslasten automatisiert bewältigen kann. Um diese Forschungsfrage zu beantworten, ist es erforderlich ein Gesamtverständnis über die Funktionsweise des U-Space-Dienstes zu gewinnen, welches die prozesstechnischen Abläufe, die benötigten Informationsgrundlagen und die

zu erfüllenden Funktionen umfasst. Die Wichtigkeit der Ausgestaltung wurde auch in anderen Projekten, wie beispielsweise in dem Schweizer SUSI-Projekt, festgestellt [9].

3. Anforderungen an einen UAS-Fluggenehmigungsdienst gem. U-Space-Verordnung

Zunächst ist es notwendig, die Mindestanforderungen zu identifizieren, die sich aus der U-Space-Verordnung und den darin referenzierten rechtlichen Rahmenbedingungen ergeben. Die Mindestanforderungen aus rechtlicher Sicht an den U-FAS lassen sich aus den elf Absätzen des Artikel 10 der U-Space-Verordnung ableiten [5]:

1. U-Space Service Provider (USSP) müssen einen UAS-Fluggenehmigungsdienst (U-FAS) anbieten und UAS-Fluggenehmigungen für jeden einzelnen Flug erteilen.
2. Sendet der UAS-Betreiber eine UAS-Fluggenehmigung an den USSP, muss dieser:
 - a. die eingereichte UAS-Fluggenehmigungsanfrage auf Vollständigkeit, Korrektheit und auf Übereinstimmung mit Annex IV der U-Space-Verordnung überprüfen,
 - b. den Antrag auf eine UAS-Fluggenehmigung akzeptieren, sofern keine Konflikte mit anderen UAS-Flügen innerhalb desselben U-Space-Luftraums identifiziert wurden,
 - c. dem UAS-Betreiber mitteilen, ob der Antrag angenommen oder abgelehnt wurde,
 - d. bei erfolgreicher Annahme der UAS-Fluggenehmigung den UAS-Betreiber benachrichtigen und die zulässigen Schwellenwerte für die Abweichung von der erteilten UAS-Fluggenehmigung angeben.
3. USSPs müssen bei der Erteilung einer UAS-Fluggenehmigung gegebenenfalls die vom Wetterinformationsdienst nach Artikel 12 bereitgestellten Wetterinformationen nutzen.
4. Musste die UAS-Fluggenehmigung seitens USSP abgelehnt werden, kann dieser dem UAS-Betreiber eine alternative UAS-Fluggenehmigung vorschlagen.
5. Bei Erhalt einer UAS-Flugaktivierungsanfrage nach Artikel 6 Absatz 5 muss der USSP die Aktivierung der UAS-Fluggenehmigung ohne ungerechtfertigte Verzögerung bestätigen.
6. USSPs müssen geeignete Vorkehrungen treffen, wenn Konflikte bei Anträgen von UAS-Fluggenehmigungen, die von den UAS-Betreibern bei verschiedenen USSPs eingereicht wurden, vorliegen.
7. USSPs müssen Luftraumbeschränkungen des U-Space und vorübergehende Begrenzungen des U-Space-Luftraums bei der Prüfung eines Antrags auf UAS-Fluggenehmigung berücksichtigen.
8. Die eingereichten UAS-Fluggenehmigungen müssen vom USSP auf Anträge überprüft werden, die nach Artikel 4 der Durchführungsverordnung (EU) Nr. 923/2012 besonderen Vorrang haben.

9. Zur Verteilung der Luftraumzuweisung bei gleichpriorisierten Anträgen auf Erteilung einer UAS-Fluggenehmigung wird das Windhundverfahren angewendet (first-come-first-served).
10. USSPs haben die Pflicht, bestehende UAS-Fluggenehmigungen kontinuierlich im Hinblick auf sich verändernde U-Space-Luftraum-Bedingungen zu überprüfen und sofern erforderlich, erteilte Genehmigungen entsprechend zu aktualisieren oder zurückzunehmen.
11. USSP müssen für jede erteilte UAS-Fluggenehmigung eine eindeutige Genehmigungsnummer vergeben, aus welcher der genehmigte Flug, der UAS-Betreiber und der USSP, welcher die UAS-Fluggenehmigung erteilt hat, identifiziert werden kann.

4. Konzept für eine automatisierte Fluggenehmigung und –freigabe für Drohnenflüge

Der UAS-Fluggenehmigungsdienst muss auf zwei unterschiedliche Ereignisse reagieren können: Zum einen auf Fluggenehmigungsanfragen, welche von einem UAS-Betreiber an den USSP bzw. den UAS-Fluggenehmigungsdienst gesendet werden (E1), und zum anderen auf statische und dynamische Änderungen im U-Space-Luftraum (E2), welche von zuständigen Behörden aktualisiert werden. Beide Ereignisse im U-FAS haben unterschiedliche Auslöser und unterschiedliche Informationen für den Prozessbeginn. Aus diesem Grund werden unterschiedliche Abläufe benötigt. Wie bereits in der Einleitung beschrieben, adressiert diese Veröffentlichung den Fluggenehmigungs- und Aktivierungsprozess.

Aufbau und Phasen des UAS-Fluggenehmigungs- und Aktivierungsprozesses

Der Fluggenehmigungs- und Aktivierungsprozess kann in zwei aufeinanderfolgende Phasen aufgeteilt werden. Phase 1 beschreibt hierbei den UAS-Fluggenehmigungsprozess und Phase 2 den UAS-Flugaktivierungsprozess (Freigabeprozess). Bild 2 stellt den vereinfachten sequenziellen Ablauf in der ersten Detaillierungsstufe des Prozessmodells dar.



Bild 2: UAS-Fluggenehmigungs- und UAS-Aktivierungsprozess

Wie in Bild 1 skizziert, können UAS-Flüge seitens des UAS-Betreibers langfristig, mittelfristig oder kurzfristig (prä-taktisch) geplant werden, wodurch die vorhandene Informationsgrundlage Einfluss auf den durchzuführenden Genehmigungsprozess von UAS-Fluggenehmigungsanfragen hat. Der UAS-Aktivierungsprozess hingegen erfolgt kurz vor Flugbeginn, wodurch zum

Zeitpunkt der Aktivierungsanfrage seitens UAS-Betreiber eine vollständige Informationsgrundlage angenommen werden kann (Vgl. Bild 1).

Phase 1: Der Fluggenehmigungsprozess

Phase 1 beschreibt den UAS-Fluggenehmigungsprozess vom Eingang einer UAS-Fluggenehmigungsanfrage seitens des UAS-Betreibers beim UAS-Fluggenehmigungsdiensts bis hin zu der erfolgreichen Genehmigung oder Zurückweisung. In diesem Teilprozess erfolgt die Prüfung, Konfliktdetektion und -lösung (Konfliktmanagement) sowie bei erfolgreicher Genehmigung die Ausstellung einer UAS-Fluggenehmigung in Form einer Genehmigungsnummer. Zudem ist zu beachten, dass in den ersten beiden Subphasen der Prozess beendet wird, sofern die UAS-Fluggenehmigungsanfrage abgelehnt werden muss. Es bietet sich deshalb an, den Fluggenehmigungsprozess in weitere Subphasen aufzuteilen und zu kapseln. Jede Subphase erfüllt dabei eine Teilmenge der Aufgaben des U-FAS (Bild 3).

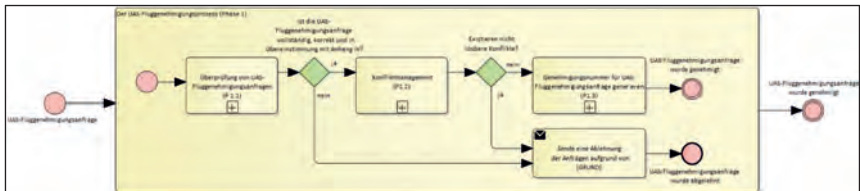
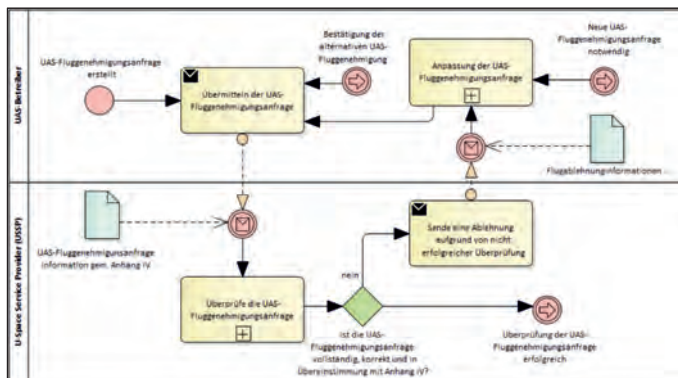


Bild 3: Der Fluggenehmigungsprozess (Phase1)

Die einzelnen Subphasen des Fluggenehmigungsprozesses, im weiteren Beitrag auch Module genannt, werden in der folgenden Tabelle weiter spezifiziert und erläutert, indem die Aufgaben eines U-FAS (Vgl. Kapitel 3) auf die Modulziele übertragen werden.

Tabelle 1: Modulziele und realisierte U-FAS Aufgaben beim UAS-Genehmigungsprozess

Phase	Modulziel	U-FAS Aufgaben
Phase 1.1	Mit dem Modul <i>Überprüfung von UAS-Fluggenehmigungsanfragen</i> wird in dem vorgestellten Konzept die Verpflichtung des USSPs zur Überprüfung eingereicherter UAS-Fluggenehmigungsanfragen erfüllt.	Artikel 10, Absätze: 1, 2 (speziell 2a und 2c) und Anhang IV der U-Space-Verordnung
Phase 1.2	Mit dem Modul <i>Konfliktmanagement</i> wird in dem vorgestellten Konzept die Verpflichtung eines USSPs zur Konfliktdetektion und -lösung unter Berücksichtigung der Prioritätsregeln erfüllt.	Artikel 10, Absätze: 2 (speziell 2b und 2c), 3 (o.), 4 (o.), 6, 7, 8 und 9
Phase 1.3	Mit dem Modul <i>Genehmigungsnummer für UAS-Fluggenehmigung generieren</i> wird die Verpflichtung des USSPs zur Erstellung einer Fluggenehmigungsnummer erfüllt.	Artikel 10 Absatz: 2 (speziell 2c und 2d) und 11



teilen und wie beim UAS-Genehmigungsprozess Module einzuführen. Jedes Modul erfüllt dabei eine Teilmenge der Aufgaben des U-FAS. Bild 5 zeigt die weitere Unterteilung in aufeinander folgende Module.

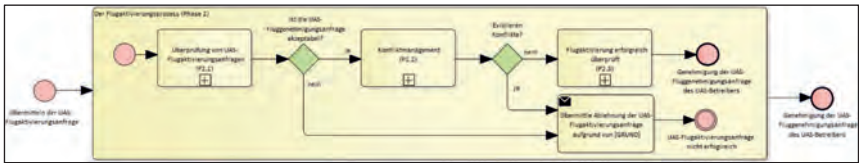


Bild 5: Aufbau und Phasen des Flugaktivierungsprozess

Die Module des UAS-Flugaktivierungsprozesses werden in der folgenden Tabelle weiter spezifiziert und erläutert, indem die Aufgaben eines U-FAS auf die Modulziele übertragen werden.

Tabelle 2: Modulziele und realisierte U-FAS Aufgaben beim UAS-Aktivierungsprozess

Phase	Modulziel	U-FAS Aufgaben
Phase 2.1	Mit dem Modul <i>Überprüfung von UAS-Aktivierungsanfragen</i> wird in dem vorgestellten Konzept die Verpflichtung des USSPs zur Überprüfung eingereicherter Flugaktivierungsanfragen erfüllt.	Artikel 10, Absätze: 5 und 6
Phase 2.2	Mit dem Modul <i>Konfliktmanagement</i> des Flugaktivierungsprozesses wird in dem vorgestellten Konzept die Konfliktdetektion und -lösung in der pre-taktischen Phase der strategischen Planung von Drohnenflügen realisiert.	Artikel 10, Absatz: 10 (optional)
Phase 2.3	Mit dem Modul <i>Flugaktivierung erfolgreich überprüft</i> des Flugaktivierungsprozesses wird in dem vorgestellten Konzept die Erstellung und Übermittlung der Flugaktivierungsgenehmigung seitens USSPs realisiert.	Artikel 10, Absätze: 5 und 6

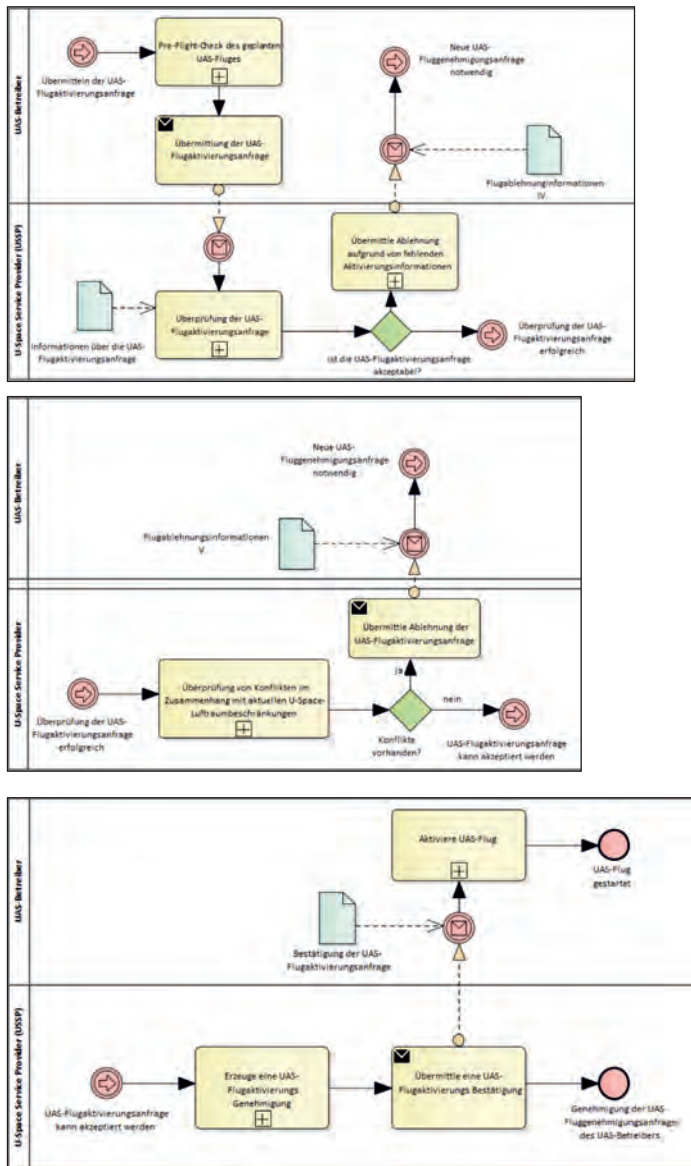


Bild 6: Detaillierte Phasenbetrachtung des UAS-Flugaktivierungsprozesses - Phase 2.1; Phase 2.2 und Phase 2.3 (von oben bis unten)

Allgemeine Herausforderungen und die Definition von Unschärfen der U-Space-VO

Die technischen Prozesse, die hinter der Inanspruchnahme der U-Space-Dienste zwischen UAS-Betreiber und USSP liegen, bedingen sich rechtlich nach den Vorgaben der U-Space-Verordnung, Details beinhaltet die U-Space-Verordnung hierzu allerdings nicht. Dies stellt die technische Umsetzung vor die Herausforderung, die rechtlich abstrakten Vorgaben oder teilweise nicht eindeutigen Bestimmungen (zusammengefasst als *Unschärfen*) zunächst zu konkretisieren und die Lösungsoptionen auf ihre Praktikabilität hin zu prüfen.

Ein Beispiel zur Verdeutlichung des Problems von Unschärfen, bietet der Vergleich der UAS-Fluggenehmigungsanfrage und der UAS-Flugaktivierungsanfrage eines UAS-Betreibers, indem die Informationsgrundlagen aus der U-Space-Verordnung beider Prozesse gegenübergestellt werden. Während sich bei der UAS-Fluggenehmigungsanfrage Mindestanforderungen an Informationen aus Artikel 10 und dem Anhang IV der U-Space-Verordnung ableiten lassen, können für die UAS-Flugaktivierungsanfrage keine Informationen oder Indizien daraus entnommen werden. Diese fehlende Spezifizierung führt dazu, dass der Interpretationsspielraum für die Realisierung des Aktivierungsprozesses sehr groß ist. Betrachtet man beispielsweise das Aktivierungsfenster, welches dem UAS-Betreiber nach erfolgreicher UAS-Aktivierungsanfrage zur Verfügung stehen kann, stellt man fest, dass die Ausgestaltung dieses Fensters einen großen Einfluss auf die gesamte strategische Planung von Drohnenflügen in einem U-Space haben kann. Ist das Aktivierungsfenster groß, kann der UAS-Betreiber deutlich vom eigentlich geplanten Flugstart abweichen, sodass früher oder später geplante Drohnenflüge davon betroffen sein können.

5. Auszug technischer Realisierungen von Teilmodulen des Konzepts

Im Rahmen des vom BMDV geförderten Verbundprojekts UDVeO (siehe <https://udveo.eu/>) wurden der Verordnungstext zum UAS-Fluggenehmigungsdienst auf Unschärfen untersucht und für die identifizierten Unschärfen Konkretisierungslösungen erarbeitet. Auf Basis der Konkretisierungslösungen wurde eine erste technische Umsetzung für den UAS-Aktivierungsprozess (Phase 2) umgesetzt.

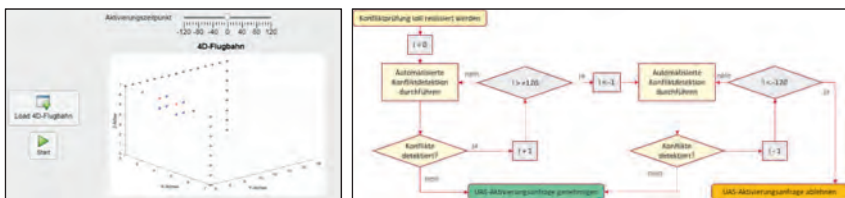


Bild 7: Manuelle Konfliktlösung (links), automatisierte Funktionsweise des Algorithmus (rechts)

Die Konfliktbehandlung beim Aktivierungsprozess kann anhand zweier Betriebsmodi untersucht werden, einem manuellen und einen automatisierten Betrieb. Ziel ist es jeweils, zu überprüfen, ob zum geplanten Flugstartzeitpunkt entsprechend dem UAS-Fluggenehmigungsantrag ein Konflikt vorliegt. Dies bedingt, dass der UAS-Betreiber bereits eine UAS-Flugaktivierungsanfrage an den U-FAS bzw. USSP gesendet hat und der UAS-Flug in den nächsten Minuten durchgeführt werden soll. Der Algorithmus prüft die bereitgestellte 4D-Flugbahn auf Konflikte mit anderen Konfliktelementen, auf Basis der zugrundeliegenden Informationsbasis. Wird ein Konflikt detektiert, kann mit Hilfe des Algorithmus die optimale Abweichung vom strategisch geplanten Flugstartzeitpunkt auf Basis der aktuellen U-Space-Lage ermittelt werden. Der Algorithmus prüft dabei, zu welchem Zeitpunkt der aktuell betrachtete UAS-Flugplan keine Auswirkungen (Konflikte) auf andere strategisch geplante und bereits genehmigte UAS-Flüge im System hat und keine anderen Konfliktelemente den betrachteten UAS-Flug selbst beeinträchtigen, indem er den geplanten Startzeitpunkt und die zugrundeliegende 4D-Flugbahn um jeweils 1 Sekunde nach vorne oder hinten verschiebt, bis zu einer maximalen Verschiebung von zwei Minuten $[-120 < t < 120]$. Konnte der Algorithmus einen Zeitpunkt identifizieren, an dem die 4D-Flugbahn realisiert werden kann, wird dem UAS-Betreiber ein UAS-Aktivierungszeitpunkt übermittelt und die Aktivierung freigegeben. Sofern im betrachteten Zeitfenster von 4 Minuten kein geeigneter Zeitpunkt gefunden werden, wird dem UAS-Betreiber keine UAS-Flugaktivierung genehmigt, sodass dieser den UAS-Flug nicht starten kann.

6. Zusammenfassung und Ausblick

In diesem Beitrag wird ein BPMN-basiertes Prozessmodell für einen automatisierten Fluggenehmigungs- und Aktivierungsprozess vorgestellt, welches die rechtlichen Mindestanforderungen aus der U-Space-Verordnung erfüllt und die Aufgaben dieses Teilbereichs eines U-FAS vollständig abdeckt. Durch die Modularisierung der Prozessschritte und somit der U-FAS Aufgaben kann eine schrittweise softwaretechnische Umsetzung realisiert werden. Zudem können dadurch einzelne Module im späteren Gesamtsystem leicht ausgetauscht werden, um beispielsweise verschiedene Konfliktdetektions-Algorithmen zu vergleichen und deren Einfluss auf den Gesamtprozess zu untersuchen.

Neben den Unschärfen gibt es noch weitere Schlüsselfaktoren, die sich auf die strategische Planungsphase von Drohnenflügen und somit auf die Ausgestaltung eines U-FAS auswirken können. So kann beispielsweise die in einem bestimmten U-Space-Luftraum gewählte Luftraumstruktur einzelne Prozessschritte bei der technischen Realisierung beeinflussen. Eine Luftraumgestaltung mit freiem Flugverkehr würde dabei weniger Beschränkungen mit sich

bringen als eine Korridorgestaltung. Darüber hinaus kann auch die Wahl eines anderen Prioritätsansatzes als das First Come First Serve Prinzip aus der U-Space-Verordnung einen großen Einfluss auf die Ausgestaltung der strategischen Planungsphase von Drohnenflügen sowie auf die Funktionsweise des Konfliktmanagements seitens USSP haben. Diese Forschungsgegenstände sollen im Weiteren näher untersucht werden.

Literaturverzeichnis

- [1] S. Törsleff et al.: Towards a U-space compliant UTM for urban airspaces. In: ITS WORLD CONGRESS, 11-15 October 2021, Hamburg.
- [2] Verband Unbemannte Luftfahrt (VUL). Analysis of the German Drone Market, 2019.
- [3] J. Rios: UAS Traffic Management (UTM) Project Strategic Deconflictions: System Requirements Final Report 2018. NASA.
- [4] SESAR JU: U-Space Initial Services – Study Report. Technical Report, SESAR Joint Undertaking, Edition 01.02.00, 2020.
- [5] CORUS: U-space Concept of Operations volume 2. Technical Report, SESAR Joint Undertaking, Edition 3.00.02, 2019.
- [6] Commission Implementing Regulation (EU) 2021/664 of 22 April 2021 on a regulatory framework for the U-space.
- [7] T. Grebner et al.: Service and Capability-based Conflict Management Framework for U-space. In: ITS WORLD CONGRESS, 11-15 October 2021, Hamburg
- [8] CORUS: Final Architecture and Solutions. Technical Report, SESAR Joint Undertaking, Edition 1.00.00, 2019, pp 35.
- [9] Amanda Boekholt, Benoit Curdy: UAS Flight Authorization Operator Experience Journey, SUSI Swiss U-Space Implementation, 2021

Erweiterte Risikobeurteilung für Maschinen und Betriebsmittel mit künstlicher Intelligenz

Eine Methode zur systematischen Erweiterung der Risikobeurteilung

Dipl.-Ing. **Werner Varro**, TÜV SÜD Product Service GmbH, München

Kurzfassung

In Maschinen und elektrischen Betriebsmittel werden vermehrt SW-Lösungen implementiert, um einerseits der Digitalisierung Rechnung zu tragen und andererseits die Potentiale in der Fertigung zu heben, die mit Hilfe der Digitalisierung erst möglich werden. Dadurch werden aber die Grenzen der heutigen Maschinen erweitert und somit auch die möglichen Gefährdungen und Risiken. Um diese bereits in der Konstruktionsphase zu erfassen, ist eine Erweiterung der Risikobeurteilung nach ISO 12100, wie sie heutzutage üblicherweise durchgeführt wird, notwendig.

Aufbauend auf der ISO 12100 wird im Folgenden eine Möglichkeit der „Erweiterten Risikobeurteilung“ (ERA) vorgestellt, um die möglichen, zusätzlichen Gefährdungen und Risiken systematisch zu erfassen sowie Maßnahmen zur Risikominderung zu definieren, implementieren und dokumentieren. Die „erweiterte Risikobeurteilung“ ist geeignet, die Anforderungen der neuen Maschineverordnung gerecht zu werden und ein Beitrag zur Erhöhung der Transparenz solcher komplexen Systeme und dadurch auch deren Vertrauenswürdigkeit und Akzeptanz!

Maschinen mit künstlicher Intelligenz

Künstliche Intelligenz (KI) ist einer der Trends im Maschinenbau und wird unser Leben und Arbeiten massiv verändern, sagen die einen, KI ist ein Hype, der seine Versprechungen nicht erfüllen wird, sagen die anderen. Wieder andere halten den Einsatz von KI für gefährlich und fordern dringend eine Regulierung. Doch was muss beachtet werden, wenn KI in Maschinen eingesetzt wird?

Es gibt bis heute keine allgemeingültige Definition von KI, nur Vorschläge, um den Begriff zu umschreiben. Beispiele:

„Problemlösungsmethoden, darunter Logik und Planungsverfahren, für die menschliche Intelligenz erforderlich wäre“. Fraunhofer-Institut für Produktionstechnik und Automatisierung:

„KI ist der Versuch, menschliches Denken auf Computer zu übertragen, sie soll eigenständig Antworten auf unterschiedliche Probleme finden.“ <https://kirevolution.com>

Unabhängig von obigen Definitionen, ist die „*künstliche Intelligenz*“ ein vom Menschen gemachtes System, dass die menschliche Intelligenz nachahmt, aber auch deren Beschränkungen ausgesetzt ist. D.h. im engeren Sinne, KI ist eine Software, die von Programmierern erstellt wurde, auf einer Hardware (Steuerung, Computer oder Netzwerk von Computern) läuft, um mit Hilfe von Regeln oder Algorithmen bestimmte Aufgaben zu lösen. Deswegen kann die KI nur im Zusammenhang mit der gestellten Aufgabe und im Verbund mit der Ziel-Hardware, z.B. einer Maschine bzw. Maschinensteuerung bewertet werden. In diesem Kontext darf es folgende Aussagen nicht geben:

- „Das kann ich mir nicht vorstellen!“
- „Das kann nie passieren!“

Anforderungen an Maschinen mit KI

Auch für Maschinen mit KI gelten grundsätzlich dieselben gesetzlichen Anforderungen zur Inverkehrbringung von Maschinen wie für Maschinen ohne KI. In der Europäischen Gemeinschaft müssen die Hersteller zum Inverkehrbringen einer Maschine die technischen Regeln und Gesetze beachten, die auf eine Maschine zutreffen. Das sind die relevanten EG-Richtlinien, ihre nationalen Umsetzungen (Gesetze) und die harmonisierten Normen (freiwillig), von denen die „Konformitätsvermutung“ ausgeht. Eine Übersicht über die mindestens zutreffenden Richtlinien und deren nationale Umsetzung in Deutschland zeigt Tabelle 1

Tabelle 1: Zutreffenden EG-Richtlinien und deren nationale Umsetzung in Deutschland zeigt:

EG-Maschinenrichtlinie	2006/42/EG	9.Verordnung zum ProdSG
EMV-Richtlinie	2014/30/EG	EMV-Gesetz EMVG
Funkanlagen-Richtlinie	2014/53/EG	Funkanlagen gesetz FuAG
...ggf. weitere spezifische Produktrichtlinien		
EG-Verordnung zur Cybersicherheit	881/2019	IT-Sicherheitsgesetz IT-SiG
Rechtsrahmens für Künstliche Intelligenz (KI)		

Die Richtlinien und Gesetze legen die Verantwortung sowie die Aufgaben eines Herstellers von Maschinen (Produkte allgemein) fest. Gemäß Maschinenrichtlinie (Umsetzung in D in der 9. ProdSGV) muss der Hersteller:

- Eine Risikobeurteilung erstellen (Abschnitt 1.4)
- die Anforderungen des Anhanges I der MRL erfüllen (Abschnitt 1.5)
- Eine technische Dokumentation erstellen (Abschnitt 2.1)
- Die Betriebsanleitung/Einbauanleitung erstellen und der Maschine beifügen (Abschnitt 2.2)
- Das Konformitätsbewertungsverfahren durchführen (Abschnitt 2.3)
- Die Konformitäts- bzw. Einbauerklärung erstellen und der Maschine beifügen (Abschnitt 2.4)
- die CE-Kennzeichnung sichtbar, lesbar und dauerhaft an die Maschine anbringen

Während der gesamten Lebensdauer darf von der Maschine kein höheres Risiko, als das im Rahmen der Risikobeurteilung als akzeptabel ermittelte Risiko, ausgehen. Der Hersteller muss also in der Lage sein, die von seinem Produkt ausgehenden Risiken zu beurteilen, einschließlich der Risiken, die von der KI ausgehen. Wenn über den Lebenszyklus einer Maschine die verwendete KI deren Funktionen steuert, ändert, verändert oder sogar weiterentwickelt, müssen diese Einflüsse in der Konformitätsbewertung des Herstellers und darüber hinaus in der Gefährdungsbeurteilung des Benutzers berücksichtigt werden. Beide werden dadurch komplexer und anspruchsvoller, da die Grenzen zwischen Hersteller und Benutzer zunehmend unschärfer werden!

Die klassische Risikobeurteilung auf Basis der ISO 12100 berücksichtigt die spezifischen Anforderungen einer sicheren KI (noch) nicht. Es ist aber wichtig, bereits in der Planungs- und Entwurfsphase den Einfluss der KI auf die möglichen Gefährdungen und Risiken zu untersuchen und ggf. Bewertungskriterien in die Risikobeurteilung einzufügen. Denn Sicherheit der KI

bedeutet zum einen die sichere und zuverlässige KI in der Anwendung, zum anderen aber auch eine vor Cyber-Angriffen geschützte KI. Nicht nur die Trainingsdaten oder Modelle können unzureichend sein, diese können ausgespäht oder sogar manipuliert werden. Deswegen muss in die Gesamtbetrachtung die gesamte beteiligte Infrastruktur und sogar die Cloud entsprechend geschützt sein. Dies kann nur mit einem sehr detaillierten Sicherheitskonzept, für solche Systeme erreicht werden.

Auch wenn entsprechend den bestehenden gesetzlichen Vorschriften und Normen zur Maschinensicherheit jegliche Manipulation durch Dritte, z.B. Cyber-Angriffe als krimineller Akt eingestuft wird, der nicht unter die EU-Maschinenrichtlinie 2006/42/EG bzw. unter die Gestaltungsleitsätze der EN ISO 12100 fällt, sollten die Hersteller von Maschinen aufgrund einer vorgesehenen Vernetzung mögliche Auswirkungen von Cyber-Angriffen auf die Maschinensicherheit berücksichtigen.

Obwohl in diesem Vortrag Fragen der KI-Ethik und der Cyber-Sicherheit nicht erläutert werden, sind diese Teil der erweiterten Risikobeurteilung!

Schritte nach 2006/42/EG, Anhang I:

Bei dem iterativen Verfahren der Risikobeurteilung und Risikominderung hat der Hersteller oder sein Bevollmächtigter:

- die Grenzen der Maschine zu bestimmen, wie:
 - o bestimmungsgemäße Verwendung
 - o vorhersehbare Fehlanwendungen
 - o Räumliche Grenzen
 - o Zeitliche Grenzen
- die Gefährdungen und Gefährdungssituationen zu ermitteln;
- die Risiken abzuschätzen (Verletzungsschwere und Wahrscheinlichkeit)
- die Risiken zu bewerten (Risikominderung nötig/nicht nötig?)
- die Gefährdungen auszuschalten oder durch Anwendung von Schutzmaßnahmen die Risiken zu mindern.

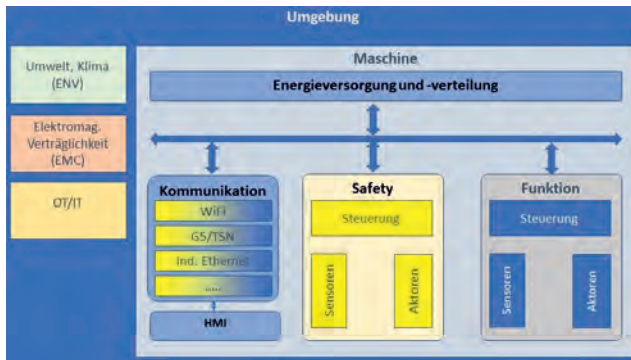


Bild 1: Funktionsdiagramm einer Maschine

Begründung für die erweiterte Risikobeurteilung

Die große Stärke der KI kann in Maschine überall dort ausgenutzt werden, wo große strukturierte oder unstrukturierte Datenmengen anfallen und auszuwerten sind. KI und maschinelles Lernen ermöglicht es technischen Systemen ihre Umwelt wahrzunehmen und durch das Erkennen von Mustern das Wahrgenommene zu verarbeiten, um daraus ein Modell für eine spezifische Aufgabe zu erstellen. Auf Basis solcher Modelle können dann Vorgehensweise, Abläufe oder Informationen erstellt werden. Durch weitere Daten im laufenden Betrieb kann ein KI-System seine Leistung im Hinblick auf eine spezifische Aufgabe kontinuierlich verbessern. Typische Anwendungen der KI in Maschinen sind: erkennen akustischer Informationen (z.B. Sprache oder Störgeräusche), Bilder oder sogar dreidimensionale Gegenstände erkennen, visuelle Informationen zu analysieren („Machine Vision“), eigenständige Orientierung im Raum (mobile Roboter, FTF, Stapler,...), erkennen von Handschriften und Umwandlung in digitale Texte, schnelle Erfassung komplexer Zusammenhänge und Erstellung von Prognosen, Plausibilitätsscheck einer Konstruktion oder eines Schaltplans durchführen, virtuelle Umgebungen generieren und vieles mehr.

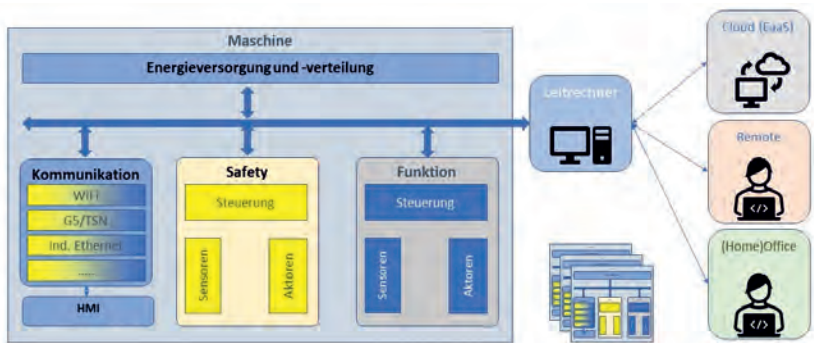


Bild 2: erweitertes Funktionsdiagramm einer IND 4.0 Maschine

Die Entwickler solcher komplexen Systeme auf Basis von KI können oft das gelernte Verhalten dieser Systeme nicht zufriedenstellend erklären. Dies ist jedoch im Sinne der Rückführbarkeit und Rückverfolgbarkeit von Fehlern auf ihre Ursachen wichtig. Auch lässt sich die Korrektheit des erlernten Modells aus Trainingsdaten schwer erbringen, weshalb Fehler, die sich aus unvollständigen, fehlerhaften oder falschen Daten während der Betriebszeit ergeben, nicht sicher ausgeschlossen werden können. Im Sinne der MRL ist es unerlässlich, bereits während der Design- und Konzeptphase einer Maschine die von der KI ausgehenden Risiken zu betrachten und zu bewerten.

In der folgenden Betrachtung wird auf den Teilaspekt der Datensicherheit und Cyber-sicherheit nicht näher eingegangen. Diese ist jedoch eine Voraussetzung für zuverlässige KI-Systeme und kann in die Risikobeurteilung mit einbezogen werden

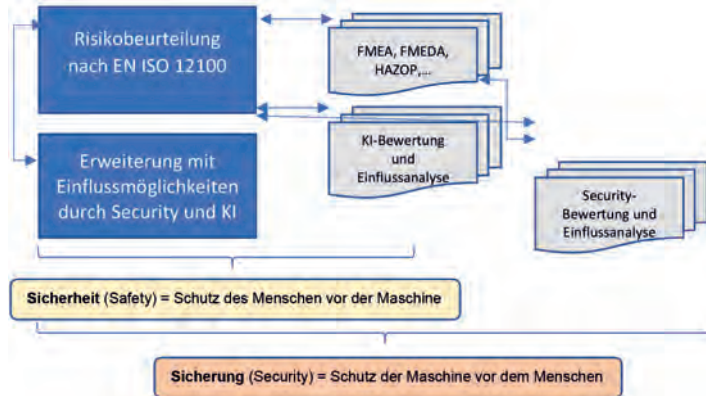


Bild 3: Erweiterte Risikobeurteilung

Neue Anforderungen durch KI

Wird in einer Maschine KI verwendet, können sich im Vergleich zu einer klassischen Lösung vor allem die Grenzen der Maschine verändern. Diese Veränderungen betreffen vorrangig die bestimmungsgemäße Verwendung der Maschinen durch den Einsatz von KI, die ggf. zu neuen Gefährdungen führen kann, z.B. durch:

- Veränderung des Verhaltens der Maschine, z.B. durch Algorithmen, Deep Learning, Machine Learning, SW-Updates, Fernzugriff, ... bis hin zu wesentlichen Änderungen
- Modulare Konfigurierbarkeit und Wandlungsfähigkeit von Maschinen, z.B. in der Matrixproduktion
- Zusätzliche Mensch-Maschine-Schnittstellen für mobile Geräte und Zugang über Internet oder Mobilfunknetz (Handy, Tablet, Notebook)
- Zusätzliche Betriebsarten bedingt durch die Verwendung und Schutz von Daten
- Bereitstellung von Komponenten und Applikationen in der Cloud, (z.B. SPS)
- Dezentralisierung der Steuerung
- IND 4.0-Fähigkeit (IIoT) und damit verbunden:
 - o Kommunikationsfähigkeit über die Betriebstechnik (Operational Technology, OT) hinaus, durch Bereitstellung von Schnittstellen
 - o Integrationsfähigkeit von und mit anderen Maschinen

- o Maschine wird zur Datenquelle und Datenträger
- o digitale Geschäftsmodelle auf Basis der Daten

Typische Fragen zur Funktionalität im Zusammenhang mit KI:

1. Welche Funktionen werden wo festgelegt und von wo angesteuert?
2. Entstehen durch die Anwendung der KI neue Betriebsarten?
3. Für welche Problemlösungen soll KI eingesetzt werden?

Durch diese Änderungen ergeben sich einige Fragen, die mit Hilfe der Risikobeurteilung systematisch bearbeitet und bewertet werden können – siehe Anhang „Checkliste zur KI in Maschinen“.

Die Veränderung der **bestimmungsgemäßen Verwendung** können auch die vorhersehbare Fehlanwendung beeinflussen. D.h. die oben genannten erweiterten Fähigkeiten der Maschinen können bei Mitarbeitern zu:

- Fehleinschätzungen durch den Menschen,
- falsche Anwendung
- zufällige oder absichtliche Manipulationen (Cybersicherheit)

führen.

Die **räumlichen Grenzen** der Maschine werden durch die neuen Funktionen, Einführung digitaler Geschäftsmodelle sowie der zunehmenden Modularität der Maschine (siehe auch Bild 2) erweitert. Einzelne HW- und SW - Komponenten müssen nicht mehr in der Maschine als abgeschlossenem physikalischen Gebilde bereitgestellt werden. Vielmehr ist deren Bereitstellung über Server, Cloud, etc. im Internet üblich und ermöglicht folgendes:

- die Kommunikationsfähigkeit, Konnektivität und Interoperabilität
- dezentrale Speicherung, Bearbeitung und Auswertung von Daten
- gemeinsame Nutzung der Daten durch Hersteller, Integrator, Betreiber und Dienstleister innerhalb eines digitalen Geschäftsmodells
- autonome Integration anderer Komponenten bzw. eigene Integration in Systeme.

Damit zusammen hängt auch die Klärung folgender Fragen ab:

1. Über welchen Raum - physikalisch und virtuell - erstreckt sich die Maschine?
2. Welche Schnittstellen stehen für welche Anwendungen bereit?
3. Welche Systeme und Komponenten können oder dürfen mit der Maschine dynamisch verbunden werden?

Die **zeitlichen Grenzen** können sich aus folgenden Gründen ändern:

- dynamische Änderung der Maschine auf Grund der Modularisierung
- maschinelles lernen während des gesamten Lebenszyklus
- regelmäßige Bewertung aus Gründen der Cyber-Sicherheit
- SW-Updates und Upgrades

Eine Änderung der Grenzen ergibt sich auch, wenn der Lernprozess (ML, Deep Learning,...) nicht einmalig vor der Bereitstellung der Maschine durch den Hersteller, sondern während dem gesamten Lebenszyklus, vor allem in der Betriebsphase erfolgt. Werden die Daten aus laufendem Betrieb für ML verwendet, geht einerseits die Verantwortung des Herstellers über die Inbetriebnahme und Bereitstellung der Maschine hinaus, andererseits gehören Daten aus laufendem Betrieb dem Betreiber. Deswegen muss bereits im Vorfeld die Nutzung von Daten und die Verantwortung dafür geklärt werden, wie z. B.:

- Wie kann die Maschine durch SW/KI verändert werden?
- Wo greift ML ein?
- Welche Prozesse und Abläufe werden optimiert?
- Welche validierten Daten werden für ML bereitgestellt?
- Können durch falsches Lernen, z.B. falsche Lernstrategie, Fehler entstehen?

Die oben beschriebenen Einflüsse der KI auf die Grenzen der Maschine können alle Lebensphasen und Betriebsarten der Maschine betreffen und zu unterschiedlichsten Gefährdungen für Menschen, Eigentum und Umwelt führen. Aufgabe der erweiterten Risikobeurteilung ist es, die zusätzlichen Gefährdungen und daraus resultierende Risiken zu betrachten, um entsprechende Maßnahmen zur Risikominderung definieren zu können, d.h.:

1. Risiken durch die Erweiterung der Grenzen der Maschine
2. Risiken durch die erhöhte Kommunikationsfähigkeit der Maschine
3. Risiken durch die Verwendung von Maschinendaten (KI, ML, Algorithmen zur Prozessoptimierung)
4. Risiken durch unbeabsichtigte Datenänderung sowie Manipulation
5. Risiken, die durch das Zusammenwirken von Mensch und digitalem Systemen entstehen.

Erweiterung der Risikobeurteilung

In Bild 4 ist der Ablauf einer Risikobeurteilung und die notwendigen Erweiterungen dargestellt.



Bild 4: Vorschlag zur Erweiterung der Risikobeurteilung

Konzept der erweiterten Risikobeurteilung

Tabelle 2: Konzept der erweiterten Risikobeurteilung

Iterativer Prozess nach ISO 12100	Erweiterte Risikobeurteilung	Mögliche Vorgehensweise zur Risikominderung
Grenzen der Maschine	Erweiterung der Grenzen um • Modulare Konfigurierbarkeit • Mobilgeräte • Kommunikation • Daten in der Cloud soziale Netzwerke (Data Leaks)	Analyse der Hardware- und Softwarespezifikation und -architektur
Identifizierung der Gefährdungen, Gefährdungsereignisse und -situationen	Gefährdungen durch: • Kommunikation & Konnektivität • Identifizierbarkeit • Interoperabilität • Modularität • Datenspeicherung/ -bearbeitung • Datenmanipulation	Erstellung einer System-FMEA Kommunikationstrategie KI-Strategie und Konzept Anfälligkeitsanalyse

Iterativer Prozess nach ISO 12100	Erweiterte Risikobeurteilung	Mögliche Vorgehensweise zur Risikominderung
Einschätzung der Risiken	Auswirkungen betrachten auf • Komponenten • Maschinen • Software • Kommunikation • Kollaboration	Kommunikationsplan Datenklassifizierung Security-Analyse
Bewerten der Risiken	Klassifizierung nach Schadensausmaß und Eintrittshäufigkeit Klassifizierung der Daten Klassifizierung der KI Bedrohungsszenarien	ISO 12100, ISO 13849-1
Risikominderung	• durch Konstruktion und Design • Security by Design • technische Maßnahmen • Benutzerinformationen	Daten-Minimierung, Auswahl der Systeme und der Mitarbeiter, Monitoring Verschlüsselung und Trennung Änderungs-, Konfigurations- und Updatemanagement

Es müssen Kontrollmechanismen entwickelt werden, die ausschließen, dass eine KI zu einer Bedrohung für den Maschinenbediener führen kann.

Tabelle 3: Einfluss der KI auf die Gefährdungen einer Maschine

Erweiterung der Gefährdungen nach ISO 12100	Möglicher Einfluss der KI auf	Zu betrachtende mögliche Gefährdungen
Gefährdungen durch die Verwendung von KI	Funktionen, die von der KI gesteuerte, geänderte verändert, weiterentwickelt werden Absichern vor Ausfälle der KI Cyberangriffen Datenschutz und Datensicherheit	Ziele und Strategie der KI Ausschlüsse Beeinflussung durch KI Freigabe durch Bediener Stop durch Bediener
Mechanische Gefährdungen	Abstand, Geschwindigkeit, Beschleunigung, Drehzahl, Richtung, Druck	Kollision Quetschen Schneiden
Elektrische Gefährdungen	Spannung, Leistung, Frequenz	Unerwarteter Anlauf, Start/Stop, Not-Halt
Thermische Gefährdungen	(Prozess-) Temperatur	Verbrennungen, Brand, Explosion
Gefährdungen durch Lärm	Erkennung von Sprachbefehle Erkennung von Anomalien	Fehlinterpretation, Fehlfunktion, unerwartetes Verhalten, Kollision
Gefährdungen durch Vibration	Erkennung von Anomalien	Fehlinterpretation, Fehlfunktion
Gefährdungen durch Strahlung	erhöhten Strahlung, Expositionsdauer, erhöhter Datenverkehr	Exposition (Laser, UV, Röntgen, EMF)
Gefährdungen durch Materialien und Substanzen	Herausschleudern, spritzen	
Ergonomische Gefährdungen	Gestaltung der HMI, Darstellung auf unterschiedlichen Geräten intuitive Bedienung, Fehlinterpretationen	
Gefährdungen im Zusammenhang mit der Einsatzumgebung	Freisetzen von Substanzen Materialverbrauch	

Erweiterung der Gefährdungen nach ISO 12100	Möglicher Einfluss der KI auf	Zu betrachtende mögliche Gefährdungen
Kombination von Gefährdungen	Kombination Gefährdungen durch KI und Security	
Gefährdung durch Steuerungen	Dynamische Änderungen von Parametern, Abläufen und Prozessen, Konnektivität, Interoperabilität	
Gefährdungen durch Kommunikation	HMI – Vernetzung, Fernbedienung, Fernwartung MMI - Vernetzung von Maschinen, Bauteilen, Transportsystemen, Gebäudefunktionen usw. i.S.d. Internet of Things (IoT) ist nicht frei von Risiken	
Gefährdungen durch die Verwendung und Änderung von Daten	Fälschungs- und Manipulations-sicherheit der Daten, Signalfälschung, Manipulation von RFIDs um Maschinen fehlzusteuern und Produktionsprozesse zu sabotieren	Möglichkeit zum schnellen Eingreifen, Notlösung zum Überstimmen der intelligenten Steuerlogik

Nutzung der Daten

Abhängig vom digitalem Geschäftsmodell kann es mehrere Nutzer der generierten Maschinendaten geben. Das können einerseits die Hersteller von Komponenten, Teilmaschinen oder Maschinen, aber auch der Integrator, Betreiber oder Dienstleister für digitale Geschäftsmodelle sein. Je nach Nutzer und Verwendung der Daten ergeben sich damit unterschiedliche Zugriffsrechte und Verantwortlichkeiten. Diese können sich auf die Gefährdungen und Risiken der Maschine auswirken. Siehe auch [11 und 12]

Die EU-Agentur für Cybersicherheit ENISA hat Bedrohungen im Zusammenhang mit KI analysiert, (siehe [13]) und eine Risikoanalyse über den ganzen Lebenszyklus einer KI-Lösung vorgeschlagen, der folgendes beinhaltet:

1. Datenebenen, von den Rohdaten bis zu den ausgewerteten und angereicherten Daten durch KI für jede Phase des Lebenszyklus hinsichtlich Schutz der Daten gegen Manipulationen, Ausspähung und Zerstörung bewertet.

2. Die Rollen aller beteiligten Personen und Nutzer von Daten definieren (Designer, Entwickler, Data Scientist, Data und Model Ownern, KI-Service Provider, Cloud-Provider).
3. Alle Prozess-Schritte vom Design über das Training bis hin zur Nutzung und Wartung auf Risiken hin untersuchen
4. genutzte Tools, Plattformen, Modelle und die Entwicklungsumgebungen in die Schwachstellenanalyse einbeziehen.

Mögliche Bedrohungen (ENISA): Missbrauch, das Abhören und Abfangen, Attacken auf Infrastrukturen wie KI-Hardware, unabsichtlichen Schaden, Fehler / Fehlfunktionen, wobei hier nochmals zahlreiche, verschiedene Typen von Angriffen, Störungen und Fehlern zu berücksichtigen sind.

Klassifizierung der KI

Ein KI-System kann auf Basis des Gefahrenpotenzials bezogen auf Gesundheit, Sicherheit und die Grundrechte einer Person, einer von vier Risikostufen zugeordnet werden:

- minimales Risiko – keine Anforderungen
- begrenztes Risiko – Anforderungen an Dokumentation und Transparenz
- hohes Risiko - Anforderungen an Transparenz und angemessene Information der Anwender notwendig
- inakzeptables Risiko – für Maschinen, ähnlich der Systematik der EN ISO 12100 werden Systeme mit inakzeptablem Risiko verboten.

Eine Klassifizierung und Entscheidung über Schutzmaßnahmen setzt Verständnis über das KI-System und dessen Grenzen voraus, vor allem durch:

- Transparenz (z.B. durch Einführung von Schwellwerten für die Genauigkeit, mit der ein KI-System entscheidet)
- menschliche Kontrolle durch einfache regelbasierten Verfahren oder „Explainable AI“ um die Entscheidungen des Systems zu erklären.

Verhältnis zwischen Maschinensicherheit und Cybersecurity

Cyber-Angriffe stellen keine zusätzliche Gefährdung im Sinne der Maschinensicherheit (und damit auch der EU-Maschinenrichtlinie) dar, sondern sind lediglich ein Auslöser für Beeinträchtigungen bzw. außer Funktion setzen von Sicherheitsmaßnahmen. Daher bietet es sich an, zunächst die Risikobeurteilung zur Maschinensicherheit (EN ISO 12100) durchzuführen, und danach detaillierte Überlegungen zur Cyber-Sicherheit anzustellen.

Intelligente Maschinen und digitale Ethik

Beim Einsatz von KI stellen sich neben technischen, auch rechtlichen und ethischen Fragen, wie z. B.:

- Kann ein künstliches Denken Verantwortung tragen?
- Kann eine künstliche Intelligenz moralischen Werten unterliegen?
- Nach welchen Kriterien entscheidet das System z.B. in einer unausweichlichen Unfallsituationen, wessen Tod oder Verletzung das kleinere Übel ist?

Dabei geht es um Themenfelder wie Autonomie und Kontrolle, aber auch Fairness, Transparenz, Verlässlichkeit, Sicherheit und Datenschutz der KI.

Ausblick

Sowohl im Entwurf der neuen Maschinenverordnung als auch im Entwurf zur Verordnung über Künstliche Intelligenz sind grundlegenden Sicherheits- und Gesundheitsschutzziele technologieunabhängig formuliert. Spezifische Festlegungen zur Erfüllung dieser Schutzziele erfolgen in den harmonisierten Normen (z. Zt. in Arbeit) unter den beiden künftigen Verordnungen. Für die Anwendung von Verlaufsstrategien für sicherheitsrelevante Steuerungsfunktionen gelten demnach folgende Voraussetzungen:

- Sie dürfen nicht zu einem Überschreiten der Grenzen führen, die in der Risikobeurteilung festgelegt sind
- Sie müssen die in der Risikobeurteilung festgelegten Zuverlässigkeitswerte (SIL, PL) der jeweiligen Sicherheitsfunktion erreichen.

Deswegen sind Methoden erforderlich, um im Rahmen der Risikobeurteilung die Zuverlässigkeit der sicherheitsrelevanten Steuerungsfunktionen bewerten zu können. Einige Anforderungen an die Software für Sicherheitsfunktionen sind in den Normen der Reihe DIN EN 61508, in der DIN EN 62061, sowie im Entwurf der DIN EN ISO 13849-1 festgelegt, wobei noch geprüft werden muss, ob die Anforderungen für KI ausreichen.

Zukünftig werden sich auch die gesetzliche IT-Sicherheitsanforderungen für smarte Produkte und Maschinen durchsetzen. Zu nennen sind vor allem:

- der **EU-Cybersecurity Act**, der einen Rahmen für freiwillige Zertifizierungen vorsieht
- die **EU-Warenverkaufsrichtlinie**, die eine Sicherheitsaktualisierungspflicht für Verbraucherprodukte mit digitalen Elementen vorsieht,
- das **IT-Sicherheitsgesetz 2.0** mit Schwerpunkt auf dem Verbraucherschutz, in dem die produktbezogenen Herstellerpflichten zur IT-Sicherheit festgelegt werden.

Bereits jetzt treffen den Hersteller von Maschinen und elektrischen Betriebsmitteln einige IT-sicherheitsbezogene Pflichten. Um Haftungsrisiken zu vermeiden, sollten Hersteller dringend ein Mindestmaß an IT-Sicherheit sicherstellen, z. B. durch:

- Anwendung bestehender technischer Normen in Bezug auf die IT-Sicherheit in der Konstruktion und Produktion (insb. Vorsehen von Maßnahmen zur Resilienz)
- Einbeziehung der IT-Sicherheit in die Risikobeurteilung, um eventuelle IT-Sicherheitslücken und Anfälligkeiten für Cyberangriffe aufzudecken
- Bereitstellung von Informationen hinsichtlich möglicher Cybergefahren und IT-Sicherheitslücken
- vertragliche Ausgestaltung der
 - o Verantwortlichkeiten zwischen Hersteller und Betreiber (ggf. nebst Abschluss eines Software-Pflegevertrags)
 - o Nutzung von Daten, z.B. das Recht zur automatisierten Datenerhebung im Rahmen der Produktbeobachtung
 - o Durchführung von sicherheitsrelevanten Firmware-Updates
 - o konkreten Anforderungen an die IT-Sicherheit mit Zulieferern von Firmware
- Produktbeobachtung der im Feld befindlichen Maschinen auch unter dem Aspekt der IT-Sicherheit
- bereitstellen sicherheitsrelevanter SW-Updates und/oder Sicherheitswarnungen im Falle festgestellter Cybergefahren
- Klärung mit der Produkthaftpflichtversicherung, ob eine Gefahren aufgrund von IT-Sicherheitslücken gedeckt sind.

Regelungen und Standards

- [1] ISO/TR 22100-5 „Sicherheit von Maschinen – Zusammenhang mit ISO 12100 – Teil 5: Einfluss von künstlicher Intelligenz auf maschinelles Lernen“
- [2] ISO/IEC AWI TR 5469 „Artificial intelligence – Functional safety and AI systems“
- [3] KI im Entwurf für eine Maschinenverordnung vom 21.04.2021 vorgestellt (https://ec.europa.eu/commission/presscorner/detail/de/QANDA_21_1683#3)
- [4] Rechtsrahmen für künstliche Intelligenz: Leitlinien für die Entwicklung vertrauenswürdiger Künstlicher Intelligenz vorgelegt und im April 2021 zum – weltweit ersten –Vorschlag für einen Rechtsrahmen für KI (https://ec.europa.eu/growth/content/europe-fit-digital-age-commission-proposes-new-rules-and-actions-excellence-and-trust_de)

- [5] KI-Paket der EU-Kommission vom 21.04.2021 mit der Verordnung für einen Rechtsrahmen zur KI und der neuen Maschinenverordnung (https://ec.europa.eu/growth/content/europe-fit-digital-age-commission-proposes-new-rules-and-actions-excellence-and-trust_de)
- [6] Plattform Industrie 4.0: Netzwerk aus Unternehmen, Verbänden, Gewerkschaften, Wissenschaft und Politik mit dem Ziel, die digitale Transformation der Produktion in Deutschland voranzubringen
- [7] Deutsche Normungsroadmap Künstliche Intelligenz, ein Projekt des BMWi mit Experten aus Wirtschaft, Wissenschaft und Gesellschaft mit dem Ziel, Handlungsempfehlungen für die Normung rund um KI zu liefern
- [8] KI in der Industrie 4.0: Orientierung, Anwendungsbeispiele, Handlungsempfehlungen (BMWi 2020)
- [9] Industrial Digital Twin Association (IDTA): Vereinigung von Unternehmen und Verbänden aus Maschinenbau und Elektroindustrie zur Förderung des digitalen Zwillings
- [10] AASX-Package-Explorer, ein Open-Source-Tool zum Erstellen einer Verwaltungsschale und zum Modellieren eines digitalen Zwillings
- [11] Whitepaper des TÜV AI LAB: Vorschlag für eine Risikoklassifizierung von KI-Systemen, www.tuev-lab.ai
- [12] Wertschöpfung aus Maschinendaten 4.0 - Die Praxisanleitung für den Mittelstand, VDMA-Verlag
- [13] ENISA Bericht: Artificial Intelligence Cybersecurity Challenges, <https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges>

Anhang: Fragen im Zusammenhang mit KI

Anforderung
Basisanforderungen
Sind die Ziele der KI für das System definiert?
Ist die Strategie der KI dargelegt?
Gibt es eine Spezifikation und Begründung für die Verwendung von KI? (z.B. Verwendung unstrukturierter Daten, große Datenmengen, bessere Vorhersagbarkeit, beste Problemlösung)
Welche Entscheidungen soll die KI treffen?
Welche Entscheidungen darf die KI nicht treffen?
Soll die KI vom Menschen überstimmbar oder abschaltbar sein?

Anforderung

Sind die Risiken einer möglichen böswilligen Verwendung, eines Missbrauchs oder einer unangemessenen Verwendung des KI-Systems bewertet?

Sind Risiken, Risikometriken und Risikostufen des KI-Systems in jedem spezifischen Anwendungsfall definiert?

Sind einen Prozess zur kontinuierlichen Messung und Bewertung von Risiken eingeführt?

Sind Endbenutzer und Betroffene über bestehende oder potenzielle Risiken informiert?

Sind getestete, ausfallsichere Fallback-Pläne definiert, um KI-Systemfehler jeglicher Herkunft zu beheben und Governance-Verfahren eingeführt, um sie auszulösen?

Sind ein geeignetes Verfahren für den Umgang mit den Fällen eingerichtet, in denen das KI-System Ergebnisse mit einem niedrigen Vertrauenswert liefert?

Gibt es einen klar definierten Prozess um das Erreichen der definierten KI-Ziele durch das System zu überwachen?

Anforderungen an Daten

Welche Daten generiert das System und wer hat Zugriff auf welche Daten?

(Komponentenhersteller, Maschinehersteller, Integrator, Eigentümer der Maschine (EaaS-Business Modell), Maschinenbetreiber, Dienstleister für Kommunikation oder Datenanalyse, Cloud-Betreiber,...)

Wie sind die Zugriffsrechte geregelt?

Ist die benötigte Datenmenge, Datenqualität, Datenzusammensetzung und Datenrelevanz für die Problemlösung definiert und validiert?

Sind die Daten klassifiziert? (z.B. unkritisch, kritisch für Funktion, vertraulich, streng geheim)

Sind Maßnahmen ergriffen, um sicherzustellen, dass die zur Entwicklung des KI-Systems verwendeten Daten (einschließlich Trainingsdaten) aktuell, von hoher Qualität, vollständig und repräsentativ für die Umgebung sind, in der das System eingesetzt wird?

Könnte eine geringe Genauigkeit des KI-Systems zu kritischen, feindlichen oder schädlichen Folgen führen?

Sind Maßnahmen ergriffen, um die Genauigkeit des KI-Systems zu überwachen und zu dokumentieren?

Kann der Betrieb des KI-Systems die Daten oder Annahmen, auf denen es trainiert wurde, ungültig machen und kann dies zu nachteiligen Auswirkungen führen?

Sind Prozesse eingerichtet, um sicherzustellen, dass die Genauigkeit des KI-Systems, die von Endbenutzern und/oder Probanden erwartet wird, ordnungsgemäß kommuniziert wird?²¹

Nutzt das KI-System kontinuierliches (Online-)Lernen?

Anforderung
Modelle
Wie sind die Modelle für maschinelles Lernen definiert?
Wie viele Modelle für maschinelles Lernen sind bereits in Verwendung?
Wofür werden diese Modelle verwendet?
Welche Daten liegen für das Modell vor?
Wie häufig werden neue Daten und Datenfelder hinzugefügt?
Sind die Daten, auf deren Grundlage das Modell Entscheidungen trifft, verifiziert und validiert?
Wird dem Kunden mitgeteilt, auf welcher Datenbasis Entscheidungen getroffen werden?
Verstoßen die Dateneingaben (direkt/indirekt) gegen Vorschriften (z. B. Datenschutzgesetz)?
Wie werden die Modellverzerrungen verringert?
Sind potenzielle negative Folgen zu erwarten, wenn das KI-System neue oder ungewöhnliche Methoden lernt, um bei seiner objektiven Funktion gut abzuschneiden?
Zuverlässigkeit
Wie soll der Nachweis der Zuverlässigkeit geführt werden? –
Wie soll der Nachweis der Resilienz geführt werden? (Robustheit gegenüber Veränderungen durch).
Sind Maßnahmen definiert (und umgesetzt), um die Integrität, Robustheit und Gesamtsicherheit des KI-Systems gegen potenzielle Angriffe während seines gesamten Lebenszyklus zu gewährleisten?
Beschreibung der Angriffsszenarien
Wird/wurde das System einem Red-Teaming/Pentest unterzogen?
Sind die Abhängigkeit der Entscheidungen eines kritischen KI-Systems von seinem stabilen und zuverlässigen Verhalten bewertet?
Sind die Zuverlässigkeits-/Testanforderungen auf die entsprechenden Stabilitäts- und Zuverlässigkeitsniveaus ausgerichtet?
Sind die Fehlertoleranz über z. B. ein dupliziertes System oder ein anderes paralleles System minimiert (KI-basiert oder „konventionell“)?
Gibt es eine Methode zur Bewertung von Änderungen und ggf. erneuten Bewertung des KI-Systems hinsichtlich Robustheit und Sicherheit?
Könnte das KI-System bei geringer Zuverlässigkeit und/oder Reproduzierbarkeit kritische, nachteilige oder schädliche Folgen (z. B. in Bezug auf die menschliche Sicherheit) haben?
Sind bestimmte Kontexte oder Bedingungen berücksichtigt worden, um die Reproduzierbarkeit sicherzustellen?

Anforderung

Sind Verifizierungs- und Validierungsmethoden und Dokumentation (z. B. Protokollierung) eingeführt, um verschiedene Aspekte der Zuverlässigkeit und Reproduzierbarkeit des KI-Systems zu bewerten und sicherzustellen?

Sind Prozesse zum Testen und Verifizieren der Zuverlässigkeit und Reproduzierbarkeit des KI-Systems klar dokumentiert und operationalisiert?

Sicherung

Kann das KI-System nachteilige, kritische oder schädliche Auswirkungen haben auf die menschliche oder gesellschaftliche Sicherheit im Falle von Risiken oder Bedrohungen wie: Konstruktions- oder technischen Fehler, Defekte, Ausfälle, Angriffe, Missbrauch, unsachgemäßer oder böswilliger Verwendung?

Sind die möglichen Bedrohungen für das KI-System und die möglichen Folgen identifiziert für: Konstruktionsfehler, technische Fehler, Umweltgefahren?

Ist das KI-System für Cybersicherheit zertifiziert (z. B. Cybersecurity Act der EU) oder erfüllt es bestimmte Sicherheitsstandards (IEC 62344)?

Wie anfällig ist das KI-System gegenüber Cyberangriffen?

Sind die mögliche Formen von Angriffen bewertet, für die das KI-System anfällig sein könnte?

Sind die verschiedenen Arten von Schwachstellen und potenzielle Einstiegspunkte für Angriffe berücksichtigt, wie z. B.: Datenverfälschung (d.h. Manipulation von Trainingsdaten), Modellumgehung (d. h. Klassifizierung der Daten nach dem Willen des Angreifers), Modellinversion (d.h. Modellparameter ableiten)

Sind die Endbenutzer über die Dauer der Sicherheitsabdeckung und Updates informiert?

Wie lang ist der erwartete Zeitrahmen, innerhalb dessen Sicherheitsupdates für das KI-System bereitgestellt werden?

Sind Sicherheitskritikalitätsstufen (z. B. in Bezug auf die menschliche Integrität) der möglichen Folgen von Fehlern oder Missbrauch des KI-Systems definiert?

Anwendung

Welche Daten/Maschinenparameter werden im System generiert, übertragen, geändert und haben Einfluss auf die Ablaufsteuerung der Maschine?

Welche Funktionen werden geändert?

Hidden Markov Models und Active Learning zur automatisierten Kennzeichnung von Batchphasen in der Prozessindustrie

Gregor Just, Dr. **Valentin Khaydarov**, Prof. Dr.-Ing. **Leon Urbas**,
Professur für Prozessleittechnik/AG Systemverfahrenstechnik,
Technische Universität Dresden;
Dr. **Benjamin Klöpper**, ABB;
Dr. **Franz David Bähner**, Bayer AG

Kurzfassung

Datengetriebene und hybride Modelle zur Analyse sowie Optimierung von Prozessen und somit zur Steigerung der Anlageneffektivität gewinnen in der Prozessindustrie zunehmend an größerer Bedeutung, wobei die Modellierung von Batchprozessen aufgrund ihrer Variabilität in Dauer und Sequenz der einzelnen Batchphasen eine große Herausforderung darstellt. In der Praxis wird die Modellierung zudem durch ungenaue oder gar fehlende Aufzeichnungen von Beginn und Ende einzelner Batchphasen erschwert. Um die große Menge an Prozessdaten mit zuverlässigen Labels (Start- und Endzeitpunkte der Phasen) zu versehen und darauf aufbauend die Modellierung der Batchprozesse zu ermöglichen, ist ein automatisierter Kennzeichnungsprozess notwendig. In einem vorhergehenden Beitrag wurde das überwachte Lernverfahren Random Forest und das unüberwachte Lernverfahren Change Point Detection vorgestellt (Merkelbach, Tan, Gärtler, Baehner, & Urbas, 2021). Dabei zeigt das überwachte Lernverfahren hinsichtlich der Interpretierbarkeit durch eine sachkundige Person eine bessere Klassifikationsleistung, benötigt eine ausreichend große Trainingsmenge mit gekennzeichneten Phasen, was in der industriellen Praxis eine Limitation der Methode darstellt. Außerdem ist das Verfahren sensitiv gegenüber Prozessvariabilität und fehlerbehafteten Daten. Um diese Probleme zu adressieren, wird in diesem Beitrag die Kombination von zwei Verfahren betrachtet: Hidden Markov Models (HMM) und Active Learning (AL). Die Untersuchungen ergeben, dass das HMM die Batchphasen sicherer erkennt, da auch die Reihenfolge vorhergehender Phasen berücksichtigt werden und das Verfahren robuster ist gegenüber Prozessstörungen. AL ermöglicht es, initial an einem kleinen Datensatz zu lernen und weitere Labels so zu ergänzen, dass mit möglichst geringem Kennzeichnungsaufwand ein gutes ML Modell erzeugt wird.

1. Einführung

Eine gute Datenbasis ist eine Grundlage für objektive Entscheidungen und Verbesserungen von industriellen Prozessen. Unterschiedlichste Ansätze wie statistische Prozesskontrolle, model-basierte Regelung und datengetriebene Modelle setzen eine gute Datensituation voraus.

In der chemischen Batchproduktion wird die Schaffung einer guten Datengrundlage durch die Flexibilität, Freiheitsgrade in der Produktionssequenz und den Einfluss der Anlagefahrer verkompliziert. Eine Analyse der Prozesse, beispielsweise mittels Methoden der statischen Prozesskontrolle, setzen verlässliche Information über die Start- und Endzeiten der einzelnen Prozessschritte voraus. Insbesondere für Anlagen mit einem geringen Automatisierungsgrad werden solche Daten, wenn überhaupt, nur manuell gesammelt und daher ungenau und lückenhaft. Oft müssen die Start- und Endzeiten anschließend anhand der Prozessdaten mühevoll bestimmt werden. Anders als bei vielen Bildklassifikationsprobleme oder natursprachlichen Daten kann die Aufgabe der Kennzeichnung nicht an eine große Zahl von Laien ausgelagert werden (Crowdsourcing), sondern erfordert wegen der anlagenspezifischen Daten verfahrenstechnische Expertise. Dieser erhebliche Aufwand steht oft einer Optimierung der Fahrweisen, der Rezepte und Anlagen entgegen.

In (Merkelbach, Tan, Gärtler, Baehner, & Urbas, 2021) wurde das Problem der Identifizierung von Start- und Endzeiten von Schritten oder Phasen in Batches als überwachtes Lernproblem (Klassifikation) oder als Change Point Detection Problem adressiert. Während der reine Klassifikationsansatz einen relative großen Trainingsdatensatz mit gekennzeichneten Daten voraussetzt, liefert die Change Point Detection nur mögliche Start- und Endzeiten, aber keine Information welcher Schritt oder Phase vorliegt. Außerdem ist das Tuning der Change Point Detection wie bei allen unüberwachten Lernverfahren kompliziert und erfordert erheblichen menschlichen Aufwand. Für die Beurteilung der Güte wird wiederum die Zeit einer sachkundigen Person benötigt.

In diesem Beitrag schlagen wir die Kombination von zwei Methoden vor, die die Unzulänglichkeiten des überwachten Lernansatzes adressieren und eine ausreichende Genauigkeit mit minimalen Trainingsdaten erreichen.

2. Kombination von Active Learning und probabilistischen graphische Modellen

Um die Notwendigkeit für einen relativ großen Trainingsdatensatz für das überwachte Lernverfahren zu vermeiden, untersuchen wir die Kombination von drei Ansätzen:

1. Zeitreihenklassifizierung als Verfahren um die Zugehörigkeit eines Zeitfensters zu einer Phase zu bestimmen. Dies entspricht im Wesentlichen den überwachten Verfahren (Merkelbach, Tan, Gärtler, Baehner, & Urbas, 2021).
2. Active Learning als Verfahren welches die Anzahl der benötigten Labels minimiert.
3. Probabilistische Graphische Modelle, um die Tatsache, dass die Anzahl der unterschiedlichen Phasensequenzen in Batchrezepten begrenzt sind, für den Lernprozess nutzen zu können.

Im Folgendem beschreiben wir zunächst kurz die drei Ansätze und anschließend ihre Kombination.

Zeitreihenklassifizierung

An dieser Stelle folgen wir der Definition für Zeitreihenklassifizierung von (Fawaz, Forestier, Weber, & Muller, 2019). Eine univariate Zeitreihe $X^u = [x_1, x_2, \dots, x_t]$ ist eine zeitlich geordnete Menge realwertiger Variablen, wobei die Anzahl t der Variablen der Länge der Sequenz entspricht. Eine multivariate Zeitreihe $X^m = [X_1^u, X_2^u, \dots, X_d^u]$ ist eine Menge univariate Zeitreihen, wobei die Länge der Zeitreihen üblicherweise identisch sind. Ein Datensatz $D = \{(X_1, Y_1), (X_2, Y_2), \dots, (X_s, Y_s)\}$ ist eine Sammlung von Paaren (X_i, Y_i) , wobei X_i eine uni- oder multivariate Zeitreihe und Y_i das entsprechende Klassenlabel ist. Die Zeitreihenklassifizierung erlernt das Mapping von Zeitreihen auf die entsprechenden Klassenlabel $X \rightarrow Y$.

Active Learning

Active Learning ist eine Variante des maschinellen Lernens, in dem der Lernalgorithmus interaktiv Labels für neue Datenpunkte bei einem Nutzer oder einer anderen Datenquelle anfragen kann (Settles, 2009). Die Grundidee ist, dass ein bereits teilweise trainiertes Modell Informationen enthält, die es ermöglichen, die Anfragen so zu stellen, dass die Verbesserung des Modells maximiert wird.

Die gebräuchlichste Form des Active Learnings ist das pool-basierte Active Learning (Settles, 2009). Hier werden die neuen Datenpunkte aus einem großen Pool ungelabelter Daten ausgewählt. Dieser Ansatz passt auch zum Problem der Erkennung von Schrittttransitionen in Batchprozessen, da in aller Regel umfangreiche historische Daten vorliegen.

Um die neuen Datenpunkte auszuwählen, kommt eine sogenannte Query-Funktion zum Einsatz. Die Query-Funktion nutzt ein Maß für den Informationsgehalt eines Datenpunkts, um die Auswahl durchzuführen. Die in der Praxis gebräuchlichste Querystrategie ist das Uncertainty Sampling (Reyes, Abdulrahman, & Ventura, 2018), bei dem die Datenpunkte ausgewählt werden, bei denen das Modell bzgl. des vorhergesagten Labels am unsichersten ist.

Probabilistische Graphische Modelle

Probabilistische graphische Modelle sind probabilistische Modelle, bei denen ein Graph aus Knoten und Kanten die bedingten Wahrscheinlichkeitsbeziehungen zwischen Zufallsvariablen beschreibt. Zu den probabilistischen graphischen Modellen zählen unter anderem Bayes'sche Netzwerke, Hidden Markov Modelle, und Markov Random Fields (Murphy, 2001). Insbesondere dynamische Bayes'sche Netzwerke, Hidden Markov Models und Conditional Random Fields (CRF) erlauben die Modellierung von Sequenzen und Übergangswahrscheinlichkeiten. Diese Fähigkeit qualifiziert diese Methoden für das gegebene Problem, da die relative Regelmäßigkeit der Sequenzen idealerweise von einem maschinellen Lernverfahren genutzt werden sollte. Die Information über die Abfolge der Schritte sollte helfen, eine bessere Klassifikationsgenauigkeit mit weniger Trainingsbeispielen zu erreichen.

In dem hier vorgeschlagenen Ansatz kommen Conditional Random Fields zum Einsatz. (Lafferty, McCallum, & Pereira, 2001) beschreiben ein CRF mit Beobachtung X und Zufallsvariablen Y wie folgt:

- $G = (V, E)$ ist ein Graph mit Knoten V und Kanten E , wobei $Y = (Y_v)_{v \in V}$, sodass Y den Knoten im Graphen G zugeordnet werden kann
- (X, Y) ist dann ein Conditional Random Field, wenn jede Variable in Y_v in Abhängigkeit von X die Markov Eigenschaft in Bezug auf den Graphen erfüllt, also die Wahrscheinlichkeit nur von seinen direkten Nachbarn abhängt: $P(Y_v | X_m \{Y_w: w \neq v\}) = P(Y_v | X, \{Y_w: w \sim v\})$ wobei $w \sim v$ bedeutet, dass eine Kante zwischen w und v besteht, die beiden Knoten also Nachbarn in G sind.

Im Allgemeinen ist die Inferenz auf CRFs nicht lösbar. Für einige spezielle Fälle ist die exakte Inferenz jedoch möglich, wenn der Graph beispielsweise eine Kette oder ein Baum ist. Dann liefern Message Passing Algorithmen eine genaue Lösung. In einer Kette kann wie beim HMM der Viterbi-Algorithmus (Viterbi, 1967) zum Einsatz kommen. Es handelt sich um ein dynamisches Programm, um die Maximum posteriori Wahrscheinlichkeit der wahrscheinlichsten Sequenz von nicht beobachtbaren Zuständen zu bestimmen. Die Parameter eines CRF werden üblicherweise mittels der Maximum Likelihood Methode gelernt (Wallach, 2004). Der Unterschied zwischen HMM und CRF liegt darin, dass die initiale Zustandsschätzung in einem HMM mithilfe einer diskreten Wahrscheinlichkeitsverteilung bestimmt wird, im CRF kommt diese Schätzung durch eine frei wählbare Feature-Funktion $f(x)$ zustande (Sutton, 2012). Das ermöglicht, wie z.B. in (Krähenbühl & Koltun, 2011), die Schätzung des initialen Zustands anhand eines Klassifikators, im angesprochenen Paper ein JointBoost-Klassifikator, in unserem

Anwendungsfall haben wir uns für einen Random Forest entschieden. Für kettenförmige Graphen wird das Modell auch als Linear Chain CRF bezeichnet.

Kombination der Ansätze

Bild 1 stellt die Kombination der Ansätze dar. Der innere Kasten mit der Beschriftung *Lerner* stellt das eigentliche Lernverfahren dar:

- Zunächst wird ein Zeitreihenklassifikator auf die multivariaten Signaldaten angewandt. Die Ausgabe ist die Entscheidungsfunktion F , die den Klassen jeweils einen Wert im Intervall $[0,1]$ zuordnet
- Da die Entscheidungsfunktion F keine tatsächliche Wahrscheinlichkeit darstellt, werden die Ausgaben kalibriert, um tatsächliche Klassen- bzw. Schrittwahrscheinlichkeiten zu erhalten. Ein übliches Verfahren für die Kalibrierung ist das so genannte Platt Scaling (Böken, 2021).
- Die Klassenwahrscheinlichkeiten gehen als Beobachtungen in das Conditional Random Field ein, das für jeden Zeitschritt die Übergangswahrscheinlichkeiten zwischen den Schritten im Batchrezept liefert.

Dieses eigentliche Lernverfahren ist in einen Active Learning Prozess eingebettet. Dieser Prozess startet mit einem kleinen Trainingsdatensatz, mit dem der Lerner initial trainiert wird. Die Query Funktion wendet den trainierten Lerner auf den Pool von ungelabelten Daten an und identifiziert die Sequenz (Batch), in dem der Lerner die größte Unsicherheit zeigt. Diese Unsicherheit wird gemessen als der negative Logarithmus der Wahrscheinlichkeit des Auftretens der Sequenz, gegeben das trainierte Modell. Diese Wahrscheinlichkeit berechnet sich aus dem Forward-Algorithmus des CRF, welcher beschrieben wird in (Rabiner, 1989). Das Orakel inspiziert die Sequenz und gibt die Start- und Endzeiten der Batchrezeptschritte an. Daraus ergeben sich sowohl die Label für den Klassifikator und das CRF. Die Signaldaten mit den Labels werden dem Trainingsdatensatz hinzugefügt und der Lerner wird erneut trainiert.

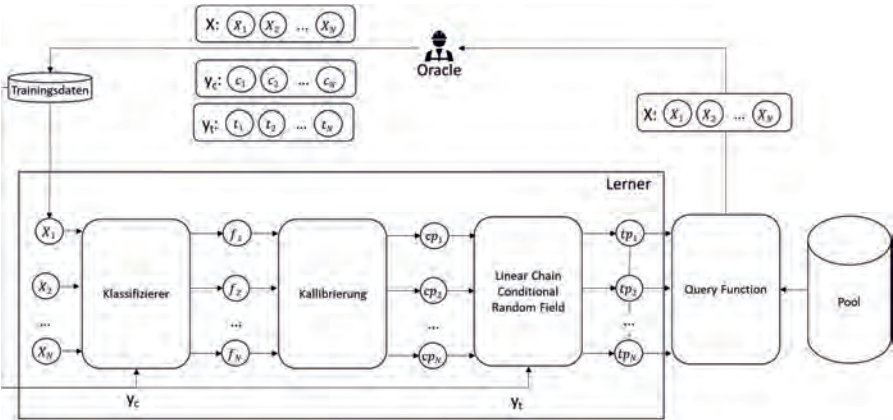


Bild 1: Kombination der Ansätze

2. Erste Experimentelle Ergebnisse

Der Prozess der in Bild 1 wurde prototypisch umgesetzt. Dabei kam als Zeitreihenklassifikator ein Random Forest aus Sklearn zum Einsatz. Zur Kalibrierung wurden ebenfalls die Sklearn Implementierung der Platt-Skalierung verwendet. Das CRF und die Query-Funktion wurden von Grund auf implementiert, wobei im CRF der Viterbi Algorithmus umgesetzt wurde.

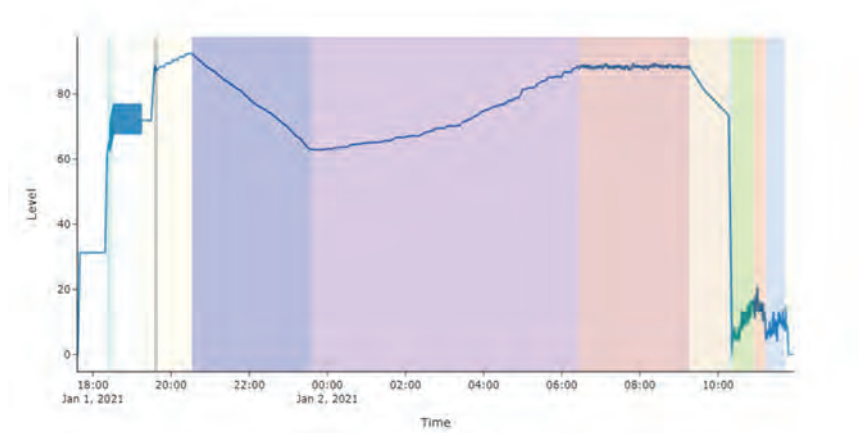


Bild 2: Füllstandmesswerte innerhalb eines Batches und Rezeptschritte (farblich hervorgehoben)

Für die Experimente wurden Daten aus einer numerischen Simulation (Dos Santos Louçada Coelho Vicente, 2021) verwendet. Für diese Daten stehen Ground-Truth Labels bereit, sodass das Orakel automatisiert umgesetzt wurde. Bild 2 zeigt den Füllstandsmesswert im Verlauf eines Batches und die farblich hervorgehobenen Rezeptschritte. Neben dem Füllstandsmesswert standen 3 weitere Werte (Dampfzufluss, Kühlwasserzufluss, und Motorumdrehungen der Pumpe für den Kühlwasserzufluss) zur Verfügung. Das verwendete Unsicherheitsmaß basiert auf dem erwähnten forward-Algorithmus.

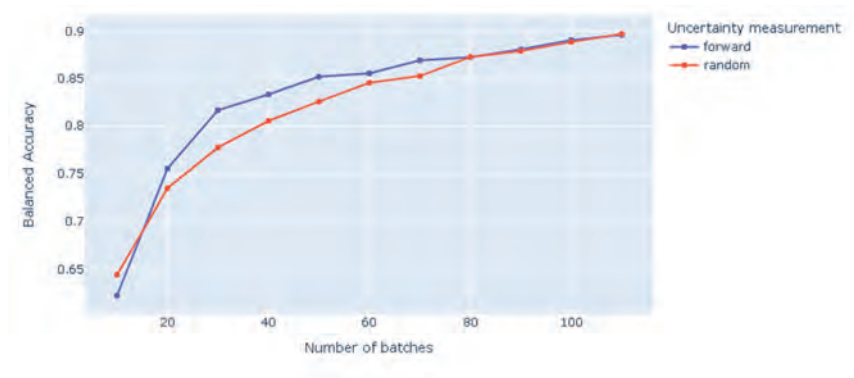


Bild 3: Vergleich von Uncertainty und Random Sampling

Bild 3 zeigt den Vergleich von Uncertainty Sampling und Random Sampling in der Queryfunktion. Es ist klar zu sehen, dass Uncertainty Sampling eine besser Labeleffizienz im Vergleich zum Random Sampling zeigt.

3. Zusammenfassung

Die Ermittlung von zuverlässigen Start- und Endzeiten von Schritten in Batchrezepten ist eine wichtige Voraussetzung zur Analyse der Batchdaten und damit zur Optimierung von Rezepten und Anlagenbetrieb. In Anlagen mit geringen Automatisierungsgrad sind die Start- und Endzeiten nicht verfügbar und müssen aufwändig in manueller Arbeit erzeugt werden.

Wir stellten eine Kombination von Zeitreihenklassifizierung, probabilistischer graphischer Modelle und Active Learning vor, die zu einer höheren Labeleffizienz, also einer besseren Performanz bei gleicher Anzahl von gelabelten Daten, führt. Erste Experimente zeigen, dass die Kombination die Labeleffizienz verbessert. Dies würde im praktischen Einsatz zu einer erheblichen Erleichterung für die Datenanalysten führen.

Acknowledgements

The German Federal Ministry for Economic Affairs and Energy (BMWi) is acknowledged for funding this research as part of the KEEN platform (reference code 01MK20014T).

Gefördert durch:



Bundesministerium
für Wirtschaft
und Energie

aufgrund eines Beschlusses
des Deutschen Bundestages

Literaturverzeichnis

- Böken, B. (2021). On the appropriateness of Platt scaling in classifier calibration. *Information Systems*.
- Dos Santos Louçada Coelho Vicente, M. (2021). Model for Batch Process Data Generation - A Benchmark for Machine Learning Testing and Comparison. Lissabon: Tecnico Lisboa.
- Fawaz, I., Forestier, G., Weber, J. I., & Muller, P.-A. (2019). Deep learning for time series classification: a review. *Data mining and knowledge discovery*, 917-963.
- Krähenbühl, P., & Koltun, V. (2011). Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. *Advances in Neural Information Processing Systems*. Curran Associates, Inc.
- Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *International Conference on Machine Learning* (pp. 282-289). Morgan Kaufmann.
- Merkelbach, S., Tan, R., Gärtler, M., Baehner, F., & Urbas, L. (2021). Automatic detection of start and end times of batch phases for the process industry. *Automation*. 2021.
- Murphy, K. (2001). *An introduction to graphical models*. Vancouver: The University of British Columbia.
- Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257-286.
- Reyes, O., Abdulrahman, H. A., & Ventura, S. (2018). Statistical comparisons of active learning strategies over multiple datasets. *Knowledge-Based Systems*, 274-288.
- Settles, B. (2009). *Active Learning Literature Survey*. Madison: University of Wisconsin.
- Sutton, C. (2012). An Introduction to Conditional Random Fields. *Foundations and Trends® in Machine Learning*, 267-373.
- Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 260-269.
- Wallach, H. (2004). *Conditional Random Fields: An Introduction*. Philadelphia, Pennsylvania: University of Pennsylvania.

Phasenidentifikation von Batch-Prozessen durch aktives Lernen

Phase Identification of Batch Processes by Using Active Learning

Dr. C. Song, Dr. Marco Gärtler, Dr. Benjamin Klöpper,
ABB Corporate Research Center, Ladenburg

Kurzfassung

Der Batch-Prozess ist eine Herstellungsmethode, bei der verschiedene Arbeitsgänge nacheinander durchgeführt werden. Im Gegensatz zum kontinuierlichen Verfahren werden die Reaktanten zu Beginn in den Reaktor eingespeist, und das Endprodukt kommt nicht kontinuierlich, sondern in Gruppen heraus. Batch-Prozesse werden häufig für hochwertige Produkte eingesetzt, z. B. bei der Herstellung von Arzneimitteln und Agrarchemikalien. Um die Produktionsleistung zu steigern, wird die industrielle Analytik (IA) mit datengesteuerten und KI-basierten Techniken angewandt, um wichtige Merkmale, wie z. B. Chargenphasen oder Start-/Endzeitpunkte, zu ermitteln. Automatisierungssysteme bieten zwar Möglichkeiten zur Aufzeichnung dieser Informationen, aber sie sind nicht allgemein verfügbar, nicht detailliert genug und enthalten Fehler bzw. Lücken.

Unser Ziel ist es, ein Modell zur Identifizierung der Phasen von Batch-Prozessen zu entwickeln, die verschiedenen Batch-Vorgängen entsprechen. Im Allgemeinen kann die Analyse von Batch-Prozessdaten mit verschiedenen Methoden durchgeführt werden [1–3], wobei die effizientesten Techniken immer noch auf maschinellem Lernen oder künstlicher Intelligenz basieren. Im Allgemeinen kann dies direkt als überwachtes Klassifikationsproblem modelliert werden, wobei die Namen der Phasen den Klassen entsprechen und entweder einzelne Zeitpunkte oder ganze Intervalle klassifiziert werden. Leider erfordern auch aktuelle Methoden eine erhebliche Menge an Trainingsdaten. Die Annotation der Daten kann nur von erfahrenen Ingenieuren oder Personen mit einem ausreichend guten Verständnis der Prozessdomäne vorgenommen werden; beides sind wahrscheinlich sehr knappe Ressourcen, was die Annotation zu einem teuren, aber langwierigen Prozess mit unzähligen zusätzlichen Möglichkeiten macht, Fehler und Verzerrungen einzuführen. Auf der anderen Seite ist Active Learning [4] ein Rahmenwerk, das die Integration von Benutzern und deren Wissen ermöglicht. In Anlehnung an Merkelbach et.al. [5] verwenden wir ein Setup mit Ground Truth und vergleichen aktives Lernen mit traditionellen überwachten Methoden. Um viele verschiedene

Szenarien und Modelleinstellungen zu testen, simulieren wir die Domänenexperten unter Verwendung der Grunddaten und messen die angeforderte Informationsmenge. Insgesamt haben wir festgestellt, dass eine geringe Anzahl von Bezeichnungen bzw. Anfragen ausreicht, um ein geeignetes Modell zu erhalten, das genauso gut oder sogar besser als überwachte Modelle funktioniert. Für den Aufbau wurden reale und simulierte Datensätze verwendet. Die erzielten Ergebnisse unterscheiden sich nicht wesentlich zwischen den verschiedenen Datensätzen, was auf die Robustheit und Anwendbarkeit des aktiven Lernens für unser Problem hinweist.

Abstract

The batch process is a method of manufacturing, where different operations are taken place consecutively. In contrast with the continuous process, reactants are fed into reactor at beginning, and final product comes out in groups and not continuously. Batch processes are often utilized for high value products, such as pharmaceutical and agrichemical manufacturing. To boost the production performance, the industrial analytics (IA) is applied using data-driven and AI-based techniques to obtain important characteristics, such as batch phases or start/end-time of those. While automation systems provide ways to record this information, it is not universally available, lacks in detail, and contains errors and gaps, respectively.

Our focus is to build a model to identify the phase from batch process, which correspond to different batch operations. In general, the way of analyzing batch process data can be carried out with different methods [1–3], the most efficient techniques are still machine learning or artificial intelligence based. In general, this can be directly modelled as supervised classification problem, where the names of the phases correspond to the classes and either individual points in time or whole intervals are classified. Unfortunately, even up-to-date methods require a significant amount of training data. The annotation of the data can only be done by experienced engineers or people with a sufficiently good understanding of the process domain; both are likely very scarce resources, which makes the annotating an expensive, yet tedious process, with countless additional ways to introduce errors and biases. On the other hand, Active Learning [4] is a framework, that allows the integration of users and their knowledge. Along the lines of Merkelbach et.al. [5], we use a setup with ground truth and compare active learning with traditional supervised methods. To test many different scenarios and model settings, we simulate the domain experts using the ground truth and measure the requested amount of information. Overall, we found that small amounts of labels and numbers of requests, respectively, are sufficient to obtain a suitable model that performs as good or even better than supervised ones. The setup uses real and simulated datasets. The

obtained results do not significantly vary between the different sets, indicating the robustness and applicability of active learning for our problem.

1. Active Learning

The main idea behind the active learning (AL) method is that the performance of machine learning models, which are constructed only on few data samples, can be improved if an interactive process established between the machine learning model and oracle. The main function of active learning algorithm is to decide which samples should be annotated during each iteration. In the other words, the active learning method can be seen as an outer loop of classical machine learning techniques, the training data are incremented by AL method. In a general active learning setting, the algorithm described above is referred as the active learner, the selected unlabeled samples for annotation are formulated as queries, these instances should be labeled by an oracle, which possess certain domain knowledge [4].

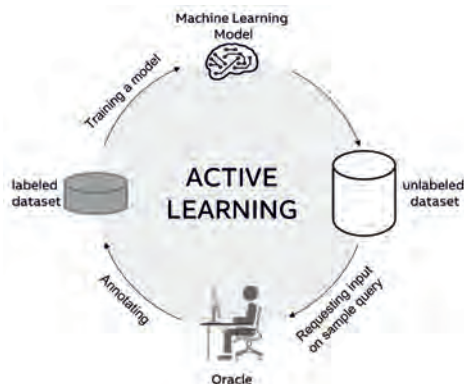


Fig. 1: Illustration of pool based active learning algorithm

In principle, there are three main settings for applying active learning: membership query synthesis; stream-based selective sampling; and item pool-based sampling. For our use case the pool-based sampling is most relevant, which focuses on choosing instances in the existing unlabeled dataset. For the other two settings, we refer the readers to [4]. For many real-world applications, the availability of a large amount unlabeled data can be assumed, the pool-based active learning tries to screen a pool of data at once and selects the most appropriate samples for further labeling. Fig. 1 illustrates the general framework of pool-based sampling, a low-accuracy machine learning model is trained with limited labeled data,

in combination with this pre-trained model, the active learner chooses one or few samples for the annotation from unlabeled sample pool. Once the training set is updated, the machine learning can be further trained. This iterative process will be altered once certain criterion are fulfilled. Obviously, the pool-based sample active learning fits better the scenario in batch phase identification, meaning that we possess certain amount raw data without labeling, and we want to train a classifier with considerable minimum effort of annotation.

The most common query selection frameworks are [4]:

- *Uncertainty sampling* extracts the indistinguishable sample data based on model's probability estimation and provide it to oracle for annotation.
- *Query by committee* is very similar to the ensemble training method, by using multiple models voting patterns, the "hard" sample data are selected.
- *Expected model change* chooses the sample, which can make the most change in current predictive model.
- *Expected error reduction* selects those sample data, which make the loss function decrease the most.
- *Variance reduction* chooses these samples, which minimize the output variance.
- *Density weighted method* selects the samples not only by judging their uncertainty, but also by considering the representative samples to avoid outliers.

In this study we consider only the uncertainty query sampling strategy because the computational effort for making a query is much lower than the other strategies and is potential more suitable for online active learning applications. Within the uncertainty sampling method, there exist three approaches for query generation: least confidence sampling [6], margin sampling [7] and entropy sampling [8]. We focus only on the first and last sampling approach in this study due to computational complexity.

The least confidence sampling tries to pick the instance, which has the least confidence in its most likely labels. The mathematical description reads:

$$x_{LC}^* = \arg \max_{x \in S} (1 - P_{\theta}(\hat{y}|x)) = \arg \min_{x \in S} P_{\theta}(\hat{y}|x),$$

Equation 1

where x^* is the query sample, S is the unlabeled data space, $\hat{y} = \arg \max_{y \in S} P_{\theta}(y|x)$, and θ is model parameters obtained based on labeled samples. $\hat{y}$ is the class with the highest model prediction probability for x .

In physics and information theory, the system uncertainty can be described by entropy, high-entropy means the higher system uncertainty. The definition of entropy sampling yields:

$$X_H^* = \arg \max_{x \in S} - \sum_i P_\theta(y_i|x) * \ln P_\theta(y_i|x),$$

Equation 2

where X_H^* is the query sample from entropy sampling, y_i is the overall possible labeling, and the definition of the rest parameters are identical in least confidence sampling.

2. Phase Identification in Batch Process

The batch production is common for manufacturing high value products, for example pharmaceuticals. The batch process can be considered as a period of time, during which a series of operations are carried out on a separate and identifiable item or parcel of materials, those operations are also known as batch phases. However, those specific operations are not always archived or accessible for process analysis, especially in the older and legacy distributed control systems. The machine learning method can be an efficient way to classify those batch phases from historical data, such that the further analysis can take place.

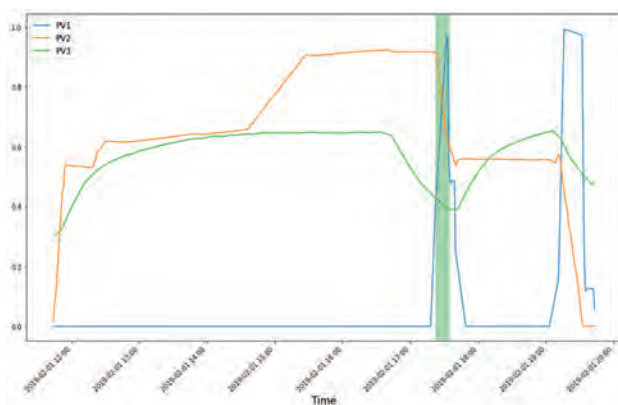


Fig. 2: Batch process cycle and batch phase

Using machine learning techniques to identify batch phases requires in general two key components: model selection and data preparation. In the past, the most research in chemical batch phase identification focuses on building efficient algorithms to classify batch events and assumes the necessary labeled data for training is accessible. However, for real-world applications the data preparation may occupy most of time in data science project life cycle. Moreover, labeling chemical batch phases can only be performed by people with domain

knowledge, e.g., chemical engineers or operation engineers. Demanding domain experts to conduct boring and monotonous labeling assignment can be demotivating and expensive. Active learning techniques can provide an interesting trade-off between an automatic labeling process, the provision of source labels and a domain expertise.

Under the framework of KEEN consortium [9] we obtain various batch process data from partners. In this paper we focus only on one batch phase from a batch process, Fig. 2 shows the considered batch process cycle with three process variables, and the highlight area is the target batch phase. This dataset consists of 190 batch process cycles, the target phase appears in each cycle.

3. Experiments and Evaluation

Our experiments are designed to assess the technical feasibility of an active learning setup. We use a fixed input dataset with a fixed phase. Both the results in [10] and our own exploration show, that they have a moderate degree of variations, which makes them a solid testbed without being trivial. The focus on the experiments is then to check that behavior and achieved quality is consistent.

In our implementation, we mainly use three numerical libraries: modAL [11], tslearn [12] and scikit-learn [12, 13]. modAL is an active learning framework, it provides the uncertainty sampling strategy and the overall active learning algorithm. tslearn is used for time series analysis, because the duration of batch event time series extracted from raw data is different, in order to compare these variable-length sequences, we use the dynamic time warping (DTW) [15] as the similarity measure in tslearn. scikit-learn is a comprehensive machine learning library containing countless models and algorithms, we mainly use the support vector machine (SVM) and k-nearest neighbors (KNN) algorithms in this experiment.

The algorithm we implemented for active learning based on the partner's data for batch phase detection can be summarized in Algorithm 1. Note that, the length of segments in X_1 is generally not equal. The segments in X_0 are selected in raw data where the batch event does not occur, they can be hence segmented with an equal length. However, in order to avoid the situation that the classifier only learns the length of time series, the length of each sample in X_0 is also chosen to be not constant, but the length of the majority of samples in X_0 and X_1 is comparable. In real world situation, the training set will be constructed with the help of oracle, in other words, experts will annotate segments for batch and not batch phases. In our current example, higher n_t means more additional annotations, hence n_t should not be very large. The ratio of X_0 and X_1 in X_{pool} can also be an issue in practice, if X_1 is less representative, active learner cannot pick good samples for improving classifier's accuracy.

Besides the data, the selection and setting of classifiers are also the key points, the classifier should be able to be trained on a limited data set for a robust performance, and it can also guide active learner to choose the most useful samples from pool data X_{pool} .

Algorithm 1 Active Learning algorithm

```

1: Segment time series  $X_1$  for a batch event in feature signals data.
2: Segment certain time series  $X_0$  in feature signals which are not in that batch event.
3: Select  $n_t$  samples from  $X_0$  and  $X_1$  respectively for building an initial training data  $X_{train}$ .  $\triangleright \#X_{train} = 2 \cdot n_t$ 
4: Select  $n_0$  samples from  $X_0$  and  $n_1$  samples from  $X_1$  to build a pool data  $X_{pool}$ .
5: Build a test data  $X_{test}$  using the rest of data in  $X_0$  and  $X_1$  with a 1 : 1 ratio.
6: Train a classifier on  $X_{train}$ .
7: procedure ACTIVE LEARNING LOOP
8:   while  $n < n_{query}$  do  $\triangleright$  Initial value of  $n$  is set to 0
9:     Active Learner selects one sample from  $X_{pool}$  for user to label.
10:    Update  $X_{train}$  with new labeled data and re-train the classifier.
11:    Test classifier with test data  $X_{test}$ .
12:   end while
13: end procedure

```

Algorithm 1 Active Learning algorithm

Therefore, we want to investigate the performance and robustness of current active learning setting regarding to the points mentioned above, the considered data is industrial data and obtained from a partner of KEEN consortium. We consider three process variables (PVs) in this dataset and focused on one batch phase for this experiment, but by using other batch phases provide similar results. Fig. 2 shows a typical batch process cycle with considered three PVs, and the highlight area is the target batch phase. In the following, the numerical experiments are divided into two parts: 1) investigation on the effect of class weight option in SVM and KNN classifiers with unbalanced pool data; 2) comparing active learning uncertainty sampling strategy and naive random sampling.

3.1. Weighted vs. not weighted

In the first experiments, we compare the prediction accuracy on the setting of class weight function in SVM and KNN classifier. As mentioned earlier, the data representing batch event is only a small portion in the unlabeled sample pool in practice. If the data with different labels are not weighted during the training process with respect to their quantity, data with less representative label may be misclassified.

We set $n_t = 2$ to build an initial training set with 4 samples. 80% of batch phases in data are preserved for testing purpose along with a same number of samples from X_0 , and the rest of X_1 samples are put into pool data X_{pool} . Moreover, in order to mimic the real-world scenario,

the amount of not batch phase samples is five times as the batch phase time series, they construct together the pool data set, i.e. $X_{pool} = X1_{pool} + X0_{pool}$, $\#(X0_{pool}) = 5 \cdot \#(X1_{pool})$.

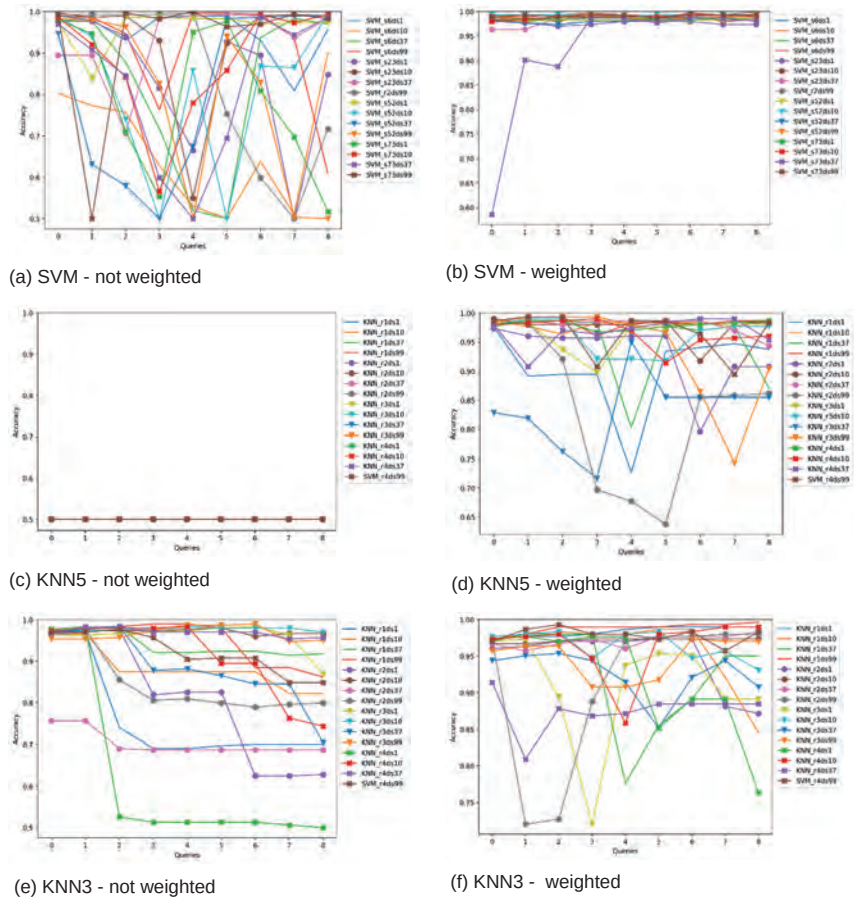


Fig. 3: Weighted vs. not weighted for SVM and KNN classifier

The least confidence classification (Equation 1) in uncertainty sampling query strategy is applied in active learning algorithm, a comparison with entropy sampling will be followed in later experiments. The selected data from $X1$ in initial training set can be biased for the classi-

fier as there are only 2 samples. Hence, we also add data shuffling during the sample selection process. Moreover, SVM classifier is generally sensible to its initial state, different random states in SVM classifier initialization are also carried out. For KNN classifiers, they are generally limited by the number of neighbors, the settings with 3 and 5 neighbors are examined. We distinguish them with KNN3 and KNN5 for the sake of convenience. Figure 3 shows the evolution of predictors' accuracy for the three classifiers mentioned above. The first row shows the results that the class weight option is not considered, and the second row is for the contrary situation. Regarding the labeling of plots in Figure 3, SVM_s6ds1 means that SVM classifier is set with random seed 6 in classifier initialization, and random seed 1 in data shuffling. KNN_r1ds1 means that the first run with random seed 1 in data shuffling. This naming convention applies also in the following experiments. Figure 3 demonstrates a significant improvement of applying weight function in model training for all three estimators. One can observe that when the class weight is not considered in SVM (Figure 3a), its performance is very unstable, the accuracy varies from 50% to 98%, and it seems not to be converged for certain runs. However, when the class weight is adjusted according to the class frequencies in pool data, SVM can converge very quickly, namely after 3 or 4 queries, with a very high accuracy around 98% (Figure 3b). Two KNN classifiers also show a similar behavior, i.e., the accuracy of model is improved after applying class weight option. KNN estimator with 5 neighbors exhibits even an extreme situation, where the classifier accuracy does not improve at all (Figure 3c). One can state that three classifiers can hardly converge when the class weight is not considered, the accuracy can drop down to 50% in the worse cases. A few of trained models in SVM and KNN3 can still preserve very high accuracy and demonstrate a monotone convergence behavior. However, they strongly depend on the initial data (namely data shuffling) and model's initialization. In the contrary, after the class weight is taken into account, SVM shows the most improvement, all models converge after 3 queries at a high accuracy. Yet, even KNN3 and KNN5 become more stable in this regard, not all models can converge eventually (e.g., Figures 3d and 3f), two KNN classifiers are not robust enough with respect to the initial data and the current active learning algorithm setting.

3.2. Naive sampling vs. entropy sampling

The section 3.1 focuses on the comparison between three machine learning classifiers, in this subsection, the effectiveness of applying query strategy in active learning algorithm is to be examined. As mentioned in Section 3.1, we are interested in uncertainty sampling because of low computational complexity and usability. However, the sampling step in Algorithm 1 can be even simplified by a naive random selection in pool data. Note that, as there

are more X_0 samples in pool data, a random selector will pick more X_0 samples expectedly, which may potentially have an impact on the learning process. We use the entropy sampling strategy (Equation 2) instead of least confidence for comparing with naive sampling.

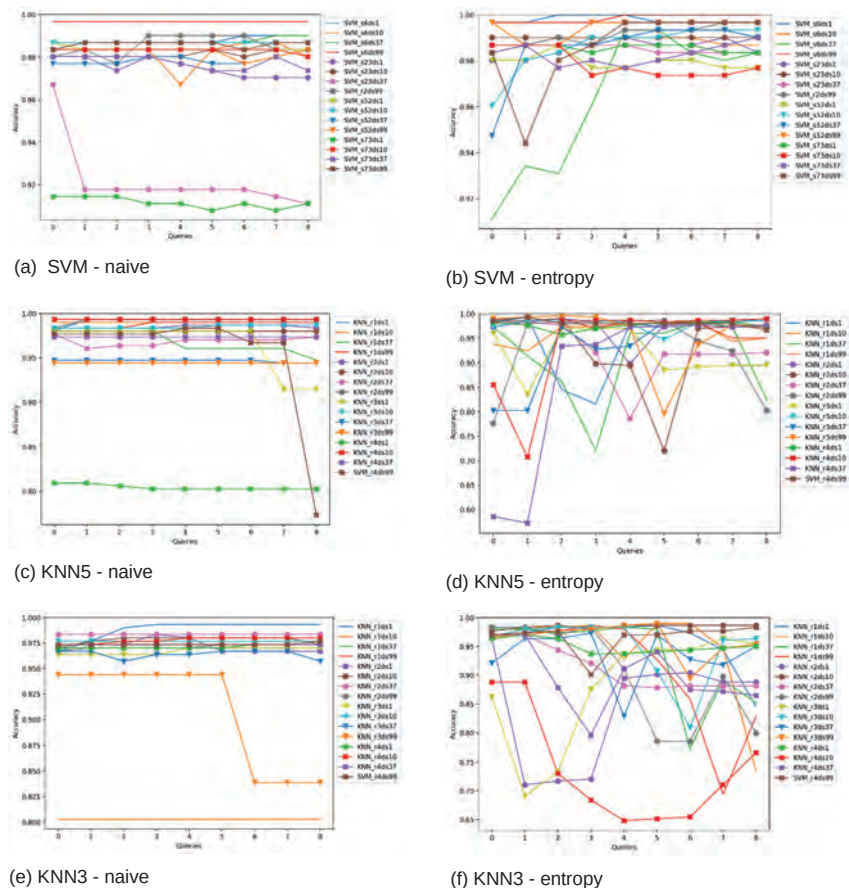


Fig. 4: Naive sampling vs. entropy sampling for SVM and KNN classifier

The construction of data sets follows the same procedure in Section 3.1, the class weight option is enabled for all three classifiers. The results on the first row in Fig. 4 are the experiments with naive random sampling, the second row is for the entropy sampling. As expected, the accuracy of three estimators is merely improved once the naive sampling strategy is applied, in certain cases they even become worse. Needless to say, in this setting the active

learner cannot learn much useful information from new labeled samples, because the most queried samples are in X_0 , the training set is not frequently updated with samples of the batch operation or phase. Hence, the precision of estimators is mostly determined by the initial training data. Nonetheless, once the uncertainty sampling is in place, SVM manifests its stability of accuracy improvement, all trained models achieve a high precision around 98%. KNN5 and KNN3 have similar behavior as in Fig. 3, not all estimators can show a monotone convergence. Obviously, the precision of most models trained based on naïve sampling can also reach a high precision, but certain models have also a very poor performance, which shows the lack of stability.

Applying sampling strategy in active learning algorithm is essential as it chooses the "right" samples for labeling, such that the precision of predictor can be further improved. However, choosing an appropriate model seems to be even more important, taking class weight into consideration can overcome on certain level the construction of pool samplings. In the previous two experiments, SVM classifier outperforms KNN in terms of overall accuracy and stability, and it can be trained with very limited data. There is no strong indication that least confidence sampling is more effective than entropy sampling, both provide meaningful query samples. Therefore, in the following tests, only the least confidence sampling is applied.

5. Conclusion

In this work, we studied a setup for active learning applied to batch phase labelling. In our experiments we could verify that with only four labels for initial training and a very small number of additional labels requested in the learning loop where sufficient to accurately label and identify the given phase. Both classifiers KNN and SVM can be efficiently trained under AL framework and the situation of unweighted samples needs to be taken into account during the training phase for better convergence. Overall, SVM combined with entropy query sampling strategy provide the most stable outcomes. These initial results are highly promising, and we plan to follow up in two streams. Firstly, using different phases and more datasets. As systematic study of the impact of different types and levels of noise could reveal when our setup will fail. Secondly, using human experts in the active learning loop to assess the usability and acceptance of such a system.

Reference

- [1] B. Schmidt, R. Tan, N. Li, M. Hollender, and M. Gärtler, "Efficient Process for Batch Analysis," *Chemie Ingenieur Technik*, Nov. 2021, doi: 10.1002/CITE.202100081.

- [2] C. A. Muñoz López, S. Bhonsale, K. Peeters, and J. F. M. van Impe, "Manifold Learning and Clustering for Automated Phase Identification and Alignment in Data Driven Modeling of Batch Processes," *Frontiers in Chemical Engineering*, vol. 0, p. 20, Nov. 2020, doi: 10.3389/FCENG.2020.582126.
- [3] P. Graniero and M. Gärtler, "Prediction of Batch Processes Runtime Applying Dynamic Time Warping and Survival Analysis," pp. 53–61, 2021, doi: 10.1007/978-3-662-62746-4_6.
- [4] B. Settles, "Active Learning Literature Survey," 2009, Accessed: Nov. 09, 2021. [Online]. Available: <https://minds.wisconsin.edu/handle/1793/60660>
- [5] S. Merkelbach *et al.*, "Automatic Detection of Start and End Times of Batch Phases for the Process Industry," Nov. 2021.
- [6] A. Culotta and A. Mccallum, "Reducing Labeling Effort for Structured Prediction Tasks", Accessed: Apr. 08, 2022. [Online]. Available: www.aaii.org
- [7] T. Scheffer, T. Scheffer, C. Decomain, and S. Wrobel, "Active Hidden Markov Models for Information Extraction," 2001, Accessed: Apr. 08, 2022. [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.2.795>
- [8] B. Settles and M. Craven, "An Analysis of Active Learning Strategies for Sequence Labeling Tasks," pp. 1070–1079, 2008, doi: 10.5555/1613715.1613855.
- [9] "KEEN – Künstliche-Intelligenz-Inkubator-Labore in der Prozessindustrie." <http://keen-plattform.de/> (accessed Apr. 04, 2022).
- [10] S. Merkelbach *et al.*, "Automatic Detection of Start and End Times of Batch Phases for the Process Industry," Apr. 2021.
- [11] T. Danka and P. Horvath, "modAL: A modular active learning framework for Python," May 2018, doi: 10.48550/arxiv.1805.00979.
- [12] "Tsllearn, A Machine Learning Toolkit for Time Series Data." <https://jmlr.org/papers/v21/20-091.html> (accessed Apr. 08, 2022).
- [13] L. Buitinck *et al.*, "API design for machine learning software: experiences from the scikit-learn project," Sep. 2013, doi: 10.48550/arxiv.1309.0238.
- [14] "Scikit-learn: Machine Learning in Python." <https://www.jmlr.org/papers/v12/pedregosa11a.html> (accessed Apr. 08, 2022).
- [15] H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978, doi: 10.1109/TASSP.1978.1163055.

Performance Evaluation of a Campus Network emphasizing the ICT-OT-relation

Lisa Underberg, Gustavo Cainelli,
Institut für Automation und Kommunikation e.V. Magdeburg;
Thomas Neugebauer, Götting KG, Lehrte/Röddensen

1. Abstract

This paper presents performance testing results from an industrial application's point of view of an early 5G standalone (Rel. 15) and a 4G campus network. These results indicate the necessity of a deeper mutual understanding of the particularities of industrial applications and the particularities of 4G and 5G networks and their devices. Moreover, problems encountered when conducting the tests emphasize the importance of not only focussing on performance parameters, but also on planning, engineering, commissioning and operating aspects.

2. Introduction

Currently, two yet disjunctive worlds are approaching each other: the information and communication (ICT) and the operational technology (OT) world. Against the background of campus networks, which can be licensed with currently low effort ranging between 3.7 and 3.8 GHz, these two worlds working jointly becomes more urgent than ever. With 5G, the ICT world claims to inherently address industrial applications' needs – not only for one specific type of application, but for various types of applications simultaneously. The OT world is well aware of these claims, but especially in the context of critical applications understandably cautious whether the communication networks are able to live up to these claims. Addressing these main points, valuable initiatives have started in both ICT and OT world separately, but most important also jointly, for example in the Industrial Radio Lab Germany (IRLG) [1] and in 5G-ACIA endorsed testbeds [2].

The present work discussed performance testing results determined by the Industrial Radio Lab Magdeburg jointly with Götting KG. Testing the communication system before and during its actual deployment is a fundamental step in determining whether the communication system can provide the features that the application requires. Thus, it is essential to perform performance tests from your application point of view, as this is the relevant perspective for producers [3].

This paper discusses the importance of performance testing of campus networks. Two campus networks, 5G and LTE are analysed with 2 different devices types. Moreover, the test system developed by ifak, the FTT, is put in contrast with other network performance tool called D-ITG.

The remainder of this paper is structured as follows. Section 3 gives an overview on related work in the context of performance testing of mobile campus network. Section 4 explains the system under tests and the test system. Moreover, the test cases and the environmental conditions are introduced. The measurement results are discussed in Section 5. Finally, Section 6 summarizes the results and gives an outlook.

3. Related work

In this section works that evaluate the performance of 5G systems are briefly presented. In [4], the authors evaluate the response time and throughput of a 5G device in 5G non-standalone (NSA) mode and 4G LTE mode. The influence of the transfer interval on the performance parameters are not evaluated. Furthermore, the user data length is also not considered during the analyses. The details about the influencing quantities used during performance testing are not given. In [5] the performance parameters data rate, transmission time, update time, jitter and packet loss of a 5G NSA network are evaluated. The authors evaluate several scenarios by varying the transfer interval and user data length. While we are considering a SA system, the authors are analysing NSA. In [6] the authors evaluate 5G SA and NSA networks in realistic industrial production environments with realistic deployment settings. Both deployments using commercial 5G systems and a 5G prototype system, including pre-commercial and standard-compliant URLLC functionality, were investigated. Latency and reliability measurements were performed for different message sizes. Results indicate that current 5G latency performance depends on packet size, transmission direction (uplink or downlink) and network configuration as well as end device design and capabilities.

4. Measurements setup

The measurement setup is composed of testing system and system under test (SUT) as shown in Fig. 1. The SUTs subjected to testing in this paper are described in Subsection 4.a. The testing system (called traffic generator in Fig. 1) sends messages in its source interface and receives it in its target interfaces. In this paper, two different testing systems are considered: D-ITG and FTT, which are described in Subsection b) and Subsection c), respectively. The SUT is everything that is between the source and target interfaces of the testing system. The

environment, where the performance testing was conducted, is described in Subsection d). The test cases used for the present analysis are specified in Subsection e).

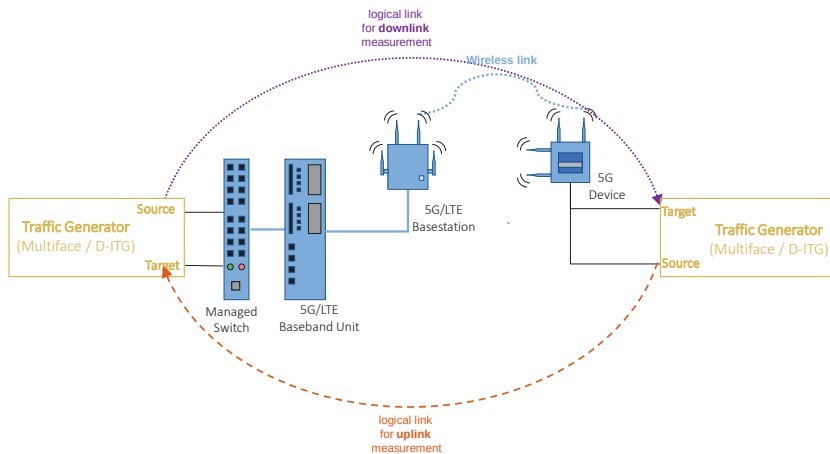


Fig. 1: Measurement setup with SUT and testing system.

a) Description of system under test

The system under test (SUT) is composed of campus network and device. The network comprises a base station, a baseband unit and a switch. Two different campus networks are evaluated: 5G SA and LTE. The devices are LTE/5G and two types are used. Here we will call them Type A and B:

Device Type A: 5G module RM500Q [7] on a development board with USB interface. This device when connected to a pc provides a mobile interface so that the PC can reach the devices connected to the 5G network.

Device Type B: 5G/LTE industrial router with Ethernet interface. This device can forward traffic received on the wireless interface to the local network (Ethernet). The devices connected to the Ethernet port can reach the devices on the network.

5G Network: It is a commercially available 5G Standalone system compliant with Release 15 of 3GPP [8]. The network uses 100 MHz bandwidth from 3,7 GHz to 3,8 GHz with a subcarrier spacing of 30 kHz, TDD ratio 1:4 (UL:DL), and TX power 40 dBm . The 5G network has been

set up and is operated by a third party. Therefore, more detailed information about the network is not available.

LTE Network: It is an LTE campus network consisting of a Mobile Edge Cloud Server (MECS) compliant with Release 15 of 3GPP and one base station compliant with Release 9 of 3GPP. It uses 20 MHz bandwidth from 3,78 GHz to 3,80 GHz, TDD ratio 1:1 (UL:DL) and TX power 24 dBm and an antenna gain of 6 dBm.

Combining these elements, three SUTs investigated are summarized in Table 1:

Table 1: System under test combining network and device.

	Device Type A	Device Type B
5G Network	SUT 1	SUT 2
LTE Network		SUT 3

SUT 1 was evaluated with the D-ITG as testing system, while the FTT was deployed with SUT 2 and 3.

b) Distributed Internet Traffic Generator (D-ITG)

The first testing system consists of two PCs. One is connected to a 5G device (type A) via USB. The other one is connected to the 5G network via Ethernet. To generate and measure data traffic, two instances of the Distributed Internet Traffic Generator (D-ITG) [9] were initiated. On each PC ran a client and a server D-ITG instance. D-ITG is configured and started through the Windows terminal.

The influencing quantities set in D-ITG were transfer intervals, message lengths and the number of messages to be transmitted. The performance parameters analysed were response time (RT) and update time (UT).

The tool has a feature to measure the transmission time. However, the accuracy depends on the synchronization of the source and the target PCs which is not guaranteed by the tool itself. In this case, the user's implementation of software or hardware is responsible for that. For this reason, in this work, the transmission time was not measured using D-ITG.

The RT refers to a service and is measured considering how long the request and response take. The UT can be measured at all target end-points as it does not depend on timing information related to the source end-point.

c) Funk Transfer Tester (FTT)

The FTT is a performance testing concept developed and implemented by ifak Magdeburg [9]. The testing system comprises a software called FTTManager and one or multiple traffic emulators called Multifaces. The system provides the influencing quantities and measures the

performance parameters. The Multifaces are implemented on an FPGA-basis which enables them to generate real-time traffic and measure performance parameters with an accuracy of less than 1 μ s. The interfaces provided by the Multifaces are typically used in industrial devices such as RS-232, Digital I/O and Ethernet.

The Multifaces are connected to a PC hosting the FTTManager, which is a customized software providing testing control functionality. The FTTManager configures and controls the Multifaces during testing and stores the test results. The Multifaces can send UDP packets, which are configured in the FTTManager, through its Ethernet interface.

The FTT testing system has several advantages compared to common UDP traffic generators like D-ITG [9] or iperf [11]. Although iperf and D-ITG are well-established tools in network performance evaluation, they are limited to certain applications. Both tools are not able to generate traffic emulating industrial applications, since industrial devices typically use interfaces like UART or digital input and outputs and may not establish TCP or UDP connections. Even if an industrial device can route UDP packets, this kind of traffic generator still depends on a PC to generate the traffic. This implies two main issues: the first one is that, if the transfer interval is too low (< 100 ms), they cannot generate traffic with a precise transfer interval. The second one is that iperf cannot measure the transmission time and although D-ITG has a feature to measure it, the accuracy depends on the synchronization of the source and the target PCs which is not guaranteed by the tool itself. Due to the lack of synchronization of source and target instances, the measurement of the transmission time by the mentioned tools is prone to be inaccurate compared to industrial applications' requirements.

According to [12] in some motion control applications, the transfer interval can be around 500 μ s and consequently a testing system with accuracy < 1 μ s is demanded. In this way, the area of evaluation of wireless solutions for the industry remains open and needs solutions that meet the different industrial applications. The measurement system comprising FTTManager and Multifaces deployed in the present work emulates the application system, i.e., the source and the target of the messages (distributed application) are external to the communication system, being closer to a real application. In addition, the source and target Multifaces are synchronized ensuring accuracy of less than 1 μ s in the measurement of transmission time. Due to their FPGA implementation, the Multifaces also generate data packets and measure the update and transmission time with a deterministic precision of less than 1 μ s.

d) Environmental conditions

The tests were carried out in a test hall recreating a factory environment with machines such as robots and cranes. Fig. 2 shows the test environment where the performance tests were conducted. Fig. 3 shows the floor plan of the test hall. The distance between the base station

(BS1) and the device is 40.66 m. This distance considers the height of the device and the base station.



Fig. 2: Test environment.

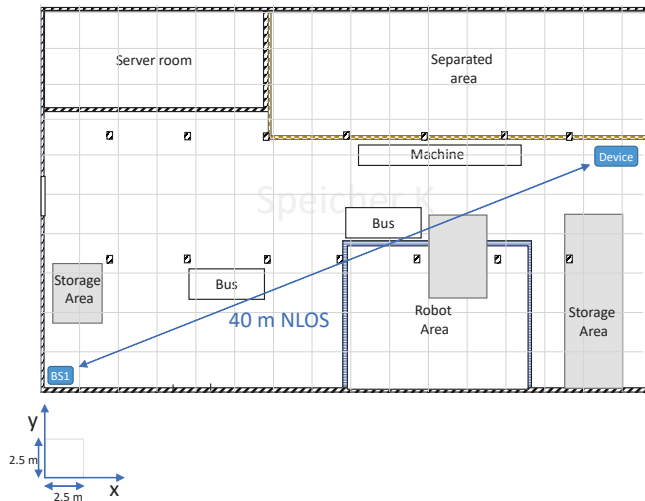


Fig. 3: Reference environment with positions of base station and device.

e) Test case specification

The performance testing is conducted for logical links (LL). A logical link represents a communication relationship between two distributed instances of an application. A single logical link may comprise multiple physical links, i.e., a wireless 5G link and a wired Ethernet link. The logical link has two endpoints: the source and the target. Considering the performance testing setup as shown in Fig. 1, the LLs are links between two traffic generators. In this paper, two logical links – one in the uplink (dashed orange line) and one in the downlink (dotted purple line) direction – are subjected to performance testing. The source interface of the traffic generator sends data according to the configured transfer interval of the logical link. It represents the logical source endpoint. The data are received by the target interface, which represents the LL's target endpoint. Between the logical source endpoint and target endpoint is the SUT [3].

Below a brief definition of influencing quantities and performance parameters evaluated is given. More detailed information can be found in [13] and [14].

Influencing quantities:

- Transfer interval: the transfer interval (TI) is the time between two transmissions on a logical link;
- User data length: it is the number of useful data in bytes of a message;
- Number of messages: the number of transmitted messages in each test case;

Performance Parameters

- Transmission time (TT): it is the time a message takes to travel from the source interface to the target interface;
- Update time (UT): it is the time between two received messages at the target interface;
- Response time (RT): It is the time interval between the request message leaves the source interface and the moment the response message is delivered to the same source reference interface.

The performance testing is structured in a test suite. The test suite contains test groups (TG), each comprising several test cases (TC). Each test case determines the influencing quantities to be implemented by the testing system and the SUT. Each test group aims to investigate a specific aspect or a specific influence on the SUT. In a test group, only one single influencing quantity is varied. By this, the varied parameter's impact on the SUT is investigated. For each SUT presented before, an extensive analysis with several test groups was performed. In this paper, we discuss only three test cases, each one performed on one SUT. TC01 is performed on SUT1, TC02 on SUT2 and TC03 on SUT3. Table 1 shows the general test conditions.

Table 2: Specification of test cases.

TC	SUT	Testing system	Logical Link	Distance [m]	TI [ms]	User data length [bytes]	Num Msgs	Performance Parameters
01	1	D-ITG	DL, UL	40	100	20	50000	RT, UT, MLR
02	2	FTT	DL, UL	40	131	20	50000	TT, UT, MLR
03	3	FTT	DL, UL	40	131	20	50000	TT, UT, MLR

TC: test case

TI: Transfer Interval

SUT: System under test

RT: Response time

UT: Update Time

MLR: message loss ratio

5. Measurements results

This section discusses the results of test cases. First the update time of the three test cases are compared in Subsection a). Then, the transmission time of TC02 and TC03 are discussed in Subsection b). As the D-ITG does not measure TT, the response time is discussed for TC01.

a) Update time

The histograms of the update time in the downlink for the three test cases are shown in Fig. 4, Fig. 5 and Fig. 6. The central value for TC01 is 100 ms, however, it has two maxima. This means, that although the mean value of the UT follows the TI (100 ms), this value only occurs with a minimum of message transmissions. This behaviour is known from wireless systems with Time Division Multiple Access scheme. Moreover, the TI of 100 ms is a multiple of the network's communication cycle. Apart from this, the testing system of TC01 (D-ITG) relies on a computer to work and, in this case, the times of processing and routing on the PCs have to be taken into account as well.

Fig. 5 and Fig. 6 show the results of TC02 and TC03. For these TCs the FTT was used as the testing system, which eliminates the uncertainty caused by the PCs' software and operating system. Moreover, the TI was set to 131 ms, which is a prime number and thus cannot be a multiple of the communication network's communication cycle. The testing results reveal only one maximum at 131 ms. The results of TC03 reveal a remarkable difference if compared to TC02. The distribution is larger in TC03 (LTE) than in TC02 (5G), but the maximum is also at 131 ms.

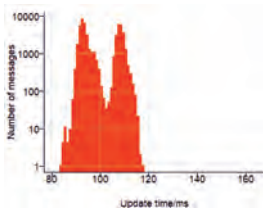


Fig. 4: Update time histogram of TC01 (D-ITG, 5G, TI = 100 ms) downlink.

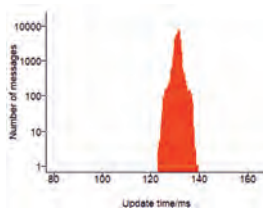


Fig. 5: Update time histogram of TC02 (FTT, 5G, TI = 131 ms) downlink.

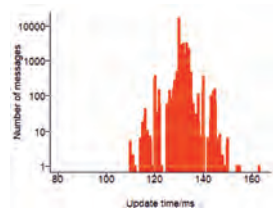


Fig. 6: Update time histogram of TC03 (FTT, 4G, TI = 131 ms) downlink.

The sequence of UT values for individual message transmissions, as depicted in Figure 5, shows that the UT value fluctuates around the value of the transfer interval (100 ms) in TC01, but hardly any UT matches this value exactly.

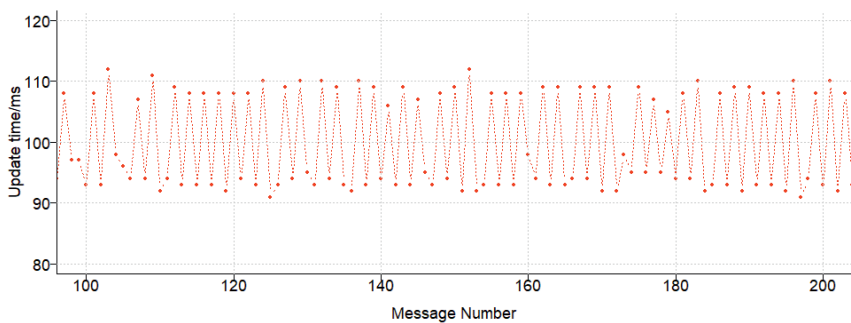


Fig. 7: Excerpt of update time sequence chart of TC01 (D-ITG, 5G, TI = 100 ms) downlink.

Figure 8 and Fig. 9 show the UT sequence measured in TC02 and TC03. As desired, the individual UT values are concentrated near the value of the TI (131 ms). However, Fig. 9 shows that the UT of some messages (beginning of the chart) oscillates in a wider range than TC02 (Figure 8).



Figure 8: Excerpt of update time sequence chart of TC02 (FTT 5G, DL, TI = 131 ms) downlink.

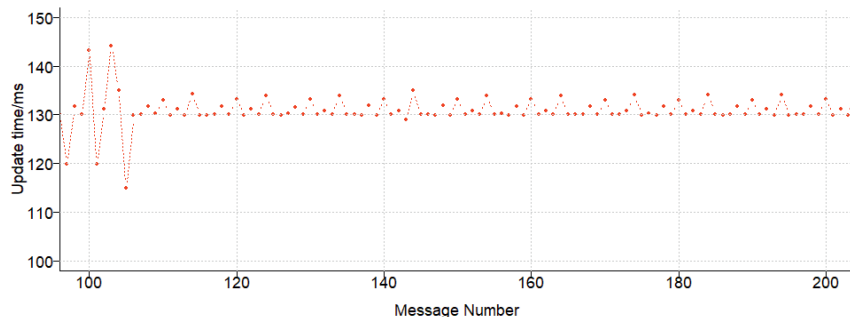


Figure 9: Excerpt of update time sequence chart of TC03 (FTT LTE, DL, TI = 131 ms) downlink.

b) Response time in TC01 and transmission time in TC02 and TC03

As D-ITG cannot measure the transmission time, the response time is presented instead. The response time includes a request and a response. In other words, the RT includes the uplink and downlink directions and is the sum of transmission times from both directions plus the time to process the message and send a response. Here we consider RT downlink when the source of the request is on the network side and RT uplink when the source of the request is on the device side. As shown in Fig. 10, the RT on downlink has two maxima, which is similar to the measured UT. Moreover, 95% of all responses are received before 31.2 ms.

Fig. 11 shows that the P95 value for transmission time in TC02 is 6.2 ms. The P95 value for TC03 is 12.9 ms as seen in Fig. 12.

The uplink results are seen in Fig. 13, Fig. 14 and Fig. 15. Although in TC01 the minimum and maximum values are similar for downlink and uplink, the distribution is quite different. While in the downlink there are two maxima, in the uplink the distribution is almost uniform.

The comparison between upper charts with LTE results (Fig. 11 and Fig. 12) with lower charts with 5G results (Fig. 14 and Fig. 15) reveals the downlink has a lower minimum value than the uplink. Moreover, the P95 is also lower with 5G. The 5G system has a better performance in downlink than uplink and it can be related to the UL:DL ratio that is 4:1. Although the UL:DL ratio in the LTE system is 1:1, the downlink exhibits a better performance, which indicates there are additional factors impacting the (industrial) end-user experience beyond the networks' properties.

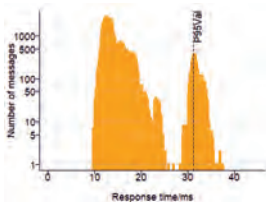


Fig. 10: Response time of downlink in TC01 (D-ITG, 5G, TI = 100 ms).

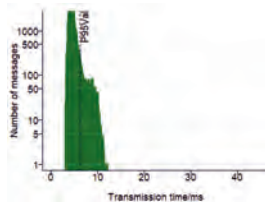


Fig. 11: Transmission time of downlink in TC02 (FTT, 5G, TI = 131 ms).

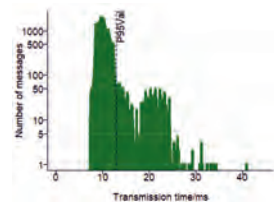


Fig. 12: Transmission time of downlink in TC03 (FTT, LTE, TI = 131 ms).

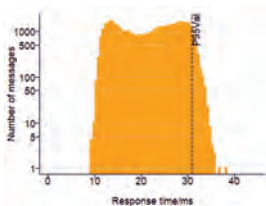


Fig. 13: Response time of uplink in TC01 (D-ITG, 5G, TI = 100 ms).

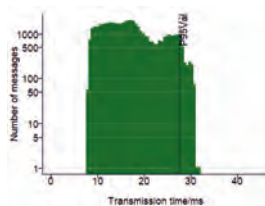


Fig. 14: Transmission time of uplink in TC02 (FTT, 5G, TI = 131 ms).

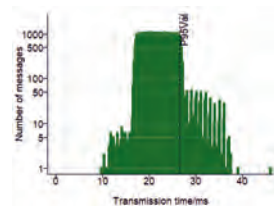


Fig. 15: Transmission time of uplink in TC03 (FTT, LTE, TI = 131 ms).

6. Conclusion and future works

This paper presented a performance evaluation of three different systems under tests with two test systems. The three system under tests analysed in this paper are mobile networks composed of a combination between two networks: 5G and LTE campus network, and two devices: an evaluation kit with 5G/LTE module and an industrial 5G router. The testing systems were the D-ITG and the FTT. While D-ITG runs on every computer, only the FTT with its specialized hardware is able to measure timing behaviour accurately. The results showed that

the LTE system achieve results similar to 5G system. Before reaching the promised potential, 5G Release 15 still has some aspects to be improved.

Although the measurements were conducted using commercially available devices and networks as SUT, we encountered several problems from a device not connecting to the network due to network software issues to a device behaving unexpectedly due to a parameter set manually deep down in the device's settings. We were able to solve these problems, but some of them only with assistance from the network or device manufacturer. Consequently, our experience reveals open issues to be addressed beyond the mutual understanding of performance parameters. In order to lift the potential of wireless communication systems in industrial applications, a comprehensive perspective has to be demanded and established by both ICT and OT world. This comprehensive perspective includes all aspects relevant to the interoperation of communication and industrial application systems - from the planning phase to operation.

7. Acknowledgement

The tests were conducted at the Digital Application Center - Mobility, Logistics, Industry - in Magdeburg, Saxony-Anhalt, Germany. This work was funded in parts by the project "Industrial Radio Lab Germany" under contract 16KIS1013, funded by the Federal Ministry of Education and Research, Germany.

8. References

- [1] "IRLG," Industrial Radio Lab Germany, [Online]. Available: <https://industrial-radio-lab.eu/>. [Accessed 20 April 2022].
- [2] 5G-ACIA, "5G-ACIA-endorsed testbeds," [Online]. Available: <https://5g-acia.org/insight/endorsed-testbeds/>. [Accessed 20 April 2022].
- [3] L. Rauchhaupt and L. Underberg, "Performance Testing of Novel Wireless Communication Networks for Industrial Automation on the Example of 5G," in *Komma 2019 - 10. Jahreskolloquium "Kommunikation in der Automation"*, Magdeburg, 2019.
- [4] A. Aijaz, B. Holden and F. Meng, "Open and Programmable 5G Network-in-a-Box: Technology Demonstration and Evaluation Results," 2021. [Online]. Available: <https://arxiv.org/abs/2104.11074>.
- [5] G. Soós, D. Ficzer, P. Varga and Z. Szalay, "Practical 5G KPI Measurement Results on a Non-Standalone Architecture," in *NOMS 2020 - 2020 IEEE/IFIP Network Operations and Management Symposium*, 2020.
- [6] J. Ansari, C. Andersson, P. de Bruin, J. Farkas, L. Grosjean, J. Sachs, J. Torsner, B. Varga, D. Harutyunyan, N. König and R. H. Schmitt, "Performance of 5G Trials for Industrial Automation," *{Electronics}*,, vol. 3, {2022},.
- [7] Quectel, "RM500Q-GL 5G Module Specification," Quectel, [Online]. Available: https://www.quectel.com/wp-content/uploads/2021/03/Quectel_RM500Q-GL_5G_Specification_V1.3.pdf. [Accessed 29 March 2022].
- [8] ETSI, "3GPP TR 21.915 version 15.0.0 Release 15," 2019.
- [9] S. Avallone, S. Guadagno, D. Emma, A. Pescapé and G. Ventre, "D-ITG distributed Internet traffic generator," in *First International Conference on the Quantitative Evaluation of Systems*, 2004.
- [10] A. Gnad, L. Rauchhaupt and S. Willmann, ""Verteiltes Testsystem zur Bewertung der Zuverlässigkeit von Kommunikationssystemen für Industrie 4.0 Anwendungen.," in *8. Jahreskolloquium "Kommunikation in der Automation - Komma "*, Magdeburg, 2017.
- [11] "iperf," [Online]. Available: <https://iperf.fr/>. [Accessed 25 April 2022].
- [12] 5G-ACIA, "5G for Automation in Industry - Primary use cases, functions and service requirements," 2019.
- [13] 5G-ACIA, "Performance Testing of 5G Systems for Industrial Automation.," February 2021.

- [14] V. 2185, "Radio-based communication in industrial automation.", part 4 "Metrological performance rating of wireless solutions for industrial automation applications.," March 1, 2019.

Konfigurierbare Controller-Controller-Kommunikation für modulare Produktionssysteme

Configurable Controller-Controller-Communication in Modular Production Systems

A. Stutz, T. Kempin, Dr.-Ing. M. Maurmaier,
Siemens AG, Karlsruhe;

M. Blumenstein, Prof. Dr.-Ing. A. Fay,
Helmut-Schmidt-Universität, Hamburg

Kurzfassung

Dieser Beitrag untersucht vier neue Varianten zur Umsetzung von Querkommunikation in modularen Produktionsanlagen auf Basis der Request/Reply und Publish/Subscribe Pattern. Bei einer qualitativen Analyse zeigt sich, dass drei Varianten mit direkter Querkommunikation deutliche Vorteile bezüglich der Übertragungszeiten aufweisen. Diese Varianten werden als Alternativen zum bereits standardisierten Verfahren weiter betrachtet und mittels OPC UA umgesetzt. Dabei wird auch eine Möglichkeit zur Vorverarbeitung von Werten aus dem Bereich der Service-Choreografien aufgezeigt. Die daraus entstehenden konfigurierbaren Mechanismen zur Querkommunikation wurden erfolgreich einer ausführlichen Evaluation unterzogen. Der Beitrag liefert somit eine Grundlage für zukünftige Arbeiten zur Spezifikation direkter Querkommunikation als Basis-Spezifikation des Module Type Package Konzepts.

Abstract

This paper examines four new variants for implementing cross communication in modular production systems based on the request/reply and publish/subscribe patterns. A qualitative analysis shows that three variants with direct cross communication have significant advantages in terms of transmission times. These variants are considered further as alternatives to the already standardized approach and have been implemented using OPC UA. In this context, a possibility for preprocessing values from the area of service choreographies is also demonstrated. The resulting configurable mechanisms for cross-communication have successfully been evaluated in detail. The paper thus provides a basis for future work on the specification of direct cross communication as a basic specification of the Module Type Package concept.

1. Motivation und Zielstellung

Eine direkte Querkommunikation zwischen Modulen (bzw. deren Controllern) in modularen Produktionssystemen ist seit längerem gefordert [1]. Als Grundlage dafür wurden in [2] bereits erste Untersuchungen zu konfigurierbaren OPC UA Client/Server-Strukturen veröffentlicht. Neuerungen der Module Type Package (MTP) Richtlinie VDI/VDE/NAMUR 2658 [3] sowie neue Erkenntnisse im Bereich der Service-Choreografien [4] machen eine Überarbeitung bzw. Erweiterung der Konzepte aus [2] erforderlich. Darüber hinaus stellt OPC UA Publish/Subscribe, wie sich zeigen wird, eine zweite vielversprechende Kommunikationstechnologie dar, die in die Untersuchungen mit aufgenommen werden soll.

Ziel des vorliegenden Beitrags ist es daher, diese Reihe von Veröffentlichungen fortzuführen und mit einer Gegenüberstellung verschiedener Kommunikationsmechanismen eine Basis für künftige Arbeiten zur direkten Querkommunikation zu schaffen. Der Beitrag ist hierfür wie folgt strukturiert. In Abschnitt 2 erfolgt eine Einführung in das MTP Konzept und im Speziellen in den Modellaspekt der Process Values. Abschnitt 3 führt allgemeine Anforderungen an die Querkommunikation ein, gefolgt vom Stand der Technik in Abschnitt 4. In Abschnitt 5 werden die Ergebnisse einer qualitativen Untersuchung unterschiedlicher Kommunikationsmuster vorgestellt. In den Abschnitten 6 und 7 werden Umsetzungsmöglichkeiten der Kommunikationsmuster auf Basis von OPC UA Client/Server und Publish/Subscribe aufgezeigt. Abschnitt 8 führt die Erweiterung der kommunikativen Mechanismen um eine adaptive Logik zur Vorverarbeitung der ausgetauschten Werten ein. In Abschnitt 9 werden die Ergebnisse einer Evaluation in Form einer Architekturanalyse, einem Funktionstest mit Latenzmessungen sowie einem kontrollierten Experiment dargestellt. Der Beitrag schließt mit einer Zusammenfassung und einem Ausblick im Abschnitt 10 ab.

2. Einführung in das Module Type Package und den Modellaspekt der Process Values

Das MTP Konzept ist eine Spezifikation von standardisierten Laufzeitschnittstellen und einem semantischen Datenaustauschformat, dem sogenannten Module Type Package (MTP), welches in der VDI/VDE/NAMUR Richtlinien-Reihe 2658 [3] spezifiziert wird. Das MTP basiert auf einem generischen Ankermodell, welches mittels weiterer Modellaspekte unterschiedliche Automatisierungstechnische Inhalte beschreiben kann. Das MTP-Dienstekonzept [5] erweitert die Summe an Modellaspekten um drei weitere Modellteile – darunter auch das sogenannte ProcessValueSet, welches die Modellierung von Prozesswerten ermöglicht. Die Prozesswerte sind dabei als eigenständige Objekte im MTP-Modell beschrieben. Sie sind jeweils einem Modul zugeordnet und können mit darin enthaltenen Diensten relationiert werden, wenn diese den Prozesswert konsumieren oder beeinflussen. Dieser strukturellen Umsetzung von

eigenständigen Prozesswert-Schnittstellen wird im weiteren Verlauf dieses Beitrags und bei der Umsetzung der Konzepte zur Querkommunikation Rechnung getragen.

3. Anforderungen an die Querkommunikation

Die Forderung nach Querkommunikation in verteilten Automatisierungssystemen ist nichts grundsätzlich Neues. Im Forschungsprojekt ORCA [7] wurden bereits Anforderungen an die Orchestrierung von Automatisierungsdiensten gestellt. Unter Orchestrierung wird dabei das Engineering, die Steuerung und Überwachung modularer Produktionsanlagen verstanden, was auch den modul-übergreifenden Austausch von Informationen zwischen Diensten beinhaltet. In der NE 171 [1] wurden zudem die Anforderungen an die Querkommunikation im Kontext der produktionsnahen Logistik vorgestellt. In [8] beschreiben die Autoren die Forderung einer Querkommunikation über Grenzen von Standards und Ansätzen hinweg, explizit wurden hier der MTP Standard, OMAC PackML und BaSys Components betrachtet. Ausgehend von diesen Beiträgen und Erfahrungen aus bisherigen Industrieprojekten lassen sich drei Technologie-unabhängige Anforderungen an die Querkommunikation ableiten.

Anforderung 1 – Forderung zur Umsetzung eines Informationsflusses

Diese Anforderung beschreibt die grundlegende Möglichkeit zur Herstellung eines Informationsflusses von einer Quelle zu einer Senke. Aus Gründen der Standardisierbarkeit ist auf hersteller-neutrale Kommunikationstechnologien zurückzugreifen. Die konkrete Ausgestaltung des Informationsflusses obliegt der entsprechenden Lösung. Darüber hinaus ist für das Einrichten und Herstellen eines solchen Informationsflusses eine entsprechende Konfigurationsschnittstelle vorzusehen, die ein Einrichten oder Ändern eines solchen Informationsflusses ermöglicht, ohne dass ein erneutes Übersetzen und Laden der SPS-Software erforderlich ist.

Anforderung 2 – Forderung zur Vorverarbeitung auszutauschender Informationen

Um Unterschiede der Skalierung und Normierung von Quellen und Senken von Informationsflüssen auszugleichen, sind ausgehende und eingehende Informationen durch eine konfigurierbare Vorverarbeitung anzupassen. Die Vorverarbeitung soll hierzu auf gängigen logischen und arithmetischen Funktionen basieren. Auch hierbei ist eine Konfigurationsschnittstelle vorzusehen, damit kein erneutes Übersetzen/Laden der SPS-Software notwendig ist.

Anforderung 3 – Möglichkeit zur Überwachung des Informationsflusses

Um die Vielzahl von Kommunikationsverbindungen in verteilten Automatisierungssystemen überwachen zu können, sind sowohl auf der aktiven als auch passiven Seite eines Informationsflusses entsprechende Maßnahmen zu treffen. Hierzu ist auf der Seite der Senke ein Zeitstempel über die letzte Aktualisierung des Prozesswertes zu erfassen. Ein Prozesswert besteht aus dem eigentlichen Wert und dessen Quality Code. Ist die letzte Änderung älter als

eine eingestellte Zeit, so ist eben dieser Quality Code auf „bad“ zu setzen, womit dem nachfolgenden Programm signalisiert, dass der Wert schlecht ist.

4. Stand der Technik

Für die Realisierung von Querkommunikation existieren bereits einige Möglichkeiten in heutigen Steuerungssystemen. Aus diesem Grund beschränkt sich diese Betrachtung des Stands der Technik auf Hersteller-neutrale Ansätze. Wie eingangs erwähnt, baut diese Arbeit auf den beiden Vorarbeiten [2, 4] auf. Eine Umsetzung des im MTP Konzept standardisierten Verfahrens ist in [9] veröffentlicht. Auch OMAC PackML hat in [10] standardisierte Ansätze für die Querkommunikation mit Fokus auf Verpackungsmaschinen dargestellt, welche in [11] als Companion Specification für OPC UA veröffentlicht wurden. Des Weiteren veröffentlicht die OPC Foundation mit [12] eine Basis-Spezifikation, mit der OPC UA Server direkt miteinander Informationen über Publish/Subscribe-Mechanismen austauschen können. In [13] hat die Fa. Siemens eine quelloffene Bausteinbibliothek für eine minimale Umsetzung von OPC UA Publish/Subscribe veröffentlicht, die auch in dieser Arbeit Verwendung findet. Darauf aufbauend wurde 2018 die Field Level Communication Initiative gegründet, die eine weitere Spezifikation für die Querkommunikation in modularen Anlagen spezifiziert [14]. In [15] haben die Autoren erste Erfahrungen und Eindrücke eines noch unveröffentlichten Release Candidate zu OPC UA FX veröffentlicht. Diese Technologie fokussiert im Speziellen auf die Echtzeit-Anforderungen in der Feldebene und greift hierbei sowohl auf die Client/Server- als auch Publish/Subscribe Mechanismen zurück [15].

Zusammenfassend zeigt sich, dass es einige existierende und zukünftige Mechanismen gibt, auf denen sich eine Querkommunikation für die modulare Produktion umsetzen lässt. Jedoch ist gerade hinsichtlich der Vorverarbeitung-Funktionalität bei allen genannten Referenzen nichts vorhanden oder geplant, u.a. dieser Lücke widmet sich dieser Beitrag.

5. Kommunikationsmuster für die Querkommunikation in modularen Anlagen

Mit Request/Reply und Publish/Subscribe sollen in diesem Beitrag zwei in der IT bekannte und etablierte Kommunikationsmuster mit Blick auf eine Querkommunikation in modularen Anlagen untersucht werden. In Bild 1 werden dazu fünf mögliche Kommunikationsmuster dargestellt. Die Varianten 1 und 2 basieren auf dem Publish/Subscribe-Pattern. Die Varianten 3 bis 5 basieren auf dem Request-Reply-Pattern. Variante 3 und 4 stellen auf Grund der aktiven Rolle in der Querkommunikation höhere Anforderungen an die Steuerungssysteme. Variante 5 entspricht dem derzeit standardisierten Verfahren und ermöglicht eine passive Auslegung der modularen Prozesseinheiten, da der aktive Kommunikationspart vom übergeordneten Process Orchestration Layer (POL) übernommen wird.

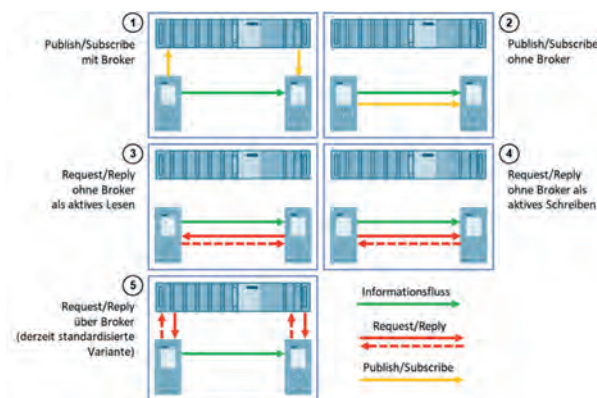


Bild 1: Varianten-Übersicht der untersuchten Muster zur Querkommunikation auf Basis von Request/Reply- und Publish/Subscribe-Mechanismen

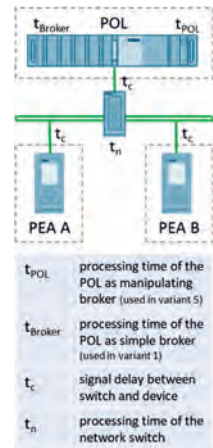


Bild 2: Schematische Darstellung der betrachteten Netzwerktopologie zur Ermittlung der Übertragungszeiten

Diese Varianten sollen im Folgenden qualitativ hinsichtlich ihrer Übertragungszeiten untersucht werden. Zur Ermittlung dieser qualitativen Größen wird die in Bild 2 dargestellte Netzwerktopologie genutzt. Hierbei werden zur Vereinfachung die Verarbeitungszeiten der Steuerungen innerhalb der beiden modularen Prozesseinheiten (PEAs) nicht berücksichtigt. Stattdessen liegt der Fokus allein auf die Querkommunikation ausgehend von den Schnittstellen der jeweiligen Steuerungseinheiten der modularen Prozesseinheit. Darüber hinaus wird ebenfalls zur Vereinfachung die Übertragungszeit zwischen den Geräten und dem Switch gleichgesetzt. Die so ermittelten Übertragungszeiten sind in

Tabelle 1 dargestellt.

Tabelle 1: Auflistung der qualitativen Größen für die fünf untersuchten Kommunikationsmuster für die Querkommunikation

Variante	Qualitative Größen
Variante 1 – Publish/Subscribe mit Broker	$T_{v1} = t_{\text{brok}} + 4 * t_c$
Variante 2 – Publish/Subscribe ohne Broker	$T_{v2} = 2 * t_c$
Variante 3 – Request/Reply mit aktiven Lesen der Senke	$T_{v3} = 4 * t_c$
Variante 4 – Request/Reply mit aktiven Schreiben der Quelle	$T_{v4} = 4 * t_c$
Variante 5 – Request/Reply mit aktiver Seite durch den POL	$T_{v5} = t_{\text{pol}} + 8 * t_c$
Annahme	$t_c \ll t_{\text{brok}} \ll t_{\text{pol}}$

Es wird ersichtlich, dass Variante 2 mit T_{v2} die kleinste Übertragungszeit besitzt, aber zusätzliche Mechanismen zur Fehlerüberwachung auf der Empfängerseite benötigt, da diese eine passive Rolle in der Kommunikation einnimmt. Variante 5 besitzt mit T_{v5} die größte Übertragungszeit. Während sie zwar geringere Hardwarefunktionalität erfordert, sind aber auch hier zusätzliche Mechanismen zur Überwachung notwendig. Die Varianten 3 und 4 zeigen keinen Unterschied hinsichtlich der Übertragungszeit, wobei Variante 3 den Vorteil bietet, dass die Senke der aktive Kommunikationspartner ist. Somit sind keine gesonderten Mechanismen zur Überwachung notwendig. Zusammenfassend zeigt sich, dass alle Kommunikationsmuster unter Berücksichtigung ihrer Vor- und Nachteile in der Lage sind, eine Querkommunikation zu realisieren. Allerdings zeigen sich deutliche Unterschiede hinsichtlich der Übertragungszeiten. Vor allem fällt auf, dass Mechanismen zur direkten Querkommunikation aufgrund kürzerer Übertragungszeiten broker-basierten Mechanismen vorzuziehen sind. Aufgrund dessen sollen im weiteren Verlauf des Beitrags lediglich die Varianten 2-4 weiter untersucht werden.

6. OPC UA Client/Server-basierte Controller-Controller-Kommunikation

Eine Kommunikationstechnologie, die sich zur Umsetzung des Request/Reply-Patterns und damit für die Varianten 3 und 4 eignet, ist OPC UA Client/Server. Ansätze für eine konfigurierbare Querkommunikation basierend auf dieser Technologie werden bereits in [2] und [4] beschrieben. Dieser Abschnitt befasst sich mit einer auf das MTP ProcessValue Konzept angepassten Weiterentwicklung dieser Ansätze. Die Lösung ist in Form eines Software-Musters beschrieben und besteht aus dem Baustein *ConnectionManager* und entsprechenden *Reader*- bzw. *Writer-Bausteinen* für die jeweiligen ein- und ausgehenden Prozesswerte (siehe Bild 3).

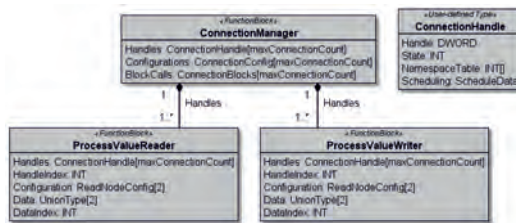


Bild 3: Darstellung des Softwaremusters für die Realisierung der konfigurierbaren Querkommunikation mittels Request/Reply-Mechanismus

Der *ConnectionManager* organisiert eine Liste von OPC UA Verbindungen, während die *Reader-/Writer-Bausteine* für die Lese- und Schreiboperationen verantwortlich sind. Die *Reader-/Writer-Bausteine* können über ihren *ConnectionIndex* einer vorhandenen Verbindung des *ConnectionManagers* zugeordnet werden. Das so entstehende adaptive Netzwerk aus Bausteinen ermöglicht es, die Informationen von beliebigen OPC UA Servern zu beziehen bzw. zu übertragen. Die Anzahl der möglichen Verbindungen ist abhängig von der eingesetzten Hardware. In gleichem Maße ist die Anzahl gleichzeitiger Lese- und Schreiboperationen möglicherweise begrenzt, wofür für jede Verbindung ein Read-/Write-Scheduler vorbereitet ist. Sowohl der *ConnectionManager* als auch die *Reader/Writer-Bausteine* nutzen bereits standardisierte Schnittstellen aus der PLC Open OPC UA Client-Baustein Spezifikation [16]. Die darin spezifizierten Schnittstellen werden über den integrierten OPC UA Server zur Konfiguration freigegeben.

In Bezug auf die Kommunikationsüberwachung ist der lesende und schreibende Zugriff zu unterscheiden. Im Falle des Reader-Bausteins ist die Senke auch gleichzeitig die aktive Seite der Kommunikation, weshalb Verbindungsabbrüche oder Fehler beim Lesen auch gleichzeitig auf der verarbeitenden Seite erkannt werden. Aus diesem Grund sind keine weiteren applikativen Maßnahmen erforderlich. Anders im Falle des Writer-Bausteins, bei dem die aktive Seite die Quelle der Information ist. In diesem Fall ist auf applikativer Ebene der Zeitstempel des letzten schreibenden Zugriffs aus dem OPC UA Server zu extrahieren und in dem Steuerungsprogramm zu verarbeiten.

7. OPC UA Publish/Subscribe-basierte Controller-Controller-Kommunikation

Mit Publish/Subscribe wird in diesem Abschnitt ein weiterer Kommunikationsmechanismus eingeführt, der für die Querkommunikation und den Austausch von Prozesswerten im Speziellen nach Variante 2 genutzt werden kann. Das Konzept führt sogenannte *ProcessValuePublisher* und *ProcessValueSubscriber* ein (siehe Bild 4). Da es sich bei Publish/Subscribe um ein

verbindungsloses Verfahren handelt, kann auf etwas Vergleichbares wie den ConnectionManager aus Abschnitt 6 verzichtet werden.

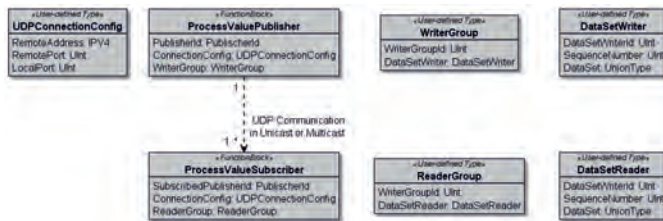


Bild 4: Darstellung des Softwaremusters zur Realisierung von Querkommunikation auf Basis eines Publish/Subscribe-Mechanismus mit UDP

Die Realisierung der beiden Funktionsbausteine basiert auf Konzepten der OPC UA Pub/Sub Spezifikation [12]. Die Kommunikationseinstellungen werden über eine Konfigurationsschnittstelle offengelegt. Hierzu dient der Datentyp *UDPConnectionConfig* mit den schreibbaren Variablen für die Remote-Adresse, den Remote-Port sowie den Port des eigenen Systems. Die Datentypen *WriterGroup*, *DataSetWriter*, *ReaderGroup* und *DataSetReader* sind analog aus [12] übernommen und auf die Übertragung eines Prozesswerts und Quality Codes ausgelegt. Eingehende Telegramme auf Seiten des Subscribers lösen ein Event aus, welches einen Zeitstempel generiert. Ist der Zeitstempel älter als eine einstellbare Dauer, so wird ein Fehler gemeldet und der Quality Code wechselt auf den Wert „Bad“. Auf diese Weise wird die Kommunikationsüberwachung in einem Pub/Sub-basierten Mechanismus realisiert.

8. Adaptive Logikfunktionen für die konfigurierbare Vorverarbeitung der Querkommunikation

Sowohl die Reader/Writer-Bausteine als auch die Publisher/Subscriber-Bausteine sind um die Funktion der adaptiven Logik erweitert. Das Konzept der adaptiven Logik stammt aus [4] und wurde im Kontext der Automatisierungsdienste-Choreografien erstmalig veröffentlicht. Es beschreibt einen Mechanismus, mit dem logische und/oder arithmetische Ausdrücke realisiert werden können. Auf diese Weise können konfigurierbare Vorverarbeitungsfunktionen für ein- und ausgehende Informationsflüsse realisiert werden. Hierzu ist der in Bild 5 dargestellte Funktionsbaustein *AdaptiveProcessValueLogic* in die Kommunikationsbausteine eingebunden. Analog zum Prinzip der Eingabe-Verarbeitung-Ausgabe ist die Vorverarbeitungsfunktion zwischen den Input- und Outputgrößen durch eine Verkettung von Logikblöcken zu bilden.

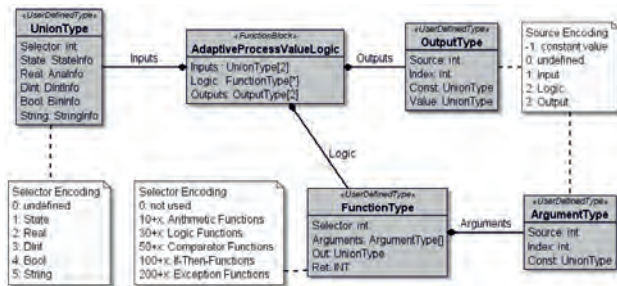


Bild 5: Darstellung des Softwaremusters zur Realisierung adaptiver Logik in Funktionsbausteinen konfigurierbarer Querkommunikation

Zur Konfiguration eines Logikblocks dient der Datentyp *FunctionType*. Ist eine Kommunikationsverbindung aktiv, wird die konfigurierte Funktion ausgeführt. Durch typischere Verarbeitung der adaptiven Logik und Prüfung von Quell- und Zieldatentypen während der Verarbeitung werden bei ungültigen Konfigurationen entsprechende Fehlercodes im System angezeigt, womit Fehler einfach zu detektieren sind. Für weitere Details zur Umsetzung und Funktionsweise der adaptiven Logik sei auf die bereits genannte Veröffentlichung verwiesen [4].

9. Evaluation der Lösungskonzepte für die Querkommunikation in modularen Anlagen

Die Evaluation der hier vorliegenden Arbeit ist dreiteilig – Teil 1 ist eine Architekturanalyse, die die gestellten Anforderungen auf konzeptioneller Ebene prüft und bewertet, Teil 2 ist eine Latenzmessung unter Laborbedingungen und Teil 3 ist ein kontrolliertes Experiment mit einer Anlagensimulation.

Teil-Evaluation 1 – Architekturanalyse

Im Rahmen einer Architekturanalyse wurden die in Abschnitt 5 vorgestellten Kommunikationsmuster in Kombination mit den in den Abschnitt 6-8 vorgestellten Umsetzungsansätzen gegen die in Abschnitt 3 identifizierten Anforderungen geprüft, mit dem folgenden Ergebnis:

- **Anforderung 1 – Forderung zur Umsetzung des Informationsflusses:** Die in den Abschnitten 6 und 7 vorgestellten Bausteine ermöglichen den grundsätzlichen Informationsaustausch über OPC UA Client/Server bzw. Publish/Subscribe. Zudem bieten sie die Möglichkeit, Informationsflüsse flexibel zu konfigurieren oder umzukonfigurieren. Somit wird diese Anforderung für alle drei Varianten als *erfüllt* betrachtet.
- **Anforderung 2 – Forderung zur Vorverarbeitung auszutauschender Informationen:** Das in Abschnitt 8 vorgestellte Konzept der adaptiven Logik ermöglicht eine Vorverarbeitung der ausgetauschten Informationen. Sowohl die OPC UA Client/Server - als auch

Publish/Subscribe-basierten Bausteine sind um diese Funktionalität erweitert. Somit wird diese Anforderung für alle drei Varianten als *erfüllt* betrachtet.

- **Anforderung 3 – Möglichkeit zur Überwachung des Informationsflusses:** Es zeigt sich, dass für alle drei betrachteten Varianten die Möglichkeit einer Überwachung des Informationsflusses besteht, allerdings mit unterschiedlichem Aufwand. Bei den Varianten 2 und 4 ist der Sender der aktive Kommunikationspartner, somit muss in diesen Fällen ein zusätzlicher Aufwand für die Überwachung betrieben werden. In Variante 3 ist der Empfänger der aktive Kommunikationspartner, somit ist eine Kommunikationsüberwachung intrinsisch sichergestellt. Nichtsdestotrotz kann diese Anforderung für alle drei Varianten als *erfüllt* betrachtet werden.

Teil-Evaluation 2 – Latenzmessung unter Laborbedingungen

Für die Latenzmessungen wurde ein Laboraufbau bestehend aus zwei S7-1511 Steuerungen gewählt, welche über einen Switch miteinander verbunden sind. Ebenfalls am Switch angebunden war ein Industrie-PC mit der Software SIMATIC TIA Portal, welcher zum Laden der SPS-Software genutzt wurde und als Network Time Protocol (NTP)-Master für die Uhrzeitsynchronisation diente. Je Variante wurde eine Messreihe von 1000 Messungen mit einem String-Prozesswert durchgeführt. Hierzu wurde ein generischer String mit 255 Bytes, bestehend aus 254 Zeichen und einem Byte für den Quality Code, übertragen. Der Zeitpunkt des Sendens und des Empfangens wurden auf den jeweiligen Steuerungen mittels Trace-Funktionen erfasst. Beide Datensätze wurden mittels Microsoft Excel zusammengeführt. Das Ergebnis ist ein Boxplot-Diagramm, das die untersuchten Varianten 2-4 sowie zum Vergleich die bestehende Variante 5 enthält (siehe Bild 6).

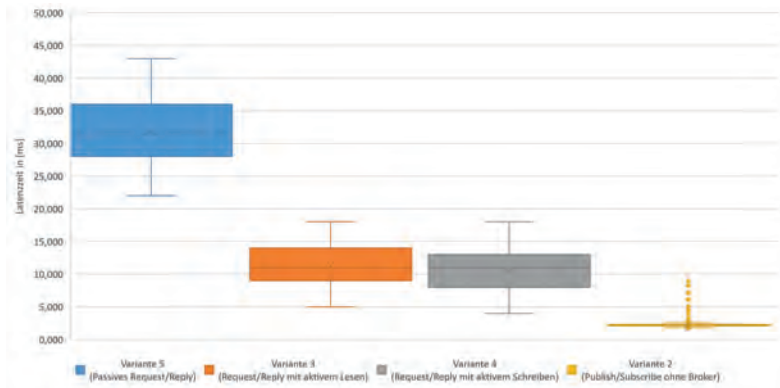


Bild 6: Darstellung der Latenzzeit-Messungen zwischen Sender und Empfänger für die 3 neuen Konzepte und im Vergleich Variante 5 mit dem derzeitig standardisierten Verfahren

Aus den Messungen zeigt sich, dass Variante 2, wie zu erwarten, die schnellste Variante darstellt. In gleichem Maße zeigt sich, dass die Varianten 3 und 4 ähnliche Latenzzeiten aufweisen. Variante 5, als die derzeitige standardisierte Variante, ist, wie zu erwarten, die langsamste Variante. Der doch deutliche Unterschied zwischen Variante 2 und Variante 3/4 liegt vermutlich an der Umsetzung der dahinter liegenden Betriebssystem-Funktionen der Steuerungen. Während die OPC UA Client-Bausteine auf Ebene der Applikation in Bezug auf das ISO/OSI-Schichtenmodell arbeiten, ist die Übertragung mittels UDP-Protokoll in Variante 2 direkt auf Ebene der Transportschicht umgesetzt. Somit hat diese Variante im Vergleich zu Variante 3 und 4 weniger Software-Schichten zu durchlaufen und ist in Summe schneller.

Teil-Evaluation 3 – Kontrolliertes Experiment am Beispiel einer simulierten Anlage

Zur Evaluation der adaptiven Logik (Abschnitt 8) im Zusammenspiel mit der konfigurierbaren Kommunikation (Abschnitt 6 und 7) wurde ein kontrolliertes Experiment an einer simulierten Prozessanlage durchgeführt. Hierzu wurde der MTP-Achema-Demonstrator (siehe Bild 7) aus dem Jahre 2018, bestehend aus einem Dosierer, einem Reaktor und einer Temperieren-Package-Unit, genutzt. Er enthält sowohl modulübergreifende Regelkreise als auch Verriegelungen zur Demonstration der direkten Querkommunikation.

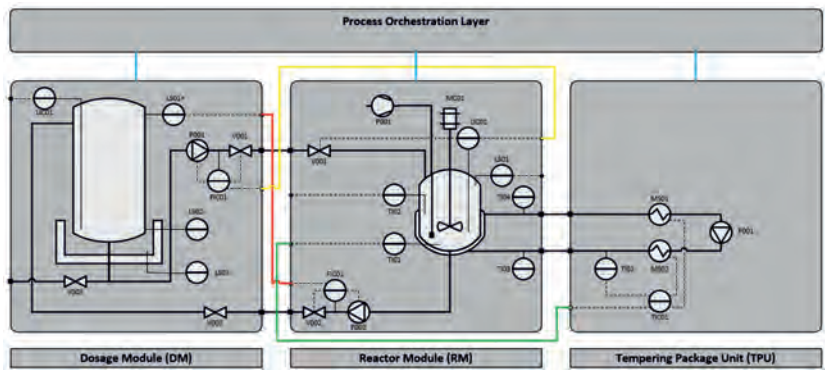


Bild 7: Schema-Architektur einer modularen Anlage mit farblich dargestellten Interaktionswegen zwischen dem Process Orchestration Layer und Modulen, sowie von Modul zu Modul

Es wurden aus dem Anwendungsbeispiel drei modulüberreifende Verbindungen mit den Varianten 2-4 umgesetzt, die im Folgenden näher beschrieben werden. Ergänzend ist die Konfiguration der adaptiven Logik angegeben.

Verbindung 1 – Temperatur-Kaskade (grün)

Zum Schließen des kaskadierten Regelkreises zwischen der Reaktor-Sumpftemperatur und der Temperieren Package Unit ist eine Interaktion zwischen dem Temperatur-Sensor TI01 des Reaktors und ExtTemp-Eingang der Temperatur-Regelung TIC01 der Temperieren Package Unit zu konfigurieren. In diesem Beispiel wird der Temperaturwert des Reaktors in Grad Fahrenheit bereitgestellt, während die Package Unit einen Wert in Grad Celsius benötigt. Zur Umrechnung von Grad Fahrenheit nach Grad Celsius ist die in Tabelle 2 angegebene Konfiguration an den Logik-Blöcken einzustellen.

Tabelle 2: Konfigurationsdaten der adaptiven Logik für die Verbindung der Temperatur-Kaskade

In- dex	Selector	IN1		IN2		IN3	IN4
		Source	Index/Value	Source	Index/Value	Source	Source
0	SUB (11)	1 (Input)	0 (PV)	-1 (Const)	32.0 (Real)	0 (not used)	0 (not used)
1	DIV (13)	2 (Logic)	0 (OUT of SUB)	-1 (Const)	1.8 (REAL)	0 (not used)	0 (not used)

Verbindung 2 – Füllstandsüberwachung des Reaktor (gelb)

Überschreitet der Reaktorfüllstand einen definierbaren Wert, so ist der weitere Zulauf durch Abschaltung der Zulauf-Pumpe zu unterbinden. Hierfür ist der Füllstandswert auf das Überschreiten eines Grenzwerts zu prüfen. Die dafür notwendige Konfiguration der adaptiven Logik ist in Tabelle 3 dargestellt.

Tabelle 3: Konfigurationsdaten der adaptiven Logik für die Füllstandsüberwachung des Reaktors

In-dex	Selector	IN1		IN2		IN3	IN4
		Source	Index/Value	Source	Index/Value	Source	Source
0	CMP_GT (51)	1 (Input)	0 (PV)	-1 (Const)	75.0 (Real)	0 (not used)	0 (not used)

Verbindung 3 – Füllstandsüberwachung des Dosierers (rot)

Diese Füllstandsüberwachung nutzt einen null-aktiven Grenzwertschalter am Tank des Dosiersystems. Die Förderpumpe des Reaktors ist, wenn der Grenzwertschalter LS01+ des Dosierers aktiviert wird, abzuschalten. Das Abschaltsignal ist Eins-aktiv und muss somit durch eine Negierung entsprechend angepasst werden, die einzustellende Konfiguration der adaptiven Logik ist in Tabelle 4 dargestellt.

Tabelle 4: Konfigurationsdaten der adaptiven Logik für die Füllstandsüberwachung des Dosierer-Tanks

In-dex	Selector	IN1		IN2	IN3	IN4
		Source	Index/Value	Source	Source	Source
0	NOT (11)	1 (Input)	0 (PV)	0 (not used)	0 (not used)	0 (not used)

Alle drei Varianten konnten für die drei umgesetzten Verbindungen im Betrieb mit einem Testrezept ihre Anwendbarkeit und Funktionsfähigkeit unter Beweis stellen.

10. Zusammenfassung und Ausblick

Querkommunikation in modularen Anlagen ist wichtig und notwendig. Sie eröffnet Freiheiten im Modulschnitt, damit Wiederverwendbarkeit und Anpassbarkeit der modularen Prozesseinheiten sichergestellt werden kann. Wie Abschnitt 5 gezeigt hat, bestehen mit dem derzeitigen standardisierten Verfahren (Variante 5) und mit den hier vorgestellten neuen Ansätzen (Variante 2-4) eine Vielzahl von Realisierungsmöglichkeiten für die Querkommunikation in verschiedenen Anwendungsfällen im Bereich modularer Anlagen.

Es hat sich gezeigt, dass durch die Offenlegung der Konfigurationsschnittstellen von Variante 2-4 eine performante konfigurierbare direkte Querkommunikation inklusive konfigurierbarer Vorverarbeitungsfunktionen umgesetzt werden kann. Im Vergleich mit dem Stand der Technik lassen sich die hier vorgestellten Lösungen technologisch zwischen dem heutigen in [5] standardisierten Verfahren und den Konzepten der Automatisierungsdienste-Choreografien [4] einordnen. Wie sich in der Evaluation gezeigt hat, stellen die Varianten 2-4 dabei eine sehr gute Alternative zum standardisierten Verfahren dar.

Seit kurzem existiert eine weitere Möglichkeit, eine direkte Querkommunikation umzusetzen: Mit OPC UA FX ist ein neuer, durchaus für das MTP-Konzept interessanter Ansatz geschaffen worden. Die darin eingeführten Offline-Beschreibungen und funktionalen Schnittstellen weisen einige Parallelitäten zum MTP-Konzept auf, weshalb bereits an Fortsetzungen dieses Beitrags gearbeitet wird, in denen OPC UA FX und MTP-Konzepte gegenübergestellt werden. Die daraus entstehenden Ergebnisse sowie die in diesem Beitrag vorgestellten Ansätze werden der noch zu gründenden Taskforce zur Querkommunikation als Grundlage zur Verfügung gestellt und unter den teilnehmenden Experten weiterführend diskutiert.

Literaturverzeichnis

- [1] NAMUR NE 171:2020-12., "Anwendung des modularen Anlagenkonzeptes in der produktionsnahen Logistik"
- [2] A. Stutz, M. Maurmaier, "Dezentrale Kommunikationsarchitektur für direkte Modul-Modul-Interaktion basierend auf standardisierten Elementen der VDI/VDE/NAMUR 2658; VDI-Bericht Nr. 2351, 343- 360; Automation 2019, Baden-Baden
- [3] *VDI/VDE/NAMUR 2658: Automatisierungstechnisches Engineering modularer Anlagen in der Prozessindustrie*. <https://www.vdi.de/2658>
- [4] A. Stutz, A. Fay, M. Barth and M. Maurmaier, "Software Patterns for the Realization of Automation Service Choreographies," 2021 26th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), 2021, pp. 1-8, doi: 10.1109/ETFA45728.2021.9613201.
- [5] *VDI/VDE/NAMUR 2658-4 (2022): Automatisierungstechnisches Engineering modularer Anlagen in der Prozessindustrie – Modellierung von Moduldiensten*. <https://www.vdi.de/2658>, abgerufen am: 02.05.2022 (in print)
- [6] *VDI/VDE/NAMUR 2658-1 (2022): Automatisierungstechnisches Engineering modularer Anlagen in der Prozessindustrie – Allgemeines Konzept und Schnittstellen*. <https://www.vdi.de/2658>, abgerufen am: 02.05.2022 (in print)
- [7] A. Klose et al., "Orchestration Requirements for Modular Process Plants in Chemical and Pharmaceutical Industries", Chemical Engineering & Technology, Vol. 42, WILEY-VCH Verlag GmbH & Co. KgaA, Weinheim, 2282-2291, 2019
- [8] S. Grüner, et al., "A Reference Architecture for Modular Industrial Automation Systems," 2021 26th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), 2021, pp. 1-8, doi: 10.1109/ETFA45728.2021.9613203.
- [9] K. Stark, R. Braun, M. Hoernicke, T. Holm, A. Stutz; Modul-zu-Modul-Kommunikation in modularen Anlagen - Modul-zu-Modul Kommunikation mit erweitertem Funktionsumfang; VDI-Bericht Nr. 2351, 331- 342; Automation 2019, Baden-Baden
- [10] Guidelines for Packaging Machinery Automation V3.0, Open Modular Architecture Controls (OMAC) Users Group, 22.10.2004
- [11] OPC UA 30050 - OPC UA for PackML; OPC Foundation / OMAC; Release 1.01; 11.11.2020
- [12] OPC 10000-14 – OPC Unified Architecture Part 14: PubSub; OPC Foundation, Release 1.05.01; 22.02.2022

- [13] C2C-Kommunikation via OPC UA PubSub mit SIMATIC S7-1500 auf Basis von UDP, Siemens AG, 18.01.2021, <https://support.industry.siemens.com/cs/document/109782455>
- [14] Extending OPC UA to the field: OPC UA for Field eXchange (FX), OPC Foundation, 11.2021, Version 2, <https://opcfoundation.org/wp-content/uploads/2020/11/OPCF-FLC-Technical-Paper-C2C.pdf>, abgerufen am 03.05.2022
- [15] S. Höme, et al., Sichere interoperable Maschine-zu-Maschine Kommunikation basierend auf OPC UA Field Exchange, AUTOMATION 2022, Baden-Baden, (in print)
- [16] PLCopen OPC UA Client for IEC 61131-3, Release V1.1, PLCopen 03.05.2022, https://plcopen.org/system/files/downloads/plcopen_opcua_client_for_iec_61131-3_version_1.1.pdf

Gestaltung einer Forschungsinfrastruktur für die Anwendung digitaler Werkzeuge in Cyber-Physischen Produktionssystemen

I. Gräßler, M. Hieb, D. Roesmann,

Heinz-Nixdorf-Institut, Lehrstuhl für Produktentstehung,
Universität Paderborn

Kurzfassung

In Anbetracht des technologischen Wandels von Industrie 4.0 rückt eine vollumfassende, standardbasierte Datenintegration zunehmend in den Vordergrund. Die Erforschung von Potenzialen und Anwendungsgebieten von OPC UA in Cyber-Physischen Produktionssystemen erfordert die Gestaltung einer flexiblen Forschungsinfrastruktur für die Anwendung digitaler Werkzeuge. Die Vernetzung der digitalen und virtuellen Werkzeuge mit dem Smart Automation Laboratory des Heinz Nixdorf Instituts erfolgt mittels des Kommunikationsstandards OPC UA und ermöglicht variable Prüfstandskonfigurationen. Datengetriebene Ansätze der digitalen und virtuellen Produktentstehung können dadurch kostengünstig getestet und validiert werden mit dem Ziel der Vorbereitung des schnellen Transfers in die industrielle Praxis.

1. Einleitung

Die Einführung von Informations- und Kommunikationstechnologien führt zu neuen Möglichkeiten hinsichtlich Vernetzung und Automatisierung von Produktionssystemen. Der politisch geprägte Begriff „Industrie 4.0“ [1] steht dabei für den technologischen Wandel in der Automatisierung, der Digitalisierung und der Vernetzung der Systeme mit dem Internet der Dinge und Dienste. Die Entwicklung von Cyber-Physischen Produktionssysteme (CPPS) ist derzeit ein dominierendes Forschungsfeld in der Produktions- und Automatisierungstechnik. Der Begriff CPPS leitet sich vom Begriff der Cyber-Physischen Systeme (CPS) ab. Im Gegensatz zu konventionellen Produktionssystemen erfüllen CPPS ihre Funktion unter Verwendung von Dingen und Diensten über das Internet. Sie sind in der Lage, Prozessinformationen zu sammeln, was die Integration datengetriebener Methoden und Algorithmen ermöglicht [2]. Die Implementierung eines CPPS ist somit Grundlage einer sogenannten Smart Factory. Ressourcen, Produkte und Systemelemente sind mit geeigneten Sensoren und Aktoren ausgestattet und organisieren den gesamten Produktionsprozess autonom [3].

Die Anwendung von CPPS in der industriellen Praxis scheitert häufig an unternehmensspezifischen und zeitlichen Restriktionen. Forschungsinfrastrukturen erlauben es, Szenarien der

Produktion des betrachteten Unternehmens durchzuführen, ohne Störungen im realen Prozess und ohne aufwändige Voraussetzungen wie Personalschulung zu schaffen. Sie ermöglichen die Simulation der Struktur und des Verhaltens von CPPS in einer realen Umgebung. Dadurch können unterschiedliche Konzepte und Möglichkeiten getestet werden, um somit den Transfer in die industrielle Praxis vorzubereiten. Herausforderung des vorliegenden Beitrages ist der Aufbau einer Forschungsinfrastruktur für Cyber-Physische Produktionssysteme. Die Umsetzung erfolgt mit dem Kommunikationsprotokoll OPC Unified Architecture (OPC UA¹). Zunächst wird das Vorgehen bei der Einführung von OPC UA im Smart Automation Laboratory des Heinz Nixdorf Instituts in Paderborn, einer Forschungsinfrastruktur als Experimentier- und Validierungsumgebung für die Produktionsplanung und -steuerung, beschrieben. Die einzelnen Elemente des Produktionssystems werden mit Computersystemen sowie geeigneten Sensoren zur Erfassung des Maschinenverhaltens nachgerüstet. Das Peer-to-Peer-Funknetzwerk mithilfe von OPC UA stellt die bidirektionale Schnittstelle zur Datenintegration von realen in die virtuelle Welt dar. Echtzeitdaten werden im laufenden Produktionsprozess aus unterschiedlichen Systemelementen und Datenquellen erfasst. Anschließend werden digitale Werkzeuge für spezifische Anwendungen auf dem Shopfloor beschrieben, die auf dem Funknetzwerk beruhen. Abschließend werden die Potenziale der entwickelten Infrastruktur für die Praxis und die Forschung im Rahmen von CPPS dargestellt.

2. Stand der Technik

Die Einführung neuer Technologien in der Produktion führt zu einer neuartigen Vernetzung, Automatisierung und Digitalisierung von Produktionssystemen. Dies ermöglicht die Entwicklung von sogenannten CPPS. Für die Erforschung innovativer Konzepte ist der Aufbau einer geeigneten Forschungsinfrastruktur (Abschnitt 2.1) mit Peer-to-Peer Netzwerk (Abschnitt 2.3) für CPPS notwendig. Dies führt zu einer schrittweisen Auflösung der klassischen Automatisierungspyramide (Abschnitt 2.2).

2.1 Forschungsinfrastruktur

Nach dem Bundesministerium für Bildung und Forschung [4] bilden Forschungsinfrastrukturen die Grundlage für Spitzenforschung und sind die Voraussetzung für die Generierung neuen Wissens und neuer Technologien. Das Bundesministerium definiert Forschungsinfrastrukturen als Einrichtungen, Anlagen, Ressourcen und Dienstleistungen, die von Wissenschaftler*innen aller wissenschaftlich-technologischen Gebieten für die Forschung benötigt werden [5]. Im

¹ OPC UA – Open Platform Communicaitons Unified Architecture

Rahmen von CPPS dienen Forschungsinfrastrukturen zur Untersuchung von Struktur, Anforderung und Verhalten. Sie werden eingesetzt zur Entwicklung, Einführung und Erweiterung von CPPS. Ziel ist die Vorbereitung des Transfers in die industrielle Praxis. Anforderungen für die Anwendung eines Produktionslabors als Forschungsinfrastruktur für CPPS sind beispielsweise in Gräßler & Pöhler [6] aufgelistet. Diese Liste umfasst unter anderem ein breites Anwendungsspektrum unterschiedlicher Einsatzszenarien, eine aufwandsarme Modellierung zur Gewährleistung von Übertragbarkeit sowie eine Adaption auf spezifische Stakeholder.

2.2 Cyber-Physische Produktionssysteme

Grundlage für den Aufbau von CPPS sind CPS. Zunächst führte Gill [7] den Begriff „Cyber-Physischen System“ im Jahr 2006 ein. CPS sind stark vernetzte technische Elemente, die aus eingebetteten Systemen bestehen, digitale Informationen über Internettechnologien austauschen und weitere Dienste nutzen können. Mit dem Einsatz geeigneter Sensoren nehmen sie Daten aus der Umwelt auf, werten sie mithilfe von Diensten aus und beeinflussen die Umwelt mit Aktoren [8]. Lee [9] definiert den Begriff wie folgt: *„Cyber-Physical Systems (CPS) is an integration of computation with physical processes. Embedded computers and networks monitor and control the physical processes, usually with feedback loops where physical processes affect computations and vice versa“*. Lee et al. [10] beschreiben mit der 5C-Architektur einen schrittweisen Leitfaden für die Entwicklung und den Einsatz von Cyber-Physischen Systemen im Produktionskontext:

1. Die Smart **Connection** Ebene zielt auf die Aufnahme und das Verfügbarmachen verlässlicher Daten über Sensoren und/oder IT-Systeme ab.
2. Die Data-to-Information **Conversion** Ebene beinhaltet die Extrahierung relevanter Informationen aus den Daten,
3. Die **Cyber** Ebene umfasst den Aufbau einer gemeinsamen Informationsbasis aller miteinander vernetzten Systemelemente durch Datenaustausch,
4. Die **Cognition** Ebene schließt die Aufbereitung von Informationen zur Entscheidungsunterstützung der Anwender*innen ein und
5. Die **Configuration** Level Ebene als Kontrollinstanz zur Selbstorganisation von Systemen.

Der Einsatz von CPS auf dem Shopfloor führt zu einem grundlegenden Wandel der Planung und Steuerung der Produktion. Auf dem Shopfloor handelt es sich bei CPS beispielsweise um intelligente Werkzeugmaschinen, intelligente Transportsysteme, kamerabasierte Qualitätskontrollen und adaptive Assistenzsysteme. Sie zeichnen ihren aktuellen Zustand und den Zustand ihrer Umgebung auf und kommunizieren ihn untereinander. Die Kommunikation erfolgt mittels

Internettechnologien. Eine derartige Umsetzung auf dem gesamten Shopfloor wird als CPPS bezeichnet. Die Einführung von CPPS und der resultierenden Vernetzung der Materialien, Mitarbeiter*innen, Maschinen und Materialflusssystemen mit dem Internet der Dinge und Dienste führt zu einer disruptiven Entwicklung in der Produktion [11, 12]. CPPS zeichnen sich durch Adaptivität, Flexibilität und eine hohe Effizienz aus [13]. Die Einführung von CPPS führt zu einer Veränderung der klassischen Planungs- und Kontrollsysteme in der Produktion. Bestehende Hierarchien werden aufgebrochen und sukzessive die klassische Automatisierungspyramide aufgelöst. Dafür entstehen sogenannte CPPS-Netzwerke. Grundsätzlich wird zwischen zwei verschiedenen Arten der Vernetzung von Systemen und Diensten unterschieden: Die **vertikale Integration** ermöglicht es jedem Systemelement (z. B. Maschinen, Transportsysteme, Betriebsmittel und Software-Diensten), sich mit anderen Teilnehmern auszutauschen; sie wird in einem Unternehmen umgesetzt. Die **horizontale Integration** umfasst die Vernetzung von Systemen und Diensten entlang des Produktentstehungsprozesses über Unternehmensgrenzen hinweg. Dies ist beispielsweise der Datenaustausch mit den Kunden und Lieferanten.

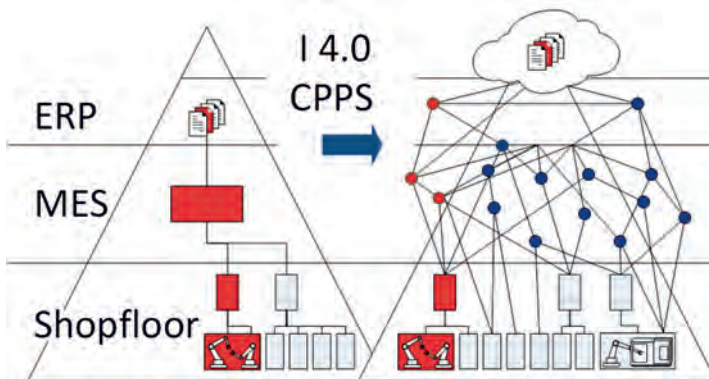


Bild 1: Änderungen der Kommunikation durch die Digitalisierung [14]

2.3 Kommunikationsprotokolle

Kommunikationsprotokolle sind notwendig für den Austausch immer größerer Datenmengen über Maschinen und Automatisierungsebenen hinweg. Die Einführung von Informations- und Kommunikationstechnologien setzt voraus, dass vernetzte Systeme Daten elektronisch übertragen und entgegennehmen können. Zur Darstellung der Kommunikationstechnik hat sich das ISO/OSI-7-Schichtenmodell [15] (englisch Open Systems Interconnection Model) etabliert. Das ISO/OSI-7-Schichtenmodell (siehe Bild 2) dient als Referenzmodell für Kommunikationsprotokolle und teilt Protokolle in Schichten, die von der Bitübertragung bis zur Anwendungsschicht

unterteilt sind. Die verschiedenen Funktionen der einzelnen Schichten werden in verschiedenen Literaturquellen im Detail beschrieben [16, 17]. Grundsätzlich teilt sich das ISO/OSI-7-Schichtenmodell in den netzwerkorientierten (bis einschließlich Ebene vier) und den anwendungsspezifischen Teil (von Ebenen fünf bis sieben) auf. Das ISO/OSI-7-Schichtenmodell gibt Auskunft darüber welche Aufgaben die einzelnen Schichten ausführen.

OPC Unified Architecture

Das Kommunikationsprotokoll OPC UA (Open Platform Communications Unified Architecture) geht auf die OPC Foundation zurück und vereint die bis dato bestehenden OPC Kommunikationsprotokolle (OPC Classic, DA², HD³ und AE⁴) [18]. Motiviert wurde OPC UA vor dem Hintergrund, ein möglichst plattformunabhängiges, robustes, skalierbares und sicheres Kommunikationsprotokoll zu schaffen. Zudem verzichtet das OPC UA Kommunikationsprotokoll auf den Einsatz von Microsoft COM/DCOM und schafft somit eine Alternative.



Bild 2: OPC UA gemappt auf ISO/OSI-7-Schichtenmodell [18]

OPC UA transportiert Daten über Firewall freundliche Web-Services (SOAP/HTTP), oder über optimierte TCP/IP-basierende Protokolle. Dies ermöglicht Daten interoperable und plattformunabhängig von Sensoren bis hin zu MES, ERP-Systemen sowie Cloud-Anwendungen zu verschicken.

² DA – Data Access

³ HD – Historical Data Access

⁴ AE – Alarm and Events

4. Evaluierung

Die Implementierung einer Peer-to-Peer-Netzwerks mithilfe von OPC UA ermöglicht unterschiedliche Anwendungsfelder im Rahmen der Digitalen und Virtuellen Produktion. Nachfolgend werden exemplarisch vier umgesetzte Beispielanwendungen dargestellt.

4.1 Produktionsplanung

Die Modellierung komplexer Produktions-, Informations- und Logistikprozesse in Simulationen dient grundlegend der Optimierung und Analyse unterschiedlicher Unternehmensprozesse. Die Software Anwendung Tecnomatix Plant Simulation von Siemens wird zur Planung und Steuerung von Produktionssystemen eingesetzt. Simulation, Visualisierung, Analyse und Optimierung von Produktions- und Logistikprozessen gehören zu den Hauptanwendungsgebieten von Plant Simulation [19]. Um das Potenzial von Simulationen zu nutzen, ist die Genauigkeit der Modellierung des Shopfloors von entscheidender Bedeutung. Für die software-technische Umsetzung der Simulation wurde das Plant Simulation Softwarepaket OPC UA genutzt, siehe Bild 4. Die Schnittstelle zwischen Simulation und der Anbindung über OPC UA ermöglicht Echtzeitkommunikation. Simulationen werden um Echtzeitdaten angereichert sowie für die Optimierung und Analyse bestehender Produktionsprozesse herangezogen. Die Simulationen ermöglicht es insbesondere anhand historischer Daten, Prognosen für zukünftige Aufträge zu bestimmen, Anomalien im Produktionsprozess zu identifizieren und Bottlenecks frühzeitig zu erkennen.

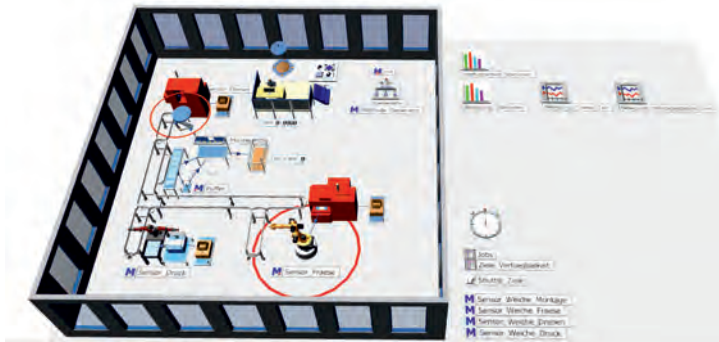


Bild 4: Simulation des Smart Automation Laboratory in Tecnomatix Plant Simulation

4.2 Andon Board

Die Verfügbarkeit von KPIs ist für Entscheidungen auf dem Shopfloor von hoher Bedeutung. Andon-Boards sind hierzu bekannte Hilfsmittel aus dem Lean Management. Andon-Boards visualisieren die Performanz eines Produktionssystems und helfen den Mitarbeiter*innen und Führungskräften als Unterstützung für Entscheidungsfindung für die Produktionssteuerung [20]. Um

das Potenzial aus dieser Visualisierungsform zu nutzen, ist das Design und der Inhalt von entscheidender Bedeutung. Für die softwaretechnische Umsetzung des Andon-Boards wurde das flow-based Visual Programming Tool Node-RED ausgewählt. Node-RED ist ein kostenloses Open Source Tool, welches ursprünglich von IBM entwickelt wurde. Es dient der Verknüpfung von Geräten, APIs und online Services im Sinne des Internet der Dinge (IoT). Bild 5 zeigt das Andon-Board, welches als Informationsbasis für die Organisation im Smart Automation Laboratory steht. Daraus können Zielabweichungen sofort erkannt und Maßnahmen eingeleitet werden. Die generierte Visualisierung wird durch die OPC UA-Schnittstelle in Echtzeit aktualisiert. Sie kann an unterschiedlichen Endgeräten im gesamten Produktionssystem angezeigt werden.

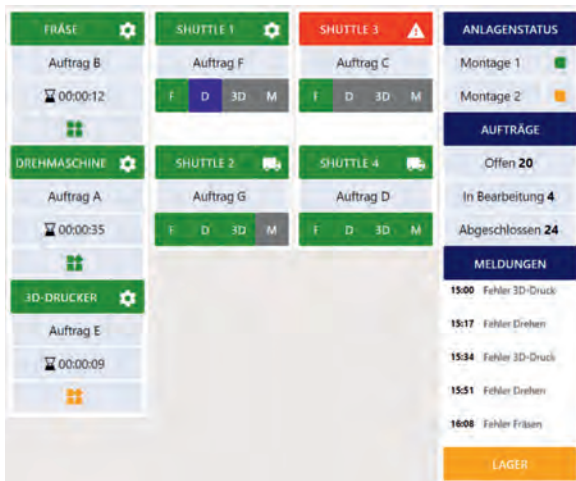


Bild 5: Homescreen des entwickelten Andon-Boards des Smart Automation Laboratory

4.3 Virtual Reality

Die Virtual Reality (VR) bietet die Immersion einer völligen synthetisch generierten Umgebung. Der Einsatz von VR in der industriellen Produktion wurde bereits in zahlreichen Anwendungen gezeigt [21]. Vielversprechende Anwendungsgebiete von VR in der Digitalen Produktion sind die Fabrikplanung, Fertigung, Montage und Services. Auch das Smart Automation Laboratory existiert als VR-Szene (Bild 6). Relevante Software-Tools sind dabei Unity VR, 3DS MAX sowie die VR-Brille Oculus Quest 2. Die VR-Szene wurde um eine OPC UA Schnittstelle erweitert und bietet somit eine Kommunikation in Echtzeit. Dadurch können aktuelle Zustände der Elemente des Produktionssystems erlebbar dargestellt werden. Beispielsweise wird die aktuelle Position der Shuttles über OPC UA kommuniziert und dem Anwender visualisiert. Die echtzeitnahe

Visualisierung kann insbesondere eingesetzt werden, um die Remote-Wartungen zu unterstützen, Kundenservices anzubieten, indem die Produktion aus der Ferne betrachtet werden kann.



Bild 6: Virtual Reality Szene des Smart Automation Laboratory

4.4 Planning & Scheduling

Das Management der Zeitwirtschaft ist eine wesentliche Aufgabe in der Steuerung von Produktionssystemen. Eine entscheidende Größe hierbei sind die Vorgabezeiten für die Maschinen sowie Mitarbeiter*innen. Um Abweichungen zu vermeiden, eine hohe Auslastung der Produktion sicherzustellen sowie eine hohe Verlässlichkeit der Vorkalkulation zu garantieren ist eine hohe Genauigkeit der Vorgabezeiten notwendig [22]. Der Einsatz des Peer-to-Peer-Netzwerks mit OPC UA ermöglicht eine einfachere Bereitstellung der anfallenden Daten. Auf Basis von auftragsspezifischen, digitalen, historischen Daten lassen sich mithilfe der Methoden der Data Analytics die Genauigkeit von Vorgabezeiten verbessern. Diese werden im ERP- und MES-System für Planning & Scheduling genutzt. Die Auswertung der Daten der einzelnen Arbeitsstationen umfasst nach Fayyad et al. [22] die fünf Schritte Selection, Preprocessing, Transforming, Data Mining und Interpretation/ Evaluation. Ein Spezialfall für das Scheduling im Smart Automation Laboratory sind die manuellen Montageprozesse [23]. Im Gegensatz zu den statischen Fertigungszeiten sind sie deutlich dynamischer und nicht nur produktabhängig, sondern hängen auch im wesentlichen Maße von den Fähigkeiten der Mitarbeiter*innen ab. Lernkurven ermöglichen die Modellierung von Leistungsverbesserungen der Mitarbeiter*innen im Laufe der Zeit. Durch die Wiederholungen von Montagevorgängen findet ein informeller Lernprozess statt. Durch die Integration von historischen Daten erhalten wir individuelle Lernkurven je Mitarbeiter*in und können genauere Prognosen über die benötigte Zeit in der Montage der zukünftigen Aufträge treffen. Um Rückschlüsse auf die individuelle Fähigkeitsentwicklung zu erhalten, wird eine Regressionsanalyse angewendet. Diese approximiert auf Basis diskreter Daten die individuelle Lernkurve der Mitarbeiter*innen. Die Umsetzung erfolgt im Softwareprogramm Matlab.

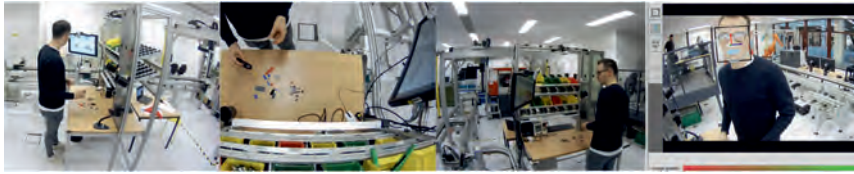


Bild 7: Montage mit adaptiven Assistenzsystem [22]

5 Zusammenfassung und Ausblick

Die Ergebnisse des Forschungsbeitrags zeigen die große Bedeutung und die vielfältigen Anwendungsgebiete einer Forschungsinfrastruktur für CPPS. Speziell die Nutzung neuer digitaler Methoden und Werkzeuge wird durch ein Peer-to-Peer-Netzwerk mithilfe des Kommunikationsprotokolls OPC UA ermöglicht. Dies wird anhand von vier Beispielanwendungen digitaler Werkzeuge zur Vernetzung und Automatisierung aufgezeigt. Dies umfasst die Software Tecnomatix Plant Simulation zur Produktionsplanung, Node-Red zum Visuellen Management auf Shopfloor, Unity3D zur Echtzeit-Visualisierung und Matlab zum Scheduling des Produktionsprozesses. Aufgrund des eingesetzten Datenstandards OPC UA existieren bereits definierte Schnittstellen zu den einzelnen Softwareprogrammen. Die bidirektionale Schnittstelle von OPC UA ermöglicht insbesondere eine umfassende Datenintegration aus Systemelementen des CPPS der Forschungsinfrastruktur mit unterschiedlichen digitalen Werkzeugen. Aktuell eingesetzte Schnittstellen fokussieren speziell die unidirektionale Verbindung von Produktionssystem zum digitalen Werkzeug. Forschungsschwerpunkte der Zukunft sind die autonome Steuerung der gesamten Produktion, die Anbindung über an Cloud-Services sowie die Integration umfangreicher Datenauswertung in menschenzentrierten Produktionssystemen aufbauend auf Gräßler & Pöhler [24].

Literaturverzeichnis

- [1] BMBF, "Bericht der Bundesregierung.: Zukunftsprojekte der Hightech-Strategie (HTS-Aktionsplan)," 2012.
- [2] I. Gräßler, A. Pöhler, and J. Pottebaum, "Creation of a Learning Factory for Cyber Physical Production Systems," *Procedia CIRP*, vol. 54, pp. 107–112, 2016.
- [3] E. M. Frazzon *et al.*, "Towards Socio-Cyber-Physical Systems in Production Networks," *Procedia CIRP*, vol. 7, pp. 49–54, 2013.
- [4] BMBF, Forschungsinfrastrukturen, Available: <https://www.horizont2020.de/einstieg-forschungsinfrastrukturen.htm>.
- [5] BMBF, Definition der Forschungsinfrastrukturen - EU-Büro des BMBF, Available: <https://www.euburo.de/infra-definition.htm>.
- [6] I. Gräßler and A. Pöhler, "Simulation von Cyber-Physischen Produktionssystemen in einer Forschungsinfrastruktur," 2019.
- [7] Gill, NSF Perspective and Status on Cyber-Physical Systems., Available: <http://varma.ece.cmu.edu/CPS/Presentations/gill.pdf>.
- [8] J. Fitzgerald, Ed., Collaborative design for embedded systems: Co-modelling and co-simulation, Berlin, Heidelberg, Springer, 2014.
- [9] E. A. Lee, "Cyber Physical Systems: Design Challenges," in 2008 11th IEEE International Symposium on Object and Component-Oriented Real-Time Distributed Computing (ISORC), Orlando, FL, USA, 2008, pp. 363–369.
- [10] J. Lee, B. Bagheri, and H.-A. Kao, "A Cyber-Physical Systems architecture for Industry 4.0-based manufacturing systems," *Manufacturing Letters*, vol. 3, pp. 18–23, 2015.
- [11] G. Reinhart, Ed., *Handbuch Industrie 4.0: Geschäftsmodelle, Prozesse, Technik*, München, Hanser, 2017.
- [12] I. Gräßler and A. Pöhler, "Intelligent Devices in a Decentralized Production System Concept," *Procedia CIRP*, vol. 67, pp. 116–121, 2018.
- [13] I. Gräßler, A. Pöhler, and J. Hentze, "Decoupling of Product and Production Development in Flexible Production Environments," *Procedia CIRP*, vol. 60, pp. 548–553, 2017.
- [14] I. Gräßler and A. Pöhler, "Produktentstehung im Zeitalter von Industrie 4.0," in *Handbuch Gestaltung digitaler und vernetzter Arbeitswelten*, G. W. Maier, G. Engels, and E. Steffen, Eds. Berlin: Springer, 2019, pp. 1–21.
- [15] ISO/IEC 7498-1: Information technology - Open Systems Interconnection - Basic Reference Model: The Basik Model, 7498-1, 1994.
- [16] A. S. Tanenbaum, *Computernetzwerke*, 4th ed., München, Pearson Studium, 2003.

- [17] M. Popp, *Das PROFINET IO-Buch: Grundlagen und Tipps für den erfolgreichen Einsatz*, 2nd ed., Berlin, Offenbach, VDE-Verl., 2010.
- [18] W. Mahnke, S.-H. Leitner, and M. Damm, *OPC unified architecture*, Berlin, Heidelberg, Springer, 2009.
- [19] Plant Simulation, Simulation, Available: <https://plant-simulation.de/simulation/>.
- [20] M. Di Gregorio *et al.*, "Designing usable interfaces for the Industry 4.0," in *Proceedings of the International Conference on Advanced Visual Interfaces*, Salerno Italy, 2020, pp. 1–9.
- [21] H. Orsolits and M. Lackner, Eds., *Virtual Reality und Augmented Reality in der digitalen Produktion*, Wiesbaden, Heidelberg, Springer Gabler, 2020.
- [22] I. Gräßler, D. Roesmann, and J. Pottebaum, "Traceable learning effects by use of digital adaptive assistance in production," *Procedia Manufacturing*, vol. 45, pp. 479–484, 2020.
- [23] I. Gräßler *et al.*, "Skill-based worker assignment in a manual assembly line," *Procedia CIRP*, vol. 100, pp. 433–438, 2021.
- [24] I. Graessler and A. Poehler, "Human-centric design of cyber-physical production systems," *Procedia CIRP*, vol. 84, pp. 251–256, 2019.

Sichere interoperable Maschine-zu-Maschine-Kommunikation basierend auf OPC UA Field Exchange

Dr.-Ing. **S. Höme**, **T. Fischer**, Dipl.-Inf. **G. Biehler**,
Dr.-Ing. **S. Kerschbaum**, Siemens AG, Nürnberg

1. Einleitung

In der OPC Foundation wurde in den letzten Jahren im Rahmen der Initiative „Field Level Communication (FLC)“ die Spezifikation für „OPC UA Field Exchange (FX)“ erarbeitet, welche das OPC UA Ökosystem auf die speziellen Anforderungen der (Echtzeit-) Kommunikation zwischen Automatisierungskomponenten erweitert (siehe Bild 1). Der Use-Case der ersten Version der OPC UA FX-Spezifikation ist die herstellerübergreifende Controller-zu-Controller (C2C) Kommunikation. OPC UA FX C2C ermöglicht es dabei zum ersten Mal in der Geschichte der Automatisierung, dass Controller unterschiedlicher Hersteller miteinander semantisch und in Echtzeit kommunizieren können. Die Daten können dabei auch sicherheitsgerichtete Informationen enthalten (Functional Safety). Informationssicherheit (Security) ist eine weitere grundlegende Eigenschaft von OPC UA und damit auch OPC UA FX.

Basierend auf der OPC UA FX Spezifikation¹ und den darin beschriebenen Techniken soll im Rahmen dieses Beitrags die Anwendererfahrung für die Projektierung dynamischer und flexibler Kommunikation zwischen autonomen Einheiten im Automatisierungskontext beschrieben werden. Dabei steht vor allem die Einfachheit der Projektierung solcher Kommunikation im Vordergrund. Eine autonome Einheit in diesem Beitrag kann eine Maschine², eine Process Equipment Assembly, eine Process Unit³ oder vergleichbares sein. Eine autonome Einheit führt einen oder mehrere Bearbeitungsschritte im Kontext der Fertigung oder Verarbeitung durch.

¹ Diese Arbeit basiert auf dem (unveröffentlichten) Release Candidate 2 (RC2) der OPC UA FX Spezifikation ([6], [3], [5], [6]). OPC UA FX ist ein Teil der OPC UA Basisspezifikation (Part 80-84).

² Der Begriff Maschine wird in der Regel im Kontext der Fabrikautomatisierung verwendet.

³ Die Begriffe Process Equipment Assembly oder Process Unit werden in der Regel im Kontext der Prozessautomatisierung verwendet.

wie möglich zu halten, soll an der Steuerungslogik und der Schnittstelle zu anderen Modulen und überlagerten Steuerungssystemen keine Änderung vorgenommen werden. Dieses Ziel wird oft auch als ein Projekt pro Modultyp bezeichnet.

Der Anlagenbetreiber möchte die verschiedenen autonomen Einheiten, welche in der Regel von unterschiedlichen OEMs geliefert werden, möglichst schnell und einfach verbinden und zu einer Produktion aufbauen. Dafür sind standardisierte Schnittstellen zwischen den autonomen Einheiten notwendig, welche auf dem gleichen unterlagerten Kommunikationssystem aufsetzen.

Zur Veranschaulichung soll im Folgenden eine Flaschenabfüllanlage genutzt werden (siehe Bild 2). Der Prozess ist dabei stark vereinfacht und abstrahiert, er dient nur dem besseren Verständnis der Ausführungen und hat keinen Anspruch auf technische Vollständigkeit oder Korrektheit⁵.

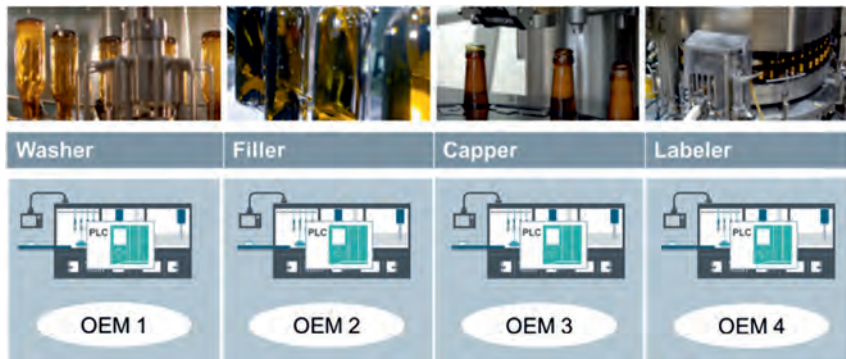


Bild 2: Vereinfachte Darstellung der modularen Flaschenabfüllanlage mit 4 autonomen Einheiten

Die Anlage besteht aus 4 verschiedenen autonomen Einheiten:

- Flaschen-Waschmaschine (Washer)
- Abfüller (Filler)
- Verschließer (Capper)
- Etikettiermaschine (Labeler)

Von jeder autonomen Einheit können im Sinne der Matrixproduktion mehrere Instanzen vorhanden sein.

⁵ Die Flaschenabfüllung wurde als Beispiel bei der OPC UA FX Demowand der OPC Foundation auf der SPS Messe 2021 und Hannovermesse 2022 ausgestellt.

3. Erstellung einer autonomen Einheit mit OPC UA FX

Eine Lösung der oben genannten Herausforderung besteht in der Nutzung von OPC UA FX für die Kommunikation zwischen den autonomen Einheiten (im allgemeinen auch Maschine-zu-Maschine-Kommunikation genannt).

Das Grundprinzip hier ist, dass jede autonome Einheit als OPC UA FX AutomationComponent modelliert wird und ein oder mehrere OPC UA FX FunctionalEntities besitzt, über welche sie mit anderen autonomen Einheiten (oder überlagerten Steuerfunktionen) kommunizieren kann. Ein wesentlicher Vorteil von OPC UA ist hier, dass sowohl die (typisierten) Datenpunkte als auch die Semantik (d.h. die Bedeutung) der Informationen beschrieben wird.

Neben den OPC UA FX FunctionalEntities als logische Schnittstellen definiert OPC UA FX ein Asset-Modell zur Beschreibung einer autonomen Einheit und ihrer Fähigkeiten. Zusätzlich sind auch steuerungstechnische und kommunikationstechnische Aspekte (z.B. Kommunikationsfähigkeiten im Bereich 5G, Ethernet TSN, o.ä.) modelliert.

OPC UA FX beschreibt die äußere Sicht auf eine autonome Einheit und ihre Schnittstellen. Der innere Aufbau ist gemäß dem Prinzip der Modularisierung versteckt. Dies bedeutet, dass eine autonome Einheit im inneren mit einer Speicherprogrammierbaren Steuerung (SPS) von Hersteller A, eine andere autonome Einheit mit einer SPS von Hersteller B und eine dritte autonome Einheit mit einem Industrie-PC gesteuert werden kann.

Für das gewählte Beispiel der Flaschenabfüllanlage baut ein Maschinenbauer beispielsweise einen Verschleißer (Capper). Er programmiert seine gewählte Steuerung mit dem zugehörigen Engineering Tool und erstellt dabei ein OPC UA FX Informationsmodell. In diesem sind neben den Konfigurations- und Assetinformationen auch zwei FunctionalEntities definiert:

- Eine FunctionalEntity für die Verbindung zur Vorgängereinheit und
- Eine FunctionalEntity für die Verbindung zur Nachfolgereinheit.

In diesen FunctionalEntities sind die Ein- und Ausgangsdaten beschrieben, welche zwischen den autonomen Einheiten ausgetauscht werden sollen. Die Ein- und Ausgangsdaten können dabei auch sicherheitsgerichtete (Safety) Datenpunkte enthalten. Weiterhin werden die Kommunikationsanforderungen wie der benötigte Quality of Service beschrieben.

Innerhalb der Steuerung der autonomen Einheit kann der Maschinenbauer diese Daten dann beispielsweise auf Datenbausteine seiner SPS abbilden und mit Funktionsbausteinen verarbeiten. Wie bereits beschrieben, ist die innere Verarbeitung der Datenpunkte von außen bewusst nicht sichtbar.

4. Orchestrierung von autonomen Einheiten mit OPC UA FX

Nachdem der Anlagenbetreiber verschiedene autonome Einheiten mit OPC UA FX gekauft hat, muss er diese in seiner Anlage installieren und zu einer Produktionslinie orchestrieren.

Für die Installation der autonomen Einheit erfolgt neben der mechanischen und elektrischen auch die netzwerktechnische Integration in die Kundenumgebung. Sie wird beispielsweise mit dem Netzwerk verbunden, Adressen werden vergeben und Zertifikate erstellt bzw. installiert. Dies soll hier nicht vertieft werden, dazu sei auf Aktivitäten wie das OPC UA Device Provisioning verwiesen (siehe [1]).

Die Orchestrierung kann direkt durch einen Anwender mit einem (graphischen) Tool erfolgen. Alternativ kann diese Orchestrierung durch ein überlagertes Produktionsplanungssystem erfolgen, welches basierend auf einer Planung autonome Einheiten dynamisch zu einer logischen Produktionslinie zusammenfasst. Die Schnittstelle zwischen den autonomen Einheiten und der überlagerten Orchestrierung ist mit OPC UA FX standardisiert, so dass beide Wege äquivalent und auch parallel einsetzbar sind.

Als Orchestrierungstool soll im Folgenden beispielhaft der sogenannte Connection Editor (siehe Bild 3) genutzt werden.

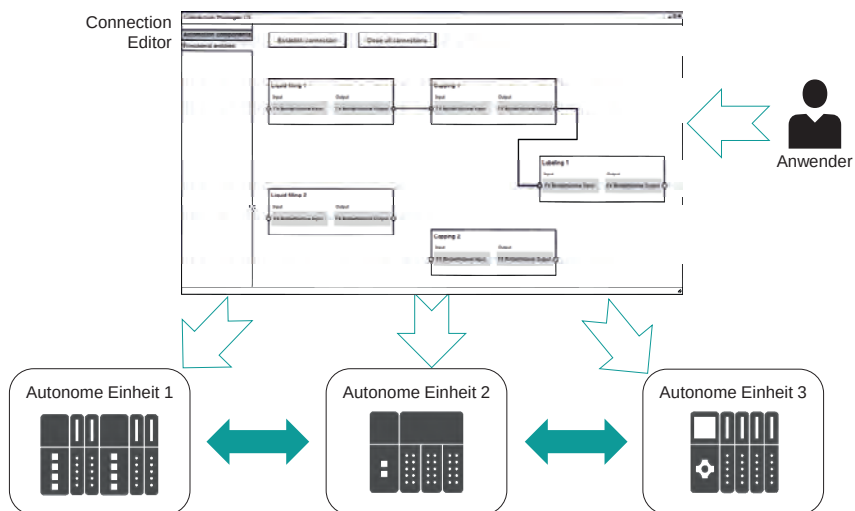


Bild 3: Orchestrierung von OPC UA FX basierten autonome Einheiten mit einem Connection Editor

Im ersten Schritt lädt der Anwender die einzelnen AutomationComponents (die FX-Repräsentation der autonomen Einheiten) entweder aus einem Offline-Descriptor⁶ oder durch Online-Browsing aus den autonomen Einheiten. Die einzelnen AutomationComponents werden eine lokale Bibliothek geladen und können dann in einem Projekt genutzt werden.

Dann verbindet der Anwender FunctionalEntities der einzelnen AutomationComponents (autonome Einheiten) durch Linien im Connection Editor. Diese Linien repräsentieren eine OPC UA FX Connection. An dieser Linie können die Eigenschaften der Verbindung (beispielsweise der verwendete Quality of Service oder Übertragungsintervall) konfiguriert werden. Wenn die autonomen Einheiten funktionale Sicherheit (Safety) für die Kommunikation erfordern, wird diese automatisch aktiviert.

Auf Knopfdruck lädt der Anwender die Orchestrierungsinformationen in die autonomen Einheiten. Nach dem Herunterladen der Konfiguration etablieren die autonomen Einheiten die Kommunikation basierend auf OPC UA PubSub, der Datenaustausch beginnt und die Produktion der Linie kann starten. Die autonomen Einheiten verbinden sich automatisch neu, beispielsweise nach dem Abschalten über Nacht. Dafür ist keine Anwenderaktion notwendig.

5. OPC UA FX C2C Multi-Vendor-Demo

Im Rahmen der Standardisierungsaktivitäten der OPC Foundation wurde eine Multi-Vendor-Demo erstellt, in der Steuerungen und Netzwerkinfrastruktur-Komponenten verschiedener Hersteller per OPC UA FX kommunizieren und den hier beschriebenen Show-Case der modularen Flaschenabfüllanlage demonstrieren (siehe Bild 4).

Jede der Steuerung implementiert eine autonome Einheit der Flaschenabfüllanlage, so dass es von jedem Typ (Washer, Filler, Capper, Labeler) mehrere Instanzen gibt. Diese können dann dynamisch verbunden werden, um gemeinsam ein Produkt zu erzeugen. Dies stellt eine einfache Variante der Matrixproduktion dar, bei der jede autonome Einheit durch die standardisierte OPC UA FX Schnittstelle mit einer beliebigen Steuerung ausgestattet sein kann.

⁶ Der Offline-Descriptor ist Teil der OPC UA FX Spezifikation und beschreibt eine AutomationComponent, deren FunctionalEntities und Capabilities in einem herstellerunabhängigen Format.

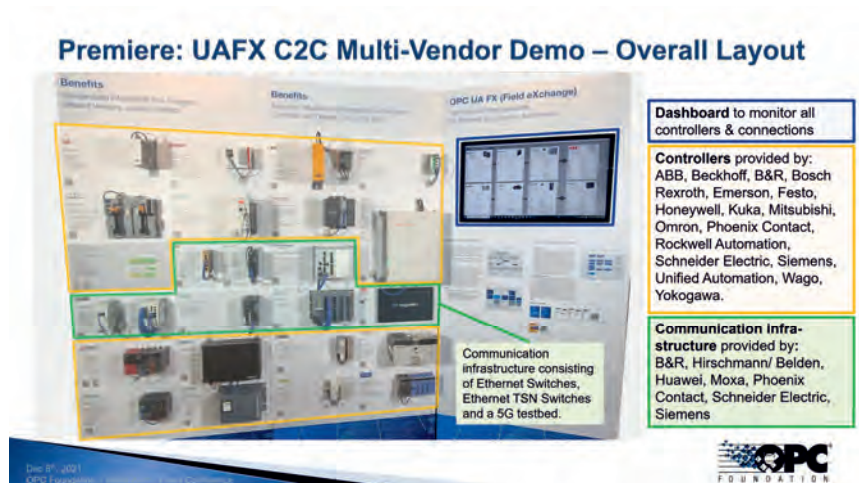


Bild 4: OPC UA FX C2C Demowand (Quelle: OPC Foundation)

6. Technische Details zu OPC UA FX

OPC UA FX setzt unterlagert auf die etablierten Kommunikationsmuster Client/Server und PubSub von OPC UA. Dadurch kann OPC UA FX die für die Automatisierungstechnik wichtigen Aspekte abdecken (siehe Bild 5):

- Safety: OPC UA unterstützt Kommunikation mit funktionaler Sicherheit sowohl für Client/Server als auch für PubSub (siehe [2])
- Security: (Informations-)Sicherheit ist ein integraler Bestandteil. So unterstützt beispielsweise OPC UA PubSub Verschlüsselung und Signatur der übertragenen Daten, aber auch die Verteilung von Zertifikaten und Schlüsselmateriale ist standardisiert.
- Quality of Service: OPC UA beschreibt im Part 14 der Spezifikation die Abbildung von PubSub auf verschiedene unterlagerte Kommunikationssysteme, welche den für die Automatisierungstechnik notwendigen Quality of Service sicherstellen können (beispielsweise Ethernet TSN oder 5G).

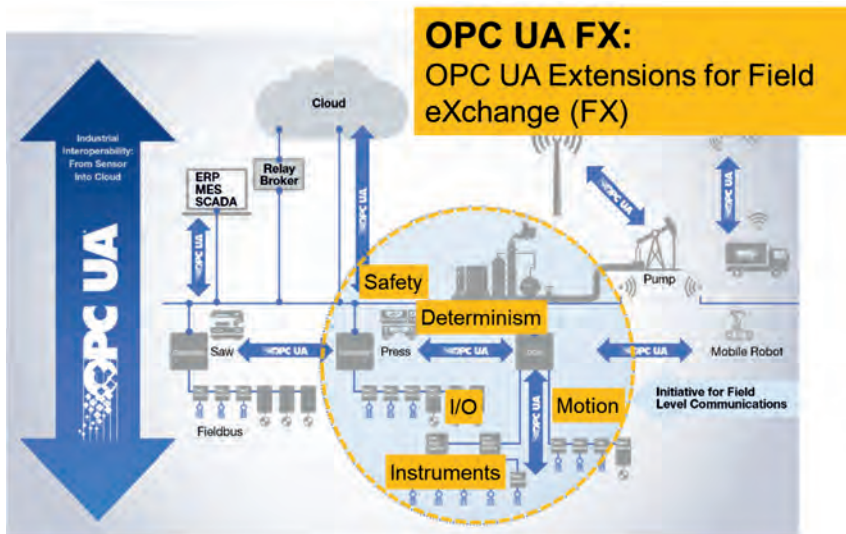


Bild 5: OPC UA Erweiterungen für die Feldebene (Quelle: OPC Foundation)

In Bild 6 ist beispielhaft ein Auszug aus dem OPC UA Informationsmodell einer solchen Steuerung dargestellt. Unterhalb des Knotens FxRoot befindet sich die AutomationComponent (BottleController), welche wiederum über Assets und Functional Entities verfügt. Das hier angelegt Asset (Simatic) wird über den Hersteller (Manufacturer), Modell (Model), Seriennummer (SerialNumber) und andere Eigenschaften beschrieben.

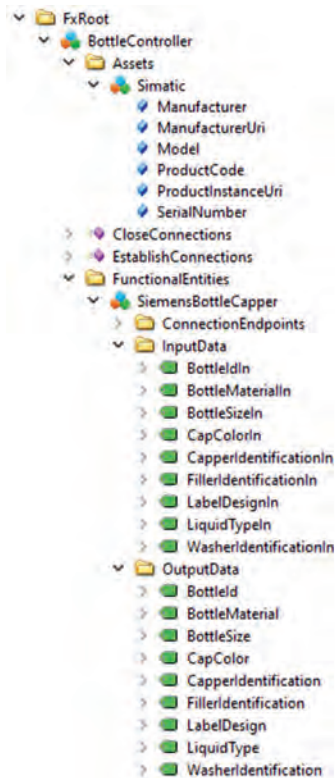


Bild 6: Auszug des OPC UA FX Informationsmodell einer Demo-Flaschenverschlussmaschine

In Bild 6 ist weiterhin eine FunctionalEntity (SiemensBottleCapper) dargestellt, welche aus Eingangsdaten (InputData) und Ausgangsdaten (OutputData) besteht. In den Verbindungspunkten (ConnectionEndpoints) sind die aktuellen Verbindungen zu anderen FunctionalEntities dargestellt.

Weiterhin enthält die AutomationComponent BottleController zwei Methoden zum Auf- und Abbau von Verbindungen: EstablishConnections und CloseConnections.

Diese Struktur des Informationsmodells ist standardisiert und im Part 81 der OPC UA FX Spezifikation [3] beschrieben.

7. Zusammenfassung

OPC UA Field Exchange (OPC UA FX) bringt ein konsistentes Informationsmodell vom Feld bis in die Cloud, welches die Durchgängigkeit der Informationen für verschiedenste Anwendungsfälle (Control, Analytics, ...) sicherstellt. Insbesondere bringt OPC UA FX hier ein harmonisiertes Informationsmodell zur Beschreibung von Assetinformationen und funktionalen Schnittstellen einer autonomen Einheit.

OPC UA FX entspricht den Anforderungen der Automatisierungstechnik an Maschine-zu-Maschine-Kommunikation durch ein konvergentes Konzept von funktionaler Sicherheit (Safety), Informationssicherheit (Security) und Echtzeit (Quality of Service).

Durch die in OPC UA existierenden Companion Specifications, welche wiederum harmonisierte Informationsmodelle für bestimmte Branchen oder Applikationen darstellen, ist so die einfache Kommunikation zwischen verschiedenen autonomen Einheiten und mit überlagerten Systemen in einer Anlage sichergestellt.

8. Literaturverzeichnis

- [1] F. Volkmann, N. Baur, S. Höme, A. Palmin und M. Madsen, „Vertrauensbasierte Integration von Geräten mittels OPC UA Device Provisioning,“ in *Automation 2022*, Baden-Baden, 2022.
- [2] OPC Foundation, „OPC 10000-15 - UA Specification Part 15 - Safety 1.05.00,“ 2021.
- [3] OPC Foundation, „OPC 10000-81 - UA Specification Part 81 - UAFX Connecting Devices and Information Model 1.00.00 RC2,“ 2022.
- [4] OPC Foundation, „OPC 10000-82 - UA Specification Part 82 - UAFX Networking 1.00.00 RC2,“ 2022.
- [5] OPC Foundation, „OPC 10000-83 – UA Specification Part 83 - UAFX OfflineEngineering 1.00.00 RC2,“ 2022.
- [6] OPC Foundation, „OPC 10000-80 - UA Specification Part 80 - UAFX Overview and Concepts 1.00.00 RC2,“ 2022.

Stärkung der Cyber-Resilienz in KMU

Strengthening cyber resilience in SME

Christian Siegwart M.Sc., **Felix Scherhag** M.Sc.,

ZeMA gGmbH, Saarbrücken;

Kai Pervözl, Fraunhofer IAIS, Sankt Augustin,

Univ.-Prof. Dr.-Ing. **Georg Frey**,

Universität des Saarlandes, Saarbrücken

Kurzfassung

Der zunehmende Einsatz digitaler Technologien stellt viele Unternehmen vor dem Hintergrund einer angespannten IT-Sicherheitslage vor große Herausforderungen. Das Bewusstsein für Anforderungen und Maßnahmen im Kontext Informationssicherheit wächst zwar kontinuierlich, es besteht jedoch weiter Handlungsbedarf bei den Unternehmen, insbesondere bei kleinen und mittleren Unternehmen (KMU). In diesem Beitrag werden Ergebnisse des Forschungs- und Transferprojektes Grundschutz^{PLUS} Aktivator vorgestellt. Ziel des Projektes ist es, auf Grundlage eines ontologiebasierten Security Management Prozesses die Cyber-Resilienz produktionsnaher KMU zu stärken.

Stichwörter: Cyber-Resilienz , IT-Security, Security-Management, IT-Sicherheitsmodell

Abstract

The increasing use of digital technologies creates major challenges for many companies against the backdrop of a tense IT security situation. Although awareness of requirements and measures in the context of information security is growing continuously, further improvement is still required for companies, especially for small and medium-sized enterprises (SMEs). This paper presents results of the Grundschutz^{PLUS} Aktivator research and transfer project. The aim of the project is to strengthen the cyber resilience of production-related SMEs based on an ontology-based security management process.

1. Einführung und Motivation

In der jüngsten Vergangenheit erleben wir einen deutlichen Anstieg der Bedrohungen durch Cyberkriminalität. Komplexe Angriffe entwickeln sich für Unternehmen zu einem immer größeren Risikofaktor. Nach [1] werden die Risiken durch Cyberkriminalität weltweit als eines

der größten Unternehmensrisiken gelistet. Das Bundesamt für Sicherheit in der Informationstechnik (BSI) beschreibt in [2] die aktuelle Lage als angespannt bis kritisch. Neben der großen Gefahr durch zunehmende Angriffe stellt die Digitalisierung die Unternehmen bereits heute und auch in Zukunft vor große Herausforderungen. Betroffen sind dabei alle Unternehmen, jedoch haben kleine und mittlere Unternehmen (KMU) besonders mit den Herausforderungen im Kontext Informationssicherheit zu kämpfen. In [3] und [4] sind die Probleme vieler KMU für die Umsetzung der notwendigen Transformationsprozesse und auch der Maßnahmen zur Erhöhung der Informationssicherheit dargestellt. Insbesondere produzierende KMU haben demnach mit den Herausforderungen zu kämpfen. Unter anderem gibt es die folgenden Herausforderungen auf KMU Seite:

- Fehlendes Bewusstsein
- Fehlende Kompetenzen zum Aufbau von Know-how
- Zeitmangel im Tagesbetrieb und zu hohe Kosten
- Überforderung durch die Komplexität der Informationen
- Mangelnde Zusammenarbeit von Geschäftsführung und IT

Ein weiteres Merkmal für die unzureichende Widerstandsfähigkeit im Kontext Cyberkriminalität zeigt die Tatsache, dass nur jedes zweite Unternehmen über ein Notfallmanagement verfügt. Auch hier bilden die kleineren Unternehmen das Schlusslicht. [5]

Unsere Erfahrungen in Gesprächen mit KMU unterstreichen die Ergebnisse aus den Studien und zeigen deutlich den Unterstützungsbedarf für KMU insbesondere zur Stärkung der Informationssicherheitsmaßnahmen.

Das vom BMWK geförderte Forschungs- und Transferprojekt Grundschutz^{PLUS} Aktivator [6] hat sich genau dieses Ziel gesetzt und möchte KMU dabei unterstützen das IT-Sicherheitsniveau zu erhöhen. Dabei sind sowohl präventive als auch reaktive Maßnahmen im Fokus des Projektes, um das Unternehmen ganzheitlich zu stärken und die Widerstandsfähigkeit produktionsnaher KMU zu erhöhen.

2. Grundlagen

Dieser Abschnitt zeigt zunächst die verwendeten Grundlagen, die für diesen Beitrag relevant sind. Dies umfasst zunächst die Einordnung des Begriffes „Cyber-Resilienz“. Anschließend wird das Projekt Grundschutz^{PLUS} Aktivator und die verwendete Wissensbasis kurz vorgestellt.

2.1 Cyber-Resilienz

Der Begriff Resilienz wird bereits in unterschiedlichen Forschungsbereichen, wie beispielsweise der Soziologie oder auch Psychologie verwendet. In der Psychologie kann der Begriff Resilienz folgendermaßen definiert werden:

“Resilience refers to a class of phenomena characterized by good outcomes in spite of serious threats to adaptation or development.” [7]

Im Zentrum dieser Definition steht eine positive Entwicklung unter dem Einfluss ernsthafter Gefährdungen. Diese Beschreibung lässt sich sehr gut auf die aktuelle Situation der Unternehmen im Kontext der Digitalisierung und den Gefährdungen durch Cyberkriminalität übertragen. In diesem Kontext existieren ebenfalls Forschungen die den Begriff Cyber-Resilienz wie folgt definieren:

“Cyber resilience refers to the ability to continuously deliver the intended outcome despite adverse cyber events.” [8]

Diese Definition beschreibt den Begriff Cyber-Resilienz durch die Fähigkeit eines Unternehmens, die geplanten Ergebnisse unter dem Einfluss unerwünschter Cyberangriffe zu erzielen.

Die dargestellten Definitionen zeigen deutlich die Herausforderung für alle Unternehmen, aber insbesondere für KMU. Auf der einen Seite gilt es, die eigenen Geschäftsmodelle anzupassen und die technologischen Potenziale der Digitalisierung zu nutzen. Auf der anderen Seite müssen im besten Fall Schäden durch Cyberangriffe verhindert werden und es sind Vorbereitungen notwendig, um im Ernstfall geeignet reagieren zu können.

Mit dem Hintergrund dieser Definitionen können wir die Cyber-Resilienz der KMU stärken, indem die Unternehmen ganzheitlich im Kontext Informationssicherheit unterstützt werden. In [9] wurde bereits gezeigt, wie ein ganzheitliches Reifegradmodell die Bewertung des Reifegrades im Kontext Security unterstützen kann. Die ganzheitliche Betrachtung fokussiert dazu im Kontext Informationssicherheit nicht nur den Bereich Technologie, sondern ebenfalls die Bereiche Organisation, Prozesse und den Faktor Mensch. Ziel des Forschungs- und Transferprojektes Grundschutz^{PLUS} Aktivator ist es, KMU ganzheitlich mit auf das KMU zugeschnittenen IT-Security-Informationen zu unterstützen.

2.2 Kurzvorstellung Grundschutz^{PLUS} Aktivator

Das Projekt Grundschutz^{PLUS} Aktivator ist Teil der Initiative „IT-Sicherheit in der Wirtschaft“ im Förderschwerpunkt Mittelstand-Digital und wird in Kooperation zwischen dem Fraunhofer Institut für intelligente Analyse- und Informationssysteme (IAIS) und dem Zentrum für Mechatronik und Automatisierungstechnik (ZeMA) durchgeführt.

Mit dem Grundschutz^{PLUS} Aktivator wird eine interaktive Plattform geschaffen, die KMU in die Lage versetzt, kompetente IT-sicherheitsrelevante Entscheidungen selbst zu treffen. Zudem wird das Verständnis für Sicherheitsfragestellungen bei den Anwendern/-innen gestärkt. KMU erhalten speziell auf ihr Unternehmen zugeschnittene IT-Sicherheitsinformationen und Hilfestellungen bei der Umsetzung.

Grundlage ist die Erfassung der spezifischen IT-Sicherheitsanforderungen des KMU. Hierzu wird eine interaktive, dialogbasierte und in der Sprache der KMU formulierte Erhebung der Anforderungen realisiert. Die Plattform führt durch einen dynamischen Dialog bis alle relevanten Informationen erfasst sind und überführt diese in ein vereinfachtes Modell des Unternehmens. Auf Basis existierender Standards wird ein IT-Sicherheitsmodell entwickelt und hierzu passende Resilienz- und Schutzmaßnahmen identifiziert. Auf dieser Grundlage entwickelt der Grundschutz^{PLUS} Aktivator ein auf das Unternehmen zugeschnittenes IT-Sicherheitskonzept. Die Umsetzung wird durch Beispielrealisierungen, Schritt für Schritt-Anleitungen und Templates unterstützt.

Dies führt dazu, dass die Unternehmen sicherer bei der Planung und Umsetzung digitalisierter Prozesse und Geschäftsmodelle werden und die Wettbewerbs- und Innovationsfähigkeit der KMU gestärkt wird.

Diese Vorgehensweise im Projekt (siehe Abbildung 1) wird unterstützt durch KMU-Feedback, das über die Veranstaltung von Workshops bereits während der Projektlaufzeit in die weitere Entwicklung der Plattform einfließt.



Bild 1: Projektablauf des Forschungs- und Transferprojektes Grundschutz^{PLUS} Aktivator

2.3 Wissensbasis

Um KMU mit den relevanten IT-Security-Informationen zu unterstützen ist ein zentraler Bestandteil des Projektes für das jeweilige KMU relevante Informationen bereitzustellen.

Für den Kontext Informationssicherheit produktionsnaher KMU stehen im Projekt die Standards BSI IT-Grundschutz und die internationale Normenreihe IEC 62443 für den Produktionsbereich im Vordergrund.

BSI IT-Grundschutz

Der IT-Grundschutz des Bundesamtes für Sicherheit in der Informationstechnik (BSI) hat das Ziel, mit einem standardisierten, ganzheitlichen Ansatz ein angemessenes Maß an Informationssicherheit in Unternehmen zu gewährleisten. Er besteht zum einen aus den BSI-Standards und zum anderen aus dem IT-Grundschutz-Kompendium ([10]). Die Standards liefern Informationen zum methodischen Vorgehen. Das IT-Grundschutz-Kompendium stellt als modular aufgebautes Nachschlagewerk das zentrale Werkzeug des IT-Grundschutzes dar. Derzeit (Stand: Februar 2022) beinhaltet das Kompendium 104 IT-Grundschutz-Bausteine, die in zehn Bereiche aufgeteilt sind und sich mit unterschiedlichsten Aspekten der Informationssicherheit beschäftigen.

Jeder Baustein enthält jeweils eine Beschreibung der betrachteten Thematik, Vorgehensweisen und IT-Systeme, gibt einen Überblick über spezifische Gefährdungen und beschreibt detailliert standardisierte Sicherheitsanforderungen, die es zu beachten gilt.

Einen Sicherheitsprozess nach IT-Grundschutz im Unternehmen zu etablieren ist vollständig kompatibel zum Standard IEC/ISO 27001 [11]. Daher dient das IT-Grundschutz-Kompodium als Prüfgrundlage für Zertifizierungen nach ISO/IEC 27001 auf Basis von IT-Grundschutz.

IEC 62443

Die Normenreihe IEC 62443 „Industrial communication networks – Network and system security“ umfasst 14 Teile, die in die vier Bereiche General, Policies & Procedures, System und Component/Product unterteilt sind. Die IEC 62443 hat sich als Branchenstandard im Bereich der industriellen Automatisierung etabliert und wird in Deutschland durch das Spiegelgremium DKE UK 931.1 „IT-Sicherheit in der Automatisierungstechnik“ betreut. Die Norm hat die klassischen Security Schutzziele Vertraulichkeit, Integrität sowie insbesondere die Verfügbarkeit von Industrial Automation and Control Systems, kurz IACS, im Fokus. Darüber hinaus stellt die Einhaltung von Anforderungen der funktionalen Sicherheit (functional safety) ein weiteres Schutzziel der IEC 62443 dar. Wichtig ist dabei, dass diese Ziele nur mit einem ganzheitlichen Security-Ansatz, der sowohl Hersteller als auch Integrator und Betreiber umfasst, erreicht werden können. Besonders relevant für die Wissensbasis und damit auch für diesen Beitrag sind die Anforderungen aus Integrator- und Betreibersicht an die Automatisierungssysteme aus dem Normenteil 3-3 ([12]).

Die Informationen aus den dargestellten Standards direkt zu verwenden ist nur eingeschränkt möglich, da die Komplexität zu groß ist und nicht die Bedürfnisse der kleinen und mittleren Unternehmen trifft. Daher wird im Rahmen des Projektes auf ein ontologiebasiertes Security Management zurückgegriffen. Die in [13] dargestellte Entwicklung einer Wissensbasis ermöglicht für das Projekt Grundschutz^{PLUS} Aktivator eine wissensbasierte digitale Plattform zu erstellen, die alle notwendige Informationen von der Unternehmensmodellierung bis zu den IT-Security-Informationen und deren Zusammenhänge beinhaltet.

3. Nutzung der Plattform Grundschutz^{PLUS} Aktivator

Ziel des Projektes Grundschutz^{PLUS} Aktivator ist es, durch die Entwicklung einer Plattform für KMU die Cyber-Resilienz der Unternehmen zu stärken. Die Herausforderung liegt insbeson-

dere darin, die relevanten Informationen herauszufiltern und für die Zielgruppe möglichst einfach darzustellen.

Für eine möglichst einfache Nutzbarkeit ist die Nutzung der Plattform auf drei wesentliche Schritte beschränkt (siehe Abbildung 2), die im Folgenden näher beschrieben werden.

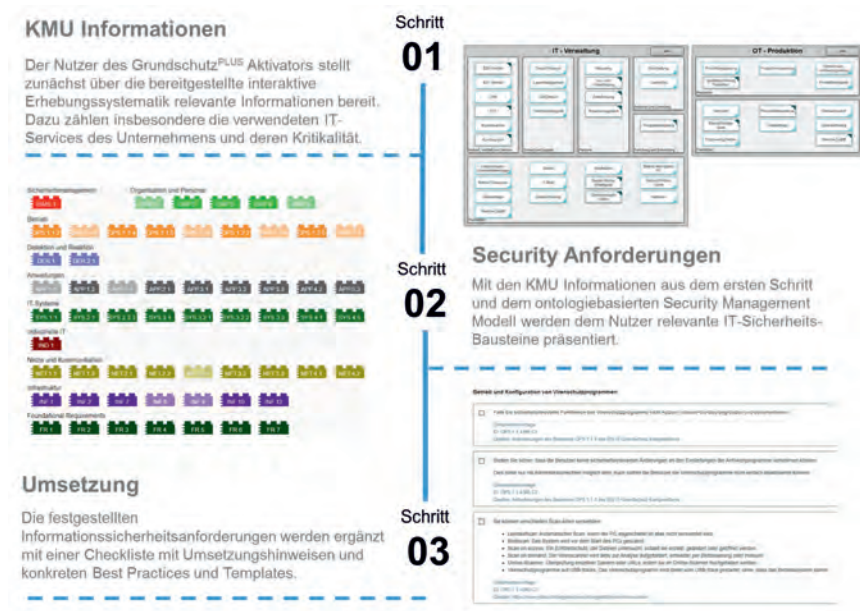


Bild 2: Vorgehensweise zur Nutzung des Grundschutz^{PLUS} Aktivators

Schritt 1: KMU Informationen

Um auf das KMU zugeschnittene Informationssicherheitsanforderungen zu bestimmen, ist im ersten Schritt die Abfrage wichtiger Unternehmensinformationen notwendig. Dazu wird im Projekt eine interaktive Erhebungssystematik erstellt, die es ermöglicht, ohne tiefere Expertise die relevanten Informationen zu hinterlegen. Dazu zählen insbesondere:

- Allgemeine Unternehmensinformationen (z.B. Branche, Anzahl Mitarbeiter)
- Auswahl der im KMU eingesetzten IT-basierten Services
- Auswahl der extern bereitgestellten Services (Outsourcing)
- Einschätzung der Kritikalität der ausgewählten Services

Schritt 2: Security Anforderungen

Die vom KMU bereitgestellten Informationen erlauben mit der im Projekt realisierten Wissensbasis die Auswahl der speziell für das jeweilige KMU relevanten Informationssicherheitsanforderungen. Wie bereits zuvor dargestellt, enthält die Wissensbasis Informationen für den Kontext IT-Security Management für KMU, die unter anderem relevante Inhalte aus den anerkannten Standards BSI IT-Grundschutz oder auch der Normenreihe IEC 62443 beinhalten.

Diese herausgefilterten Informationen werden den Nutzern/-innen zusammengefasst in Bausteinen mit kurzen Erklärungen präsentiert, sodass die Nutzer/-in sich in diesem Schritt einen Überblick über die relevanten Anforderungen machen können.

Schritt 3: Umsetzung

Das Herzstück zur Stärkung der Cyber-Resilienz produktionsnaher KMU stellt nach der Darstellung relevanter Anforderungen die Bereitstellung konkreter Umsetzungsempfehlungen dar. Diese werden den Nutzern/-innen in Form einer Checkliste präsentiert. Zu jedem Anforderungspunkt bekommen die Nutzer/-innen somit Erklärungen zu Umsetzungsmöglichkeiten, die, wenn es möglich ist um weitere Umsetzungsdokumente ergänzt werden. Umsetzungsdokumente, wie beispielsweise das Template einer Informationssicherheitsleitlinie, werden zum Download angeboten und bieten eine einfache Möglichkeit die Umsetzung der dargestellten Maßnahmen zu unterstützen.

Da je nach angegebenen Unternehmensinformationen viele Umsetzungsempfehlungen relevant sind, ist in diesem Schritt das Sortieren, Filtern oder auch Umgruppieren der Checkpunkte möglich. Damit hat der/die Nutzer/-in beispielsweise die Möglichkeit, sich nur die Checkpunkte anzuzeigen, die seiner Rolle im Unternehmen (z.B. Geschäftsführer oder IT-Administrator) zugeordnet werden. Des Weiteren ist eine Filterung der Checkpunkte nach bestimmten Bedrohungen oder auch nach der Kritikalität möglich.

4. Fazit und Ausblick

Insbesondere kleine und mittlere Unternehmen sehen sich in einem Spannungsfeld zwischen notwendigen Veränderungen der Geschäftsprozesse durch die Digitalisierung und den steigenden Risiken durch Cyberkriminalität. Dies führt dazu, dass es insbesondere bei produktionsnahen KMU dringenden Handlungsbedarf zur Stärkung der Cyber-Resilienz gibt.

Das Projekt Grundschutz^{PLUS} Aktivator hat zum Ziel, das IT-Sicherheitsniveau produzierender KMU ganzheitlich zu erhöhen und Hürden in Bezug auf die Umsetzung von IT-Sicherheitsmaßnahmen abzubauen. Dazu wird eine interaktive Plattform geschaffen, die

KMU in die Lage versetzt, kompetente IT-sicherheitsrelevante Entscheidungen selbst zu treffen und die das Verständnis für Sicherheitsfragestellungen stärkt. KMU erhalten speziell auf ihr Unternehmen zugeschnittene IT-Sicherheitskonzepte und Hilfestellung bei der Umsetzung. Die Informationen betrachten das KMU ganzheitlich, sodass es sowohl technologisch und organisatorisch als auch präventiv und reaktiv gestärkt wird.

Bei der weiteren Bearbeitung des Projektes steht die Funktionserweiterung des User Interfaces und die stetige Aktualisierung und Ergänzung der Umsetzungshinweise für die KMU im Vordergrund, da konkrete Hilfestellungen zur Umsetzung enorm wichtig für die Unterstützung der KMU sind.

Das Projekt Grundschutz^{PLUS} Aktivator ist Teil der Initiative „IT-Sicherheit in der Wirtschaft“ im Förderschwerpunkt Mittelstand-Digital. Weitere Informationen zum Projekt sind unter www.security4kmu.de zu finden. Das Mittelstand-Digital Netzwerk bietet mit den Mittelstand 4.0-Kompetenzzentren, der Initiative „IT-Sicherheit in der Wirtschaft“ und „Digital Jetzt“ umfassende Unterstützung bei der Digitalisierung.



Kleine und mittlere Unternehmen profitieren von konkreten Praxisbeispielen und passgenauen, anbieterneutralen Angeboten zur Qualifikation und IT-Sicherheit. Das Bundesministerium für Wirtschaft und Klimaschutz ermöglicht die kostenfreie Nutzung und stellt finanzielle Zuschüsse bereit. Weitere Informationen finden Sie unter www.mittelstand-digital.de und www.it-sicherheit-in-der-wirtschaft.de.

Literaturangaben

- [1] Allianz Global Corporate & Specialty SE: Allianz Risk Barometer. Identifying the major business risks for 2020, 2020. <https://agcs.wufoo.com/forms/download-the-allianz-risk-barometer-2020-report/>, abgerufen am: 29.04.2020
- [2] BSI - Bundesamt für Sicherheit in der Informationstechnik: Die Lage der IT-Sicherheit in Deutschland 2021, 2021. https://www.bsi.bund.de/DE/Service-Navi/Publikationen/Lagebericht/lagebericht_node.html, abgerufen am: 02.03.2022
- [3] Wissenschaftliches Institut für Infrastruktur und Kommunikationsdienste - WIK: Aktuelle Lage der IT-Sicherheit in KMU. Langfassung, 2017. https://www.wik.org/fileadmin/Sonstige_Dateien/IT-Sicherheit_in_KMU/WIK-Studie_Aktuelle_Lage_der_IT-Sicherheit_in_KMU_Langfassung__2_.pdf, abgerufen am: 22.04.2020
- [4] Czichos, R. u. Neuburger, R.: Digitaler Wandel: Die größten Stolpersteine für KMU. Wissensmanagement 3 (2021) 3, S. 28–31
- [5] bitkom e.V.: Sicherheitsmaßnahmen 2021. Wie sich Deutschlands Unternehmen gegen Diebstahl, Spionage und Sabotage schützen, 2021. <https://www.bitkom.org/Bitkom/Publikationen/Sicherheitsmassnahmen-2021-Chart-Bericht-zur-IT-Sicherheit-in-der-deutschen-Wirtschaft>, abgerufen am: 30.04.2022
- [6] Zentrum für Mechatronik und Automatisierungstechnik gGmbH, Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme: GrundschatzPLUS Aktivator. <https://www.security4kmu.de>, abgerufen am: 11.12.2020
- [7] Masten, A. S.: Ordinary magic. Resilience processes in development. The American psychologist 56 (2001) 3, S. 227–238
- [8] Björck, F., Henkel, M., Stirna, J. u. Zdravkovic, J.: Cyber Resilience – Fundamentals for a Definition. In: Rocha, A. (Hrsg.): New Contributions in Information Systems and Technologies. Volume 1. Advances in Intelligent Systems and Computing, Bd. 353. Cham, s.l.: Springer International Publishing 2015, S. 311–316
- [9] Frey, G., Siegwart, C., Scherhag, F. u. Krammel, M.: Security Awareness für den Mittelstand. Wie ein Reifegradmodell die Security Awareness kleiner und mittlerer Unternehmen erhöhen kann. atp magazin 63 (2021) 6-7, S. 78–85
- [10] IT-Grundschatz-Kompodium Edition 2022; Februar 2022. https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Standards-und-Zertifizierung/IT-Grundschatz/IT-Grundschatz-Kompodium/it-grundschatz-kompodium_node.html, abgerufen am: 29.04.2022

- [11] DIN EN ISO/IEC 27001. Information technology - Security techniques - Information security management systems - Requirements; Juni 2017
- [12] IEC 62443-3-3. Industrial communication networks - Network and system security - Part 3-3: System security requirements and security levels
- [13] Siegwart, C., Scherhag, F., Pervölz, K. u. Frey, G.: Ontologiebasiertes Security Management. In: Proceedings of the Kongress Automation 2020, S. 365–376

Prognostizierte Energieverbräuche durch Deep Learning als Basis zur Integration in ein Smart Grid am Beispiel des Flughafens Köln Bonn

Mathis Stange, Maximilian Both, Prof. Dr. Jochen Müller,
Institut für Technische Gebäudeausrüstung,
Technische Hochschule Köln

Kurzfassung

Die zunehmende Energieverfügbarkeit durch erneuerbare Energien erfordert effiziente und intelligente Energieversorgungsnetze. In einem Smart Grid sollen alle Akteure von der Erzeugung bis zum Verbraucher miteinander interagieren. In diesem Beitrag wird ein Vorhersagemodell für den Flughafen Köln-Bonn erstellt. Grundlage dafür sind Verbrauchsdaten der TGA-Anlagen, Erzeugerdaten (BHKW, PV und Notstromaggregate) sowie der zugekaufte Strom aus dem vorgelagerten Netz. In Kombination mit Wetterdaten und flughafenspezifischen Informationen über Passagier- und Frachtverkehr bilden diese Daten die Grundlage des Algorithmus. Dieser wurde in Form eines künstlichen neuronalen Netzes (KNN) implementiert. KNN haben im Vergleich zu Simulationen den Vorteil, Prognosen nicht auf Grundlage von Annahmen über Baueigenschaften und technischen Anlagen zu treffen. KNN tätigen die Vorhersagen aufgrund von historischen Daten. Das entwickelte KNN prognostiziert den Strom- und Erdgasverbrauch am Flughafen Köln Bonn für einen kurz-, mittel-, und langfristigen Prognosehorizont. Der mittlere absolute Fehler liegt je nach Prognosehorizont zwischen drei und zehn Prozent. Eine Vorhersage künftiger Verbräuche des Flughafens, in Kombination mit den prognostizierten Erzeugnissen, ermöglicht die Integration des Flughafens in ein Smart Grid. Die Energieerzeuger wissen, in welchen Zeiträumen Residuallasten am Flughafen entstehen und können diese durch individuelle Stromtarife oder durch intelligente Steuerung reduzieren.

1. Einleitung

Damit Treibhausgasemissionen eingespart werden können und gleichzeitig eine wachsende weltweite Nachfrage von Energie gedeckt werden kann, ist eine feingranulare Abstimmung von Energieerzeugung, -verbrauch und -speicherung in Form einer intelligenten und dezentralen Netzsteuerung notwendig [1]. Von besonderer Bedeutung für eine Dekarbonisierung und Optimierung der Netzsteuerung sind Forecasting Modelle (FM), die als Informationsgrundlage

dienen. FM, die mithilfe von Deep Learning (DL)-Methodiken entwickelt werden, haben den Vorteil, dass sie im Vergleich zu Simulationen keine Energielastprognosen (ELP) anhand von Vereinfachungen und Annahmen tätigen. Stattdessen prognostizieren sie den Bedarf auf Basis historischer Verbrauchsdaten, die im Betrieb aufgenommen wurden.

In dieser Arbeit wird ein KNN zur ELP für den Flughafen Köln Bonn vorgestellt. Die Prognose wird unter Einbeziehung exogener Parameter realisiert. Für das Energienetz des Flughafens liegen die Verbrauchsdaten der TGA-Anlagen, die Erzeugerdaten (BHKW, PV-Anlagen und Notstromaggregate) sowie der zugekaufte Strom aus dem vorgelagerten Netz, viertelstündlich, für insgesamt dreieinhalb Jahre vor. In Kombination mit Informationen zu Passagieren und Fracht sowie Wetter- und Kalenderdaten wird sowohl ein Strom- als auch Erdgasverbrauch prognostiziert.

Es wurden drei Modellausprägungen für unterschiedliche Zeithorizonte entworfen. Es wird vorgestellt, inwiefern DL dazu beiträgt Residuallasten in nachgelagerten Energieversorgungsnetzen zu reduzieren. Bild 1 veranschaulicht die Einbindung des Prognosemodells in eine Smart Grid-Umgebung. In dieser gibt es ein zentrales Energiemonitoring, dass mit allen Verbrauchern und Erzeugungsanlagen vernetzt ist. Die Einbindung des Prognosemodells erfolgt von einer übergeordneten Managementebene.

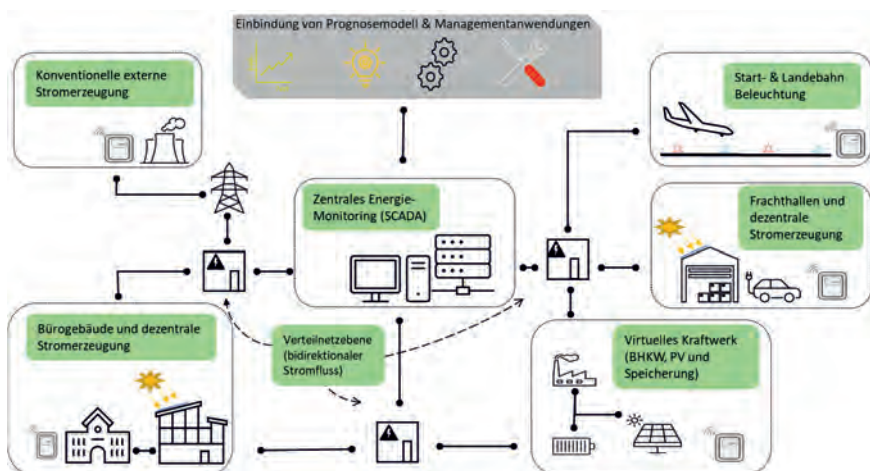


Bild 1: Einbindung des Prognosemodells in eine Smart Grid-Umgebung

2. Stand der Technik

2.1. Smart Grid

Für den Ausgleich der durch den Ausbau der volatilen Stromerzeugung entstehenden Residuallasten, gilt es, den neuen Anforderungen für Verteilnetze durch intelligente Vernetzung von Erzeugung, Verbrauch und Speicherung durch passende Informations- und Kommunikationstechnik, gerecht zu werden. Grundvoraussetzung hierbei ist eine Netztransparenz, die es ermöglicht, Kenntnis über den vergangenen, den aktuellen und den zukünftigen Netzzustand zu haben. Darüber hinaus müssen für die Nutzung der Netzzustandsdaten intelligente Messsysteme an relevanten Verbrauchern in den Versorgungsnetzen implementiert sein.

Es gilt nach der Smart Grid-Definition eine Interaktion zwischen dem Netzbetrieb und Marktaktivitäten herzustellen, um den Betrieb des Verteilnetzes zu optimieren und eine Versorgungssicherheit herzustellen. Virtual power plants, bestehend aus einer Kombination von zentralen Kraftwerken und verteilten Energieressourcen (Distributed Energy Resources, DER) gewährleisten die Schaffung von Flexibilitäten. In Kombination mit intelligenten Messgeräten sorgen vor allem marktbezogene Maßnahmen, wie z.B. dynamische Tarife oder Echtzeitübertragungen an den Verbraucher, für einen stabilen Netzbetrieb [2].

Diese Maßnahmen, die dem Oberbegriff Demand Response zugeordnet werden können, sorgen durch Verhaltensbeeinflussung der Verbraucher, für eine zeitliche Verlagerung der Lasten. Durch den Ausbau der E-Mobilität, ergibt sich für Elektrofahrzeuge mit bidirektionaler Lademöglichkeit, weiteres Potenzial, Netzengpässe auszugleichen [3]. Neben stationären Batteriespeichern, ergeben sich durch Technologien wie Power to Heat (P2H) oder Power to Gas (P2G) Möglichkeiten, Elektrische Energien zu speichern [2]. Grundsätzlich interagiert der Netzbetreiber mit den Marktteilnehmern nach einem Netzampelkonzept [4]. Dabei wird der aktuelle Netzzustand ermittelt und der zukünftige Netzzustand prognostiziert und im Anschluss einer der drei Ampelphasen zugeteilt. In der grünen Ampelphase ist kein Netzengpass prognostiziert. Hier werden Flexibilitäten ausschließlich marktdienlich benutzt. In einer orangenen Ampelphase wird ein Engpass prognostiziert. Das hat zur Folge, dass der Netzbetreiber Flexibilitäten abrufen. Für die Marktseite bedeutet dies, die Fahrpläne zur Vermeidung des Engpasses anzupassen. In der kritischen roten Phase greift der Netzbetreiber mit bedarfsgerechten Notfallmaßnahmen umgehend durch direkte Steuerung von Anlagen in den Betrieb ein [5]. Grundlegend für die Umsetzung jeglicher Maßnahmen ist die Informations- und Kommunikationstechnik. Es müssen Schnittstellen geschaffen werden, die es ermöglichen, eine interoperable herstellerunabhängige Kommunikation herzustellen [2]. In der Kommunikation mit eingebunden sind zudem FM, welche die Grundlage für zuvor genannte Maßnahmen geben.

2.2. Forecasting Modelle und deren Kategorisierung

FM können hinsichtlich ihres Einsatzzweckes, ihrer Fristigkeiten sowie dessen Inputparameter kategorisiert werden.

ELP werden entweder mit statistischen Methodiken oder mit Methodiken, die auf DL-Ansätzen basieren, entworfen. DL-Techniken können im Gegensatz zu den statistischen Ansätzen, auch nicht-lineare Zusammenhänge analysieren.

FM können zudem hinsichtlich ihrer verwendeten Eingabeparameter und deren Verwendung im Modell in drei Techniken geclustert werden [1]. Dazu zählen Modelle, die den Ansatz der Extrapolation bzw. Trendanalyse verfolgen. Diese Modelle ziehen die historischen Trends der Variablen mit ein, um Wachstumstrends zu extrahieren. Hier wird der prognostizierte Wert aufgrund historischer Trends ermittelt. Modelle mit Korrelationstechniken oder auch causal-based methods generieren ihren Output mithilfe von exogenen Parametern [6]. Schließlich gibt es hybride Ansätze, in denen Kombinationen beider Techniken verwendet werden, um eine entsprechende Prognosegüte zu erreichen [1].

In [7] werden Eingabeparameter in vier Kategorien eingeteilt. Die Eingabeparameter unterscheiden sich in historischen Daten, die numerische Informationen zu Lastmustern geben können sowie Nutzerinformationen, die Auskunft über Belegungszeiten oder Belegungsdichten geben. Kalendarische Informationen helfen den Modellen zum Beispiel, eine Beziehung zwischen dem Datum und der Ausgabe herzustellen. Zudem geben meteorologische Daten Auskunft über Temperaturen, Niederschlag oder relative Feuchte. Zusätzlich werden in [6] sozio-ökonomische Faktoren, die den Energieverbrauch beeinflussen, präsentiert. Letztere Eingabeparameter informieren über den sozio-ökonomischen Entwicklungsstand innerhalb einer Gesellschaft. Unter dieser Kategorie fallen z.B. Arbeitslosenquoten oder Bruttoinlandsprodukte [1].

Zuletzt kann eine weitere Kategorie gebildet werden, die die Fristigkeit der Modelle clustert. Eine einheitliche Definition der Fristigkeiten gibt es in den wissenschaftlichen Werken nicht [8]. In [8] und [9] werden jedoch zeitliche Annäherungen vorgestellt, um unterschiedliche Fristigkeiten der Modelle zu kategorisieren. Folgend werden die Zeithorizonte unterschiedlicher Fristigkeiten zur Kategorisierung von FM dargestellt. (Tabelle 1)

Tabelle 1: Prognosehorizonte

Fristigkeit	Zeitlicher Prognosehorizont	Prognosezweck
sehr kurzfristig (very short term)	einige Minuten bis zu einer Stunde	Netzbetrieb in Echtzeit Strompreis
kurzfristig (short term)	eine Stunde bis zu mehreren Stunden	Betriebssicherheit auf dem Strommarkt
mittelfristig (medium term)	mehrere Stunden bis zu einem Monat	Netzausbauplanung
langfristig (long term)	einen Monat bis zu einem Jahr oder mehr	Instandhaltung Optimale Betriebskosten

Dabei gibt es Modelle, die Vorhersagen für die nächsten Minuten tätigen sowie Modelle, die Prognosen über Jahre tätigen. Kurzfristige Prognosen (very short term und short term) sind für die Steuerung von Stromnetzen im Echtzeitbetrieb erforderlich. Hierdurch kann zum Beispiel das bedarfsgerechte Zu- oder Abschalten von Energieerzeugungseinheiten optimiert werden, sodass die Netzversorgungssicherheit gesteigert wird. Außerdem haben die Prognosen einen wirtschaftlichen Nutzen. Kurzzeitige Preisprognosen sorgen durch optimierten Einsatz von Energieerzeugern und Energieverbrauchern für Betriebskosteneinsparungen bei Energieversorgern und Netzbetreibern [10]. Die mittelfristige Strombedarfsprognose kann den Betrieb und die Ausbauplanung des Energienetzes optimieren. Sie ist besonders hilfreich für Unternehmen, die in einem deregulierten Energiemarkt tätig sind. Durch mittelfristige Lastvorhersagen können Energiezukäufe sowie Energieverkäufe besser geplant werden, was die Chance bietet, vorteilhafte Verträge auszuhandeln [11]. Langfristige Lastprognosen werden verwendet, um Kapazitätsplanungen zu tätigen. Sie unterstützen Energieversorger bei Investment-Entscheidungen sowie Strategieentwicklungen zum Ausbau [12].

3. Datengrundlage

Auf dem Flughafenbetriebsgelände, mit einer Fläche von 1.000 Hektar, findet der Flugbetrieb rund um die Uhr an sieben Tagen der Woche statt. Versorgt wird das Gelände mit Wärme, Kälte und Strom von dem sich auf dem Betriebsgelände befindlichen Kraftwerk. In dem Kraftwerk wird die Wärmeenergie im Grundlastbetrieb von BHKW-Modulen und im Spitzenlastbetrieb zusätzlich von Heißwasserkesselanlagen erzeugt. Die Brennstoffe hierfür sind Erdgas und Heizöl. Der Brennstoff Heizöl wird zur Redundanzsicherung bereitgestellt und kann lediglich von den Kesselanlagen verbrannt werden. Kältekompressions- und

Absorptionskältemaschinen versorgen das Areal mit Kälteenergie. Den Strombedarf decken die BHKW-Module, die auf dem Gelände verbauten PV-Anlagen sowie Notstromaggregate. Für den Fall, dass der Strombedarf die erzeugte elektrische Energie aus den zuvor genannten Anlagen übersteigt, wird das Gelände über den vorgelagerten Energieversorger (vEVU) bedarfsgerecht versorgt. Das entwickelte Modell soll Vorhersagen für den Strom- sowie den Erdgasverbrauch mithilfe von exogenen Parametern tätigen. Das Modell wird mit den Daten der letzten dreieinhalb Jahre trainiert. Die zeitliche Auflösung ist viertelstündlich. Nachdem das Data-Preprocessing inklusive Datenexploration und -visualisierung umgesetzt worden ist, ergibt sich ein aufbereiteter Datensatz. In Abbildung Bild 2 sind die Kategorien der exogenen Parameter dargestellt.



Bild 2: Kategorien der Inputparameter

In den historischen Daten sind Stromerzeugung durch das BHKW, die PV-Anlagen und die Notstromaggregate inkludiert. Die Nutzerinformationen sind flughafenspezifisch und geben viertelstündlich Informationen über zu- und ausgeladene Fracht sowie zu- und ausgestiegene Passagiere. Weitere Eingabeparameter für das Modell sind Kalenderdaten. Das Modell wird mit Informationen zur jeweiligen Uhrzeit gespeist. Zwei weitere Parameter geben Auskunft, ob für den jeweiligen Zeitstempel, Schulferien oder Feiertage in Nordrhein-Westfalen vorliegen. Die meteorologischen Daten sind der Plattform des Deutschen Wetterdienstes entnommen und umfassen die Lufttemperatur, die relative Feuchte, den Luftdruck und die Niederschlags-höhe.

4. Modellarchitektur

Um den gesamtheitlichen Energieverbrauch am Flughafen zu prognostizieren, werden der prognostizierte Strom- und Erdgasverbrauch subsumiert. Im Vergleich zu einer Simulation, werden für eine ELP auf Basis von historischen Daten, keine Eigenschaften von Baustoffen

und der Technischen Gebäudeausrüstung benötigt [13]. Zudem werden Annahmen zu einem Nutzerverhalten nicht benötigt [14]. Bei einer Simulation des Energieverbrauchs für Zusammenschlüsse von Gebäuden mit heterogenen Nutzungsformen, müssen die zuvor genannten Parameter zusammengefasst werden, was zur weiteren Unschärfe der ELP führt. Die angewandte Methode wird als eine causal-based method kategorisiert. Das Modell kann abhängig von den exogenen Parametern sowohl für kurzfristige-, mittelfristige- als auch für langfristige Prognosen eingesetzt werden. Da zwischen den exogenen Parametern nicht-lineare Zusammenhänge bestehen, wird eine Methode des DL angewendet. Für die Vorhersage des Strom- sowie Erdgasverbrauchs wird ein Feed Forward Artificial Neural Network entwickelt. Abbildung Bild 3 zeigt den Aufbau des entwickelten KNN.

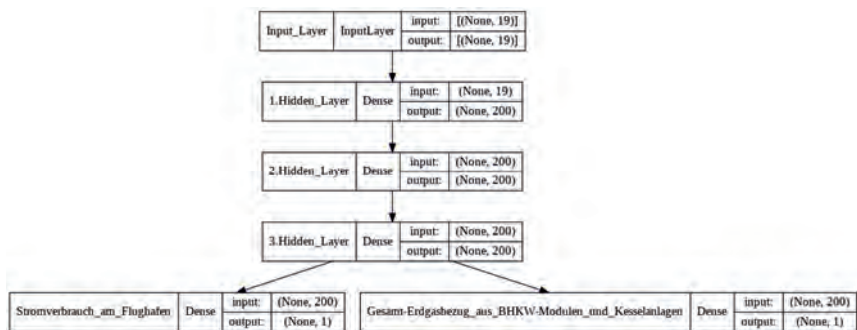


Bild 3: Schema der entwickelten Modellarchitektur

Das KNN besteht aus einem Input Layer, in dem alle Inputparameter aus Kapitel 3 gespeist werden. Darauf aufbauend sind drei Hidden Layer implementiert. Jeder Hidden Layer besitzt 200 Hidden Units. Als Aktivierungsfunktion in den Hidden Layern wird die ReLU Funktion verwendet. Um mehrere Modelle für die ELP zu vermeiden, gibt es im Output Layer zwei Ausgangsgrößen, die simultan vorhergesagt werden. Als Metrik zur Evaluation der Prognosegüte wird der mittlere absolute prozentuale Fehler (MAPE) verwendet. Der Algorithmus für die Back Propagation ist der ADAM-Optimierungsalgorithmus. Außerdem wird die Batch Größe auf 32 Trainingsbeispiele pro Batch festgelegt. Ein Trainingsdurchlauf wird bei 50 Epochen beendet. Die entwickelte Architektur des Modells sowie die exogenen Parameter ermöglichen einen variablen Prognosehorizont. Durch variable Prognosezeiträume kann das Modell unterschiedliche Anwendungsfälle abbilden. Es werden drei Modellausprägungen vorgestellt, die für unterschiedliche Zwecke verwendet werden können. Jene unterscheiden sich durch ihre

Inputparameter. Folgende Abbildung Bild 4 gliedert die Modellausprägungen zwischen kurzfristigen, mittelfristigen- und langfristigen Prognosehorizonten. Die kurzfristige Energielastprognose soll im Anwendungsfall Prognosen für die kommenden 48 Stunden geben. Bei der Entwicklung wurde bestimmt, dass die mittelfristige Energielastprognose (MELP) mithilfe ihrer Parameter, Prognosen für einen Zeitraum von bis zu einem Monat tätigen kann. Die langfristige Energielastprognose (LELP) wird für ein Jahr festgelegt.

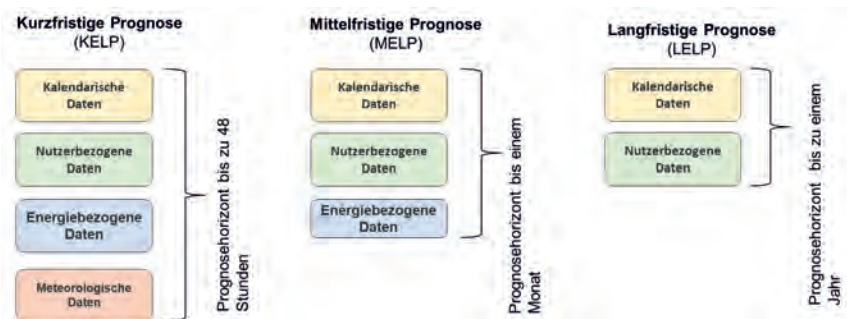


Bild 4: Prognosehorizonte für die entwickelten Modellausprägungen

5. Ergebnisse

In Tabelle 2 werden die unterschiedlichen Prognosegenauigkeiten dargestellt. Die Ausprägungen gliedern die Tabelle. Für jede Ausprägung des KNN ist die Prognosegüte sowohl für den Strom- als auch den Erdgasverbrauch dargestellt. Es ist ersichtlich, dass der Fehler mit kürzer werdenden Prognosehorizont geringer wird. Eine größere Anzahl von Parametern führt zu einer besseren Prognosegüte. Die KELP hat infolgedessen sowohl beim Strom- als auch beim Erdgasverbrauch die beste Prognosegüte.

Tabelle 2: Prognosegüten für einzelne Modellausprägungen

	Prognosegüte (MAPE für Testdatensatz)	
	Stromverbrauch am Flughafen	Erdgasverbrauch am Flughafen
	[%]	[%]
MAPE für KELP	3,92	3,48
MAPE für MELP	4,97	6,02
MAPE für LELP	5,22	9,71

Die Ergebnisse sind mit den Untersuchungen der aktuellen Forschung vergleichbar. Diese erreichen je nach Prognosehorizont eine Prognosegüte von 2% bis 16% [6,12,15]. Dabei ist zu berücksichtigen, dass die Prognosegenauigkeiten der Untersuchungen nicht vollumfänglich mit dem in dieser Arbeit entwickelten Modell vergleichbar sind. Es werden andere exogene Faktoren, zeitliche Auflösungen und Zeithorizonte betrachtet. In Abbildung Bild 5 werden exemplarisch die Stromverbräuche aus dem Testdatensatz mit den dazugehörigen Prognosen verglichen. Auf der Abszisse befinden sich die tatsächlichen Stromverbräuche. Auf der Ordinatennachse sind die prognostizierten Werte abgebildet. Unterschiedliche rötliche Farbtiefen ergeben sich durch ein gehäuftes Auftreten einzelner Punkte an bestimmten Stellen.

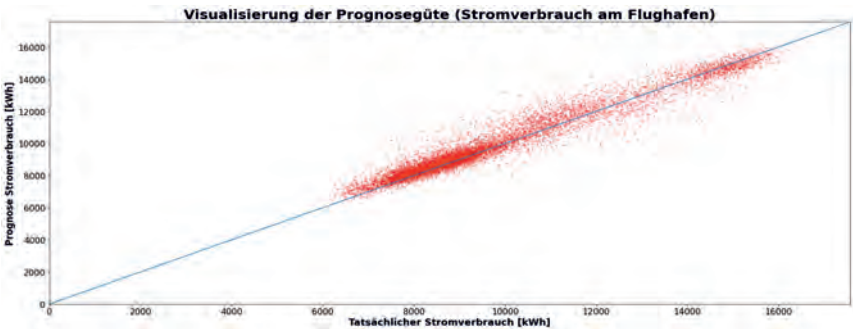


Bild 5: Visualisierung der Prognosegüte für den Stromverbrauch

Der Abbildung ist ein linearer Verlauf zu entnehmen. Der Großteil der Prognosen stimmt mit den realen Werten überein. Es ist erkennbar, dass sich parallel zur in blau gezeichneten Proportionalen in beide Richtungen ein rötlicher Bereich bildet. Es wird ersichtlich, dass das Modell auf Verbräuche, die häufiger auftreten, präzisere Prognosen geben kann. Die

Abweichungen zur blauen Linie sind für Verbräuche von ca. 6.000 bis 10.000 kWh geringer als die von 10.000 bis 16.000 kWh.

6. Einbindung des entwickelten Modells am Flughafen

Zur Umsetzung eines Smart Grid am Flughafen wird derzeit die dafür benötigte Infrastruktur errichtet. Das verhindert noch den Einsatz des entwickelten Modells für eine intelligente Steuerung von Energieerzeugung und Verbrauch in der Praxis. Es wird das im Rahmen dieser Arbeit entwickelte Konzept des Energiesystems am Flughafen Köln Bonn vorgestellt.

Dem BHKW kommt eine zentrale Rolle zu, da es ganzjährig im Grundlastbetrieb wärmegeführt betrieben wird. In der Zukunft soll es mit CO₂ neutralem Biogas betrieben werden. Die Heißwasserkesselanlagen werden weiterhin mit Erdgas gespeist. Es gilt den Betrieb jener Anlagen aus ökologischer Perspektive zu vermeiden.

Konkret wird durch den Ausbau der PV-Anlagen um die Mittagszeit mehr regenerative Energie zu Verfügung stehen, als benötigt. Die Möglichkeit BHKW-Module entsprechend zu drosseln, um Residuallasten auszugleichen, wird der Flughafenbetreiber nicht nutzen, da ansonsten die Wärme- und Kälteversorgung gefährdet wird. Betriebsbedingt benötigt der Flughafen auch in den Nachtstunden elektrische Energie. In diesen Stunden ist der erzeugte Strom durch die BHKW-Module nicht ausreichend. Das Ziel ist es, möglichst wenig Strom vom vEVU zu verbrauchen, stattdessen mit der Erzeugung von Strom aus PV-Anlagen und dem Strom der BHKW-Module eine Versorgung sicherzustellen.

Durch die Sektorenkopplung des Verkehrs in das Stromnetz wird die elektrifizierte Automobilflotte in das Smart Grid eingebunden. In Zeiten, in denen die Erzeugung den Bedarf übersteigt, erhalten die Betreiberfirmen von elektrifizierten Fluggastbrücken, Flugzeugschleppern sowie Personenkraftwagen eine Benachrichtigung, dass sie, falls es der Betrieb zulässt, die Akkumulatoren der Fahrzeuge an entsprechenden Ladesäulen aufladen sollen. Da das Smart Grid Informationen darüber hat, welche Fahrzeuge am nächsten Tag benutzt werden sollen, kann durch Steuerung der Ladesäulen ein bidirektionaler Energiefluss erfolgen.

Durch die Prognose wird mittels Demand Response außerdem ein Anreiz geschaffen, Strom in Zeiten zu verbrauchen, in denen ein großes Angebot von elektrischer Energie aus regenerativen Energien zur Verfügung steht. Es findet eine Lastverschiebung durch unterschiedliche Preissignale statt. Wenn diese für den Verbraucher nicht verpflichtenden Maßnahmen nicht wahrgenommen werden, wird der überschüssige Strom mithilfe der P2G-Technologie zu Wasserstoff umgewandelt. Der erzeugte Wasserstoff kann gespeichert, mit dem Biogas gemischt und im Anschluss durch die BHKW-Module verbrannt werden.

Außerdem vollzieht der Flughafenbetreiber auch strategische Geschäftsentscheidungen, die sich positiv auf die ökonomische Performance auswirken. Hierbei sind besonders die MELP oder LELP zu berücksichtigen. Energiemengen, welche trotz der oben genannten Maßnahmen zugekauft werden müssen, sind durch eine rechtzeitige Prognose bereits bekannt. Dies gilt insbesondere auch für den Fall, dass die vorhandenen Energieerzeugungsanlagen gewartet werden müssen. Da der vEVU im Voraus Informationen über die bereitzustellenden Energiemengen hat, können preiswerte Verträge geschlossen werden.

7. Fazit

Das entwickelte Modell prognostiziert den Strom- und Erdgasverbrauch am Flughafen Köln Bonn. Es wurde ein KNN zur Vorhersage ausgewählt, da es im Vergleich zur einer Simulation keine Annahmen über Eigenschaften von Baustoffen, Nutzerverhalten oder technischer Gebäudeausrüstung tätigen muss. Das KNN prognostiziert die Verbräuche auf Basis von historischen Daten für unterschiedliche Ausprägungen, die sich abhängig vom Prognosehorizont ergeben. Die Fehler der unterschiedlichen Ausprägungen liegen für den Strom- und Erdgasverbrauch zwischen drei und zehn Prozent. Für die Einbindung des Modells in ein Smart Grid, bedarf es vorerst noch grundlegender Infrastrukturarbeiten. Dabei ist insbesondere zu beachten, dass heterogene Komponenten, die sowohl als aktive als auch als passive Akteure implementiert sein können, kommunikationsfähig sind. Aufgrund der Ergebnisse und praktischen Umsetzbarkeit, wird der Modellansatz der MELP in die Praxis für die nächste Ressourcenbeschaffung bei präventiven Instandhaltungsmaßnahmen der Energieerzeugungsanlagen eingebunden. Da frühzeitig Kenntnis über den benötigten Strombedarf besteht, können preiswerte Verträge für die bereitzustellenden Energiemengen geschlossen werden. In Zukunft werden weitere DL-Methodiken für eine Trendanalyse des Strom- und Erdgasverbrauchs am Flughafen entworfen und mit der aktuellen Methodik verglichen. Zudem wird der Datensatz in weiteren Schritten sukzessive erweitert.

8. Literaturangaben

- [1] Nti, I. K., Teimeh, M., Nyarko-Boateng, O. und Adekoya, A. F. (2020): Electricity load forecast-ing: a systematic review. Journal of Electrical Systems and Inf Technol (7), 13
- [2] Komarnicki, P., Lombardi, P., Styczynski, Z. A (2021): Elektrische Energiespeicher-systeme, S.9-39, Berlin, Heidelberg: Springer Verlag.
- [3] Aichele, C., Doleski, O. D. (2014): Smart Market: Vom Smart Grid zum intelligenten Energiemarkt, S.3-59, Wiesbaden: Springer Fachmedien Wiesbaden.

- [4] BDEW Bundesverband der Energie- und Wasserwirtschaft e.V. (2017): Konkretisierung des Ampelkonzepts im Verteilungsnetz, <https://www.bdew.de/service/stellungnahmen/konkretisierung-ampelkonzept-verteilungsnetz/>, abgerufen am 30.04.2022
- [5] Etezadzadeh, C. (2020): Smart City–Made in Germany, S.291-301, Wiesbaden: Springer Fachmedien Wiesbaden.
- [6] Hakob, Grigoryan (2021): Electricity Consumption Prediction using Energy Data, Socio-economic and Weather Indicators. A Case Study of Spain. 9th International Conference on Control, Mechatronics and Automation (ICCMA), 2021, S. 158-164
- [7] Mariano-Hernández, D., Hernández-Callejo, L., García, F. S., Duque-Perez, O. und Zorita-Lamadrid, A. L. (2020): A Review of Energy Consumption Forecasting in Smart Buildings: Methods, Input Variables, Forecasting Horizon and Metrics. Applied Sciences, Jg. 10, Nr. 23), 8323
- [8] Amarasinghe, K., Marino, D., L. und Manic, M. (2017): Deep neural networks for energy load forecasting. IEEE 26th International Symposium on Industrial Electronics (ISIE), S. 1483–1488
- [9] Srinivasan, D., Swee Sien Tan, Cheng, C.S. und Eng Kiat Chan (1999): Parallel neural network-fuzzy expert system strategy for short-term load forecasting: system implementation and performance evaluation. IEEE Transactions on Power Systems, Nr. 3, S. 1100–1106
- [10] Kyriakides, E. und Polycarpou M., (2007): Short Term Electric Load Forecasting: A Tutorial in Studies in Computational Intelligence-Bd. 35, S. 391–418, Berlin, Heidelberg: Springer Verlag
- [11] Gonzalez-Romera, E., Jaramillo-Moran, M.A. und Carmona-Fernandez, D. (2006): Monthly Electric Energy Demand Forecasting Based on Trend Extraction. IEEE Transactions on Power Systems, S. 1946–1953
- [12] Ekonomou, L. (2010): Greek long-term energy consumption prediction using artificial neural networks. Energy, S. 512–517
- [13] Zhao, H. und Magoulès, F. (2012): A review on the prediction of building energy consumption. Renewable and Sustainable Energy Reviews, Nr. 6, S. 3586–3592
- [14] de Wilde, P. (2014): The gap between predicted and measured energy performance of buildings: A framework for investigation. Automation in Construction, Nr. 14, S. 40–49
- [15] Elkamel, M., Schleider, L., Pasiliao, E. L., Diabat, A. und Zheng, Q. P (2020): Long-Term Electricity Demand Prediction via Socioeconomic Factors-A Machine Learning Approach with Florida as a Case Study. Energies, Jg. 13, Nr. 15, 3996

Verbesserte Workflow-Prozesse in der Gebäudetechnik und Gebäudeautomation im Kontext der Transformationspfade Energiewende und Digitalisierung

Dipl.-Ing. (FH) **Peter Knoll**, Prof. Dr.-Ing. **Martin Becker**,
Hochschule Biberach (HBC), Biberach

Kurzfassung

Die Gebäude- und Energietechnik steht vor großen Herausforderungen aufgrund der zunehmenden Digitalisierung im Bauwesen, sowie der gewaltigen Anforderungen hinsichtlich eines möglichst nachhaltigen bzw. klimaneutralen Gebäudebetriebs. Hierzu sind neue Konzepte und Methoden für ein nachhaltiges Planen, Bauen und Betreiben von Gebäuden erforderlich.

Ein wichtiger Baustein auf diesem Weg ist ein möglichst durchgängiges automationsgestütztes Energie- und Gebäudemanagement, das den Weg vom systematischen Datenmanagement über das Informationsmanagement bis hin zu einem automationsgestützten Wissensmanagement ermöglicht. Grundlage des Datenmanagements ist wiederum ein standardisiertes Kennzeichnungssystem (KS), das durchgängig von der Planung über die Ausführung bis in den Betrieb genutzt wird.

An der Hochschule Biberach wurde dazu am Institut für Gebäude- und Energiesysteme bereits vor einigen Jahren ein Kennzeichnungssystem auf Basis der VDI Richtlinie 3814 Blatt 4.1 definiert und konsequent umgesetzt.

Diese Erfahrungen wurden und werden nunmehr in dem aktuell laufenden Forschungsprojekt EnMa-HAW für das Land Baden-Württemberg genutzt, um für alle Hochschulen der angewandten Wissenschaften (HAW) ein automationsgestütztes Energiemanagement zu konzipieren. Zunächst soll dieses in einer ersten Phase an vier Pilothochschulen umgesetzt werden. Ein zentrales Ziel des Projektes ist es, die automatisierte Erfassung und Archivierung von Messdaten in einer herstellerunabhängigen Datenbank nach einer standardisierten Datenstruktur mit Bezug zum Kennzeichnungssystem der Richtlinie VDI 3814, Blatt 4.1 zu erreichen. Aufbauend auf dieser Datenbank (MySQL bzw. Maria DB) können über einen sicheren Lesezugriff (VPN) unterschiedlichste Softwaretools auf diese Daten zugreifen, ohne dass aufwändige und kostenintensive Exportfunktionen, oder projektspezifische Schnittstellen eingerichtet werden müssen. Dieses Konzept wird derzeit in Form eines Prototyps an der Hochschule Biberach umgesetzt und getestet.

1. Einleitung

Aufgrund der zunehmenden Digitalisierung im Bauwesen sowie der gewaltigen Anforderungen hinsichtlich eines möglichst nachhaltigen bzw. klimaneutralen Gebäudebetriebs steht die Gebäude- und Energietechnik vor großen Herausforderungen. Hierzu sind u.a. innovative Konzepte und neue Methoden für ein nachhaltiges Planen, Bauen und Betreiben von Gebäuden erforderlich. Hierbei besteht noch ein großes Optimierungspotenzial, indem die zurzeit noch häufig vorliegenden Brüche während der Planungs-, Ausführungs- und Betreiberprozesse durch einen durchgängigen Engineering-Prozess und verbesserte Workflow-Prozesse nach Bild 1 verbessert werden.

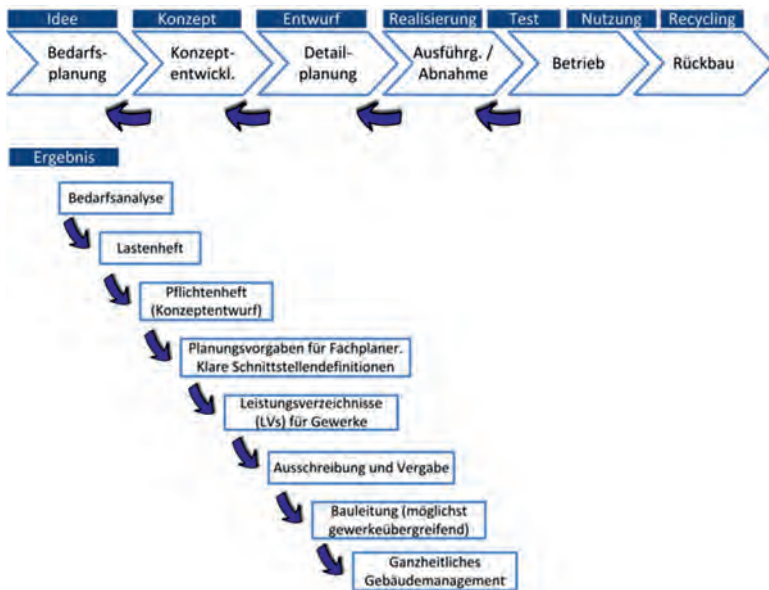


Bild 1: Verbesserter, möglichst durchgängiger Engineering- und Workflow-Prozess in der Gebäudetechnik

Ein wichtiger Baustein bei diesem Workflow-Prozess ist auch ein möglichst durchgängiges automationsgestütztes Energie- und Gebäudemanagement, das den systematischen Weg vom Daten- über das Informations- zum Wissensmanagement ermöglicht, siehe Bild 2. Hierbei gilt es auch im Sinne eines verbesserten Informationsmanagements den erforderlichen

Informationsbedarf in der Planung mit dem späteren in der Ausführung bereitgestellten Informationsangebot möglichst gut mit dem im Betrieb genutzten Informationsnutzen (z.B. Monitoring, Energiemanagement) in Überlappung zu bringen.



Bild 2: Vom Datenmanagement über das Informationsmanagement zum automationsgestützten Wissensmanagement

Grundlage des Informationsmanagements ist wiederum ein systematisches Datenmanagement, welches wiederum ein standardisiertes Kennzeichnungssystem (KS) benötigt, das durchgängig von der Planung über die Ausführung bis in den Betrieb genutzt werden kann.

Diese grundlegenden Aspekte eines verbesserten Engineering-Prozesses bzw. Workflows bei Planung, Ausführung und Betrieb von Gebäuden werden in einem Pilotprojekt für ein verbessertes Energie- und Gebäudemanagement für die Anwendungsdomäne Hochschulen für angewandten Wissenschaften (HAW) des Landes Baden-Württemberg berücksichtigt und erstmals exemplarisch an vier Pilothochschulen getestet.

Das Land Baden-Württemberg hat sich zum Ziel gesetzt, eine annähernde Klimaneutralität in der Landesverwaltung bis zum Jahr 2030 zu erreichen. Der neu verabschiedete Hochschulfinanzierungsvertrag schreibt vor, dass zukünftig in den Struktur- und Entwicklungsplänen der Hochschulen ein Kapitel für den Klimaschutz integriert sein muss, in dem realisierbare und messbare Ziele zur Energieeffizienz festgelegt werden. Somit wird durch den Hochschulfinanzierungsvertrag, als auch in dem Energie- und Klimaschutzkonzept Baden-Württemberg die

besondere Bedeutung von Maßnahmen zur Effizienzsteigerung, durch die flächendeckende Einführung von Energiemanagementsystemen, hervorgehoben. Um die Fortschritte der Energieeffizienzsteigerung und somit die CO₂-Einsparung messbar zu machen, wird an den Hochschulen zudem die Einführung eines Monitoring-Systems in enger Abstimmung mit dem Finanz- und Wissenschaftsministerium des Landes Baden-Württemberg erarbeitet.

Diese Rahmenbedingungen stellen die HAW vor enorme Herausforderungen, da viele Hochschulen dieser Hochschulart aufgrund ihrer Größe und historischen Entwicklung nicht mit einem durchgehenden Gebäudeautomationsmanagement (GA-M) und/oder einem Energiemanagementsystem (EmS) ausgestattet sind und keine durchgehende Struktur bzgl. ihres Energie- und Gebäudemanagements haben. Sie sind vielmehr historisch gewachsen und weisen eine Vielzahl von unterschiedlichsten Technologien bezüglich Zählerstrukturen, Automationskomponenten und Kommunikationsschnittstellen auf. Vielfach werden auch noch Zähler manuell abgelesen und erfasst, was zu einem erhöhten Zeitaufwand des Personals führt und fehleranfällig ist. Dies führt z.B. dazu, dass viele Anlagen der technischen Gebäudeausrüstung (Heizung- Lüftungs- und Klimaanlage, Kälteanlagen, Beleuchtungsanlagen, ...) über lange Zeiträume unentdeckt mit einem schlechten Wirkungsgrad bzw. schlechter Energieeffizienz betrieben werden und Optimierungsmaßnahmen an der Anlagentechnik nicht oder nur erschwert entdeckt und zeitnah umgesetzt werden können. Des Weiteren besteht in vielen Liegenschaften das Problem, dass bereits automatisch ausgelesene Zähler (z.B. durch M-Bus) in einer proprietären Datenbank abgelegt werden. Somit sind diese Daten an den entsprechenden Softwarehersteller gebunden, was die technologie- und herstellerunabhängige Weiterverarbeitung der Energiedaten behindert.

Diese Rahmenbedingungen erschweren ein automationsgestütztes Energie- und Gebäudemanagement mit standardisierten Energieanalysen und -berichten und letztlich einen nachhaltigen Betrieb der Liegenschaften aus Betreibersicht.

Im folgenden Kapitel 2 wird auf die aktuelle technologische Umsetzung eingegangen. Kapitel 3 erläutert das Konzept des hierbei verwendeten Kennzeichnungssystems (KS) als wichtiger Baustein für ein systematisches und herstellerunabhängiges Daten- und Informationsmanagement. Für weitere Details zur Konzeption und Umsetzung sei auf die Veröffentlichung von *Knoll, Becker*, hingewiesen.

2. Technologische Umsetzung

Ein zentrales Hauptziel des Pilotprojektes ist es, eine automatisierte Archivierung von Messdaten in einer herstellerunabhängigen Datenbank nach einer standardisierten Datenstruktur mit Bezug zum Kennzeichnungssystem der Richtlinie VDI 3814, Blatt 4.1 zu erreichen. Aufbauend auf dieser Datenbank (MySQL bzw. MariaDB) können über einen sicheren Lesezugriff (VPN) unterschiedlichste Softwaretools auf diese Daten zugreifen, ohne dass aufwändige und kostenintensive Exportfunktionen, oder projektspezifische Schnittstellen eingerichtet werden müssen. Hierbei handelt es sich zunächst um die strukturierte und vollständige Zählerdatenerfassung (Strom, Wärme, Kälte, Gas, Wasser), die in Zukunft noch um passende Prozessdaten aus Anlagen und Räumen für einen optimierten Gebäudebetrieb erweitert werden können. In der Pilotphase wurde das ModbusTCP Protokoll verwendet, wobei die Protokolle BACnet und MQTT in einer weiteren Ausbaustufe ergänzend mitberücksichtigt werden können. Auf Grund der offenen Systemarchitektur der Datenbank können beliebige Analysetools (z.B. Matlab/Simulink, Power BI, Excel, Python usw.) parallel zur EM Software verwendet werden. Die EM Software beschreibt auf Basis der importierten Messdaten aus der offenen Datenbank dann wieder eine separate Datenbank für optimierte Zugriffszeiten im Frontend zur Visualisierung von Berichten, Dashboards usw.

Bild 3 zeigt das Systemkonzept im EnMa-HAW Projekt auf. Ein wichtiger Baustein bildet hierbei die Smart Energy Box (SEB), die über Ethernet eingebunden ist. Prinzipiell ist auch ein direkter M-Bus Anschluss an die SEB möglich. Im Projekt wird aber mit M-Bus / Modbus Gateways über Ethernet gearbeitet, weil eine direkte M-Bus Verbindung zwischen verschiedenen Gebäuden auf Grund der großen Leitungswege nicht optimal ist. Zähler und Komponenten, die direkt über Modbus TCP kommunizieren, können ebenfalls direkt eingebunden werden. Über die SEB werden die Messdaten in eine MySQL bzw. Maria Datenbank auf Basis des Kennzeichnungssystems abgelegt, über die dann ein flexibler Zugriff für die spätere Auswertung möglich ist. In der Pilotphase wurden eine Reihe von verschiedenen Direktzugriffen (Excel, Matlab/Simulink, Power BI, MySQL Workbench usw.) getestet. Der übergeordnete Zugriff auf Basis einer cloudbasierten Energiemanagementsoftware erfolgt im Projekt über einen VPN-Tunnel, mit welchem die Messdaten aus der MySQL bzw. Maria DB auch in die Datenbank eines spezifischen Energiemanagementsystems zur weiteren Analyse und Auswertung (hier EDL Portal der Firma MST) übertragen werden können.

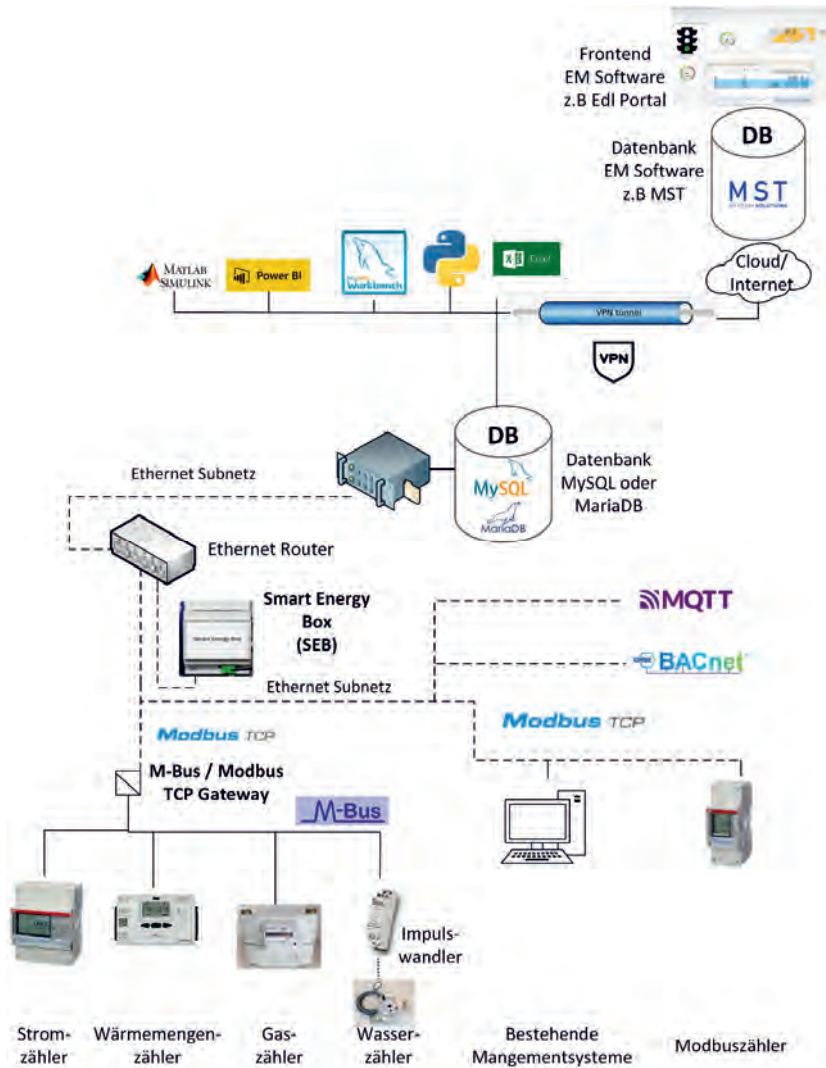


Bild 3: Systemkonzept der Zählerstruktur mit Datenbank und Visualisierung
 [Bildquellen Zähler: Stark Elektronik GmbH]

Folgende Hauptfunktionen wurden im Rahmen des Projektes als Anforderungen gestellt:

- a) Schnittstellen:
 - M-BUS (mit Autoscanfunktion)
 - ModBus TCP (Zähler, M-BUS Land Gateways, IoT-Geräte)
 - CSV Export mit Mailfunktion
 - MQTT (für zukünftige Erweiterung geplant)
 - BACnet (für zukünftige Erweiterung geplant)
- b) MySQL bzw. Maria Datenbank
 - Automatische Exportfunktion der Messdaten
 - Verwendung eines Kennzeichnungssystems nach VDI 3814, 4.1
 - Übernahme von Flächenbezeichnungen des Landes
 - Diverse Beschreibungen zur Übernahme in die MySQL Datenbank
 - Pufferung der Messdaten auf der Smart Energy Box, falls aktuell keine Verbindung zur MySQL Datenbank besteht. Bei der Wiederherstellung der Verbindung werden die Messdaten dann synchronisiert.
- c) Berechnungen
 - Erstellung von virtuellen Messstellen
 - Berücksichtigung von Korrekturfaktoren
- d) Visualisierung
 - Browserbasierte Visualisierung zur schnellen Überprüfung bzw. Analyse von Datenpunkten
 - historischer Darstellung mit verschiedenen Aggregationsstufen
- e) Update und Dokumentation
 - Automatische Updatefunktion im Betrieb aller Funktionen der Box
 - Dokumentation sämtlicher Funktionen

Im Folgenden wird die Grundstruktur der Smart Energy Box (SEB) erläutert, die auf der Symbox Technologie der Fa. Symcon basiert, siehe Abb. 4. „SymOS“ umfasst als Kern das Betriebssystem der SEB. Das Modul „Smart Energy Box“, welches spezifisch für das Projekt EnMa-HAW entwickelt wurde, ist an „IP-Symcon“ angegliedert, welches der Konfiguration dient und eine eigene Webvisualisierung besitzt. „IP-Symcon“ ist eine frei programmierbare Visualisierungs-, Archivierungs- und Automatisierungssoftware mit der Lauffähigkeit auf vielen Hardware-Systemen. „IP-Symcon“ kann auf der SymBox, Microsoft Windows, Apple Mac OS, Ubuntu Linux, Raspberry Pi, Synology NAS, QNAP NAS und Docker installiert werden.

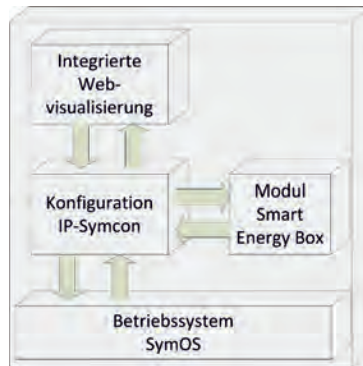


Abb. 4: Softwareaufbau der Smart Energy Box (SEB)

2. Kennzeichnungssystem (KS)

Ein wichtiger Baustein des Konzeptes ist ein einheitliches Kennzeichnungssystem (KS) für alle Zähler-, Anlagen- und Raumdaten, welches an die Richtlinie VDI 3814 Blatt 4.1 angelehnt wurde. Das Kennzeichnungssystem besteht aus insgesamt 31 Feldern und ist als Beispiel für die Hochschule Biberach (HBC) als Liegenschaft in der Bild 5 dargestellt. Unterschiedliche Campus-Standorte (hier ST=Campus Stadt) werden über die Spalte 3 Liegenschafts-Detail unterschieden. Der Installationsort des Datensammlers (M-Bus oder Modbus Gateway) wird mit der Spalte 5 und 7 festgelegt. Die Anlagennummer in Spalte 9 setzt sich aus der jeweiligen Kostengruppe der DIN 276 und einer dreistelligen laufenden Anlagennummer zusammen. Somit spiegeln die Ziffern 1-9 das Anlagenkennzeichnungssystem (AKS) der VDI 3814, 4.1 wieder. Die anschließende Betriebsmittelkennzeichnung (BKS) Ziffer 1-15 gibt an, wo der zugehörige Zähler oder das Gateway installiert ist. Mit der Ziffer 11 und 13 wird das Gebäude und der Installationsraum vorgegeben. Über die Ziffer 15 lässt sich das Betriebsmittel in den Stromlaufplänen identifizieren. Hier wird die Seitenzahl des Stromlaufplans (15), die Art des Betriebsmittels (P=zählende Messeinrichtung) und die Spaltenzahl des Installationsplans des Betriebsmittels (1) angegeben. Die Benutzeradressierung (BAS) von Ziffer 1-31 beschreibt die zugehörigen GA-Funktionen. Die Ziffer 17 gibt die Funktion des Datenpunkts (ZW=Zählwert) an. Mit der Ziffer 19 wird definiert, was genau gemessen wird (E=Elektrische Leistung). Die restlichen Ziffern von 21-31 bieten die Möglichkeit über Zusatzfelder dem Datenpunkt weitere Informationen zuzuweisen. In Bild 5 werden als Zusatzinformationen der gemessene Anlagenteil für einen Heizkreis (HZK) am Vorlauf (VL) und das Medium Heizungswasser (HW) vorgegeben. Zur Unterscheidung gleichartiger Komponenten an derselben Stelle dient die Ziffer 27

(g1). Das verwendete Bus-Protokoll M-Bus (M) wird mit der Ziffer 29 vorgegeben. Ebenfalls wird dieser Datenpunkt auf die Änderung der Werte überwacht. Sollte sich der Wert für eine bestimmte Zeit nicht ändern, wird automatisch eine E-Mail als Störmeldung generiert. Dies wird mit der letzten Ziffer 31 gekennzeichnet.

Liegenschaft:	Trennzeichen	Liegenschaft - Detail	Trennzeichen	Zentrale-Gebäude	Trennzeichen	Zentrale-Raum	Trennzeichen	Anlagennummer	Trennzeichen	Einbauort-Gebäude	Trennzeichen	Einbauort-Raum	Trennzeichen	BMK-Seite	BMK	BMK-Nummer	Trennzeichen	Funktion	Trennzeichen	Datenpunkt	Trennzeichen	Zusatzfeld_01	Trennzeichen	Zusatzfeld_02	Trennzeichen	Zusatzfeld_03	Trennzeichen	Zusatzfeld_04	Trennzeichen	Zusatzfeld_05	Trennzeichen	Zusatzfeld_06
HBC	-	ST	-	B	-	112	#	420046	+	R	-	112	-	15	P	1	/	ZW	-	E	-	HZK	-	VL	-	HW	-	g1	-	M	-	Dpu
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31		

AKS nach VDI 3814 4.1

BKS nach VDI 3814 4.1

BAS nach VDI 3814 4.1

Bild 5: Kennzeichnungssystem (KS) bestehend aus Anlagenkennzeichnungssystem (KAS), Betriebsmittelkennzeichnungssystem (BKS) und Benutzeradressierungssystem (BAS), angelehnt an die VDI 3814 Blatt 4.1.

The screenshot displays the 'Kennzeichnungssystem v0.8.0' interface. On the left, a sidebar lists categories: KS Datenpunkt, Liegenschaft, Gebäude, Raum, Anlagennummer, Kostengruppe, BMK, Funktion, Datenpunkt, and various Zusatzfeld (01-06). The main area shows a 'Liste KS Datenpunkt' table with columns ID, KS-Datenpunkt, and a description. Three entries are visible, with the first one (ID 312) highlighted in red. To the right, a detailed view of the selected data point is shown, featuring a 'KS Datenpunkt > Listenreingabe bearbeiten' form. This form includes fields for 'ID', 'Liegenschaft', 'Gebäude', 'Raum', 'Anlagennummer', 'Funktionsgruppe', 'Funktion', 'Datenpunkt', and various 'Zusatzeid' (01-06) fields, each with a dropdown menu and a 'Zusatzfeld' label. A red box highlights the 'ID' field and the 'Zusatzeid 01' field.

Bild 6: Übersicht zu PHP Kennzeichnungstool mit Eingabeebene

Um eine einheitliche und automatisierte Generierung des Kennzeichnungssystems (KS) zu gewährleisten, wird im Projekt ein neu entwickeltes PHP-Tool verwendet. Dieses Tool bietet die Möglichkeiten über ein vordefiniertes Eingabeformular das Kennzeichnungssystem für die jeweilige Liegenschaft zusammenzustellen. Somit kann das Kennzeichnungssystem für die jeweiligen Zähler zügig und fehlerfrei erstellt werden. In dem Tool werden Eingabefelder und Drop Down Menüs verwendet, die wiederum auf verschiedene Listeninhalte (Gebäude, Liegenschaften usw.) zurückgreifen. Dadurch werden Fehleingaben vermieden.

Bild 6 veranschaulicht die Eingabe und die automatische Generierung des Kennzeichnungssystems. Funktionen, wie z.B. „Duplizieren“, vereinfachen die Eingabe sehr, weil dann beispielsweise bei verschiedenen Phasen nur ein Eingabepunkt abgeändert werden muss.

3. Strukturierte und cloudbasierte Ausführung

Um eine möglichst reibungslose Ausführung und Inbetriebnahme zu gewährleisten wurde in auf die cloudbasierte, kommerzielle Software Planradar zurückgegriffen, mit der jederzeit der aktuelle Ausführungsstand überprüft werden kann.

Auch das Kennzeichnungssystem hat hier eine besondere Rolle gespielt. Für die erste Ausführungsstufe wurde nur der BKS (siehe Bild 7, Markierung 1) vorgegeben, welcher aber schon in dem Kennzeichnungstool hinterlegt wurde. Im weiteren Projektierungsverlauf wurden dann die jeweiligen Einzelfunktionen der Zähler (Energie, Leistung usw.) auf Basis des BKS ergänzt.

The screenshot displays the 'ELZ Mensa' software interface, which is used for managing energy metering data. The interface is divided into several sections:

- Formular: Zähler** (Meter Form): This section contains fields for entering meter information.
 - Titel** (Title): ELZ Mensa
 - Raum** (Room):
 - Zählertyp** (Meter Type): Stromzähler (Current Meter)
 - Impulsconverter** (Impulse Converter): ☐
 - EVU** (Energy Supplier): ☐ (Marked with a red circle 4)
 - Adresse** (Address):
 - Zählernummer** (Meter Number):
 - Wandlerverhältnis einstellt auf** (Transformer ratio set to):
 - Bezeichnung III Sufzer** (Designation III Sufzer): HBC D-EO-169-1E-03 Mensa 200A
 - BKS Hochschule Biberach** (BKS Hochschule Biberach): HBC-CST-D4-169-443003-D4-169-45P1
 - Kommentar** (Comment): Zähler tauschen, Vorhanden, Nur tauschen, Abgang Mensa
 - Stand** (Status):
 - ☒ Verbaut (Installed)
 - ☐ Verkabelt (Wired)
 - ☐ Angeschlossen (Connected)
 - ☐ Markierung angebracht (Marking applied)
 - ☐ Inbetriebnahme (Commissioning)
 - Status**: Offen (Open)
- Projektbild** (Project Image): A floor plan diagram of the building with a red circle 2 indicating the location of the meter.
- Anhänge** (Attachments): A section for adding attachments, showing two images of the meter with a red circle 3.
- Journal** (Journal): A log of events, with a red circle 1 highlighting the first entry. The entries include:
 - L. Pfänder** (L. Pfänder): Angeschlossen gesetzt auf Ja (Connected set to Yes), Verbaud gesetzt auf Ja (Installed set to Yes)
 - T. Pfänder** (T. Pfänder): BKS Hochschule Biberach gesetzt auf HBC-CST-D4-169-443003-D4-169-45P1 (BKS Hochschule Biberach set to HBC-CST-D4-169-443003-D4-169-45P1), Titel gesetzt auf ELZ Mensa (Title set to ELZ Mensa)
 - T. Pfänder** (T. Pfänder): BKS Hochschule Biberach gesetzt auf HBC-CST-D4-169-443003-D4-169-45P1 (BKS Hochschule Biberach set to HBC-CST-D4-169-443003-D4-169-45P1)
 - L. Pfänder** (L. Pfänder): Verkabelt gesetzt auf Ja (Wired set to Yes)
 - P. Knoll** (P. Knoll): HBC-CST-D4-169-443003-D4-169-0A7 ist der BKS (HBC-CST-D4-169-443003-D4-169-0A7 is the BKS)
 - T. Pfänder** (T. Pfänder):

Bild 7: Umsetzung einer cloudbasierten Planungssoftware

Durch den Planbezug mit Verortung (siehe Bild 7, Markierung 2), das Projektbild (siehe Bild 7, Markierung 3), die Kommentarfunktion und die eindeutige Kennzeichnung war es auch bei einem Umfang von 98 Zählern jederzeit möglich, eine ständig abgestimmte Kommunikation im Engineering-Prozess zwischen allen Projektbeteiligten herzustellen.

Im weiteren Ausführungsverlauf können dann in diesem Tool auch sämtliche Zusatzinformationen, wie z.B. Wandler Verhältnis, Adressierung usw. ergänzt (siehe Bild 7, Markierung 4) werden, so dass eine lückenlose Inbetriebnahme- Dokumentation erfolgt.

4. Datenmanagement und Energiemanagementsoftware

Ein wichtiger Aspekt bei einem systematischen Datenmanagement war die Entkopplung der erfassten Messdaten von der Archivierung und Auswertung für das darauf aufbauende Informations- und automationsgestützte Wissensmanagement. Daher wurde, wie bereits erläutert, eine MySQL bzw. Maria Datenbank verwendet, welche durch die Smart Energy Box (SEB) als „Datensammler“ mit einer eindeutigen Struktur per Kennzeichnungssystem beschrieben wird. Basierend auf einer ausführlichen Bedarfsermittlung und Erstellung eines umfangreichen Kriterienkatalogs bezüglich der benötigten Funktionen, die durch die Energiemanagementsoftware abgedeckt werden müssen, wurde in einer europaweiten Ausschreibung die Firma MST mit dem EDL-Portal für das Projekt ausgewählt. Hierbei handelt es sich um eine cloudbasierte Softwarearchitektur und der Kompatibilität zu MySQL / Maria DB über VPN-Verbindungen.

Bild 8 veranschaulicht ein Dashboard-Beispiel aus dem EDL-Portal der Firma MST. Durch das direkte Einlesen der Datenpunkte aus der Datenbank werden sämtliche Informationen, wie z.B. das Kennzeichnungssystem, Flächen, Einheiten, Zählernummer usw. übernommen. Dadurch ist eine Filterung nach der entsprechenden Bezeichnung möglich. Die rot markierte Datenpunkt-Kennzeichnung zeigt diesen Datenpunkt passend zu dem in den Abbildungen 6 und 7 gezeigten Beispiel.

Ein Vorteil des Portals ist, dass bezogen auf unterschiedlichste Nutzergruppen (z.B. Hausmeister, Energiemanager, Studierende, wissenschaftliche Mitarbeiter, Professoren) flexibel zielgruppenspezifische Dashboards und Zugriffsberechtigungen erstellt werden können.

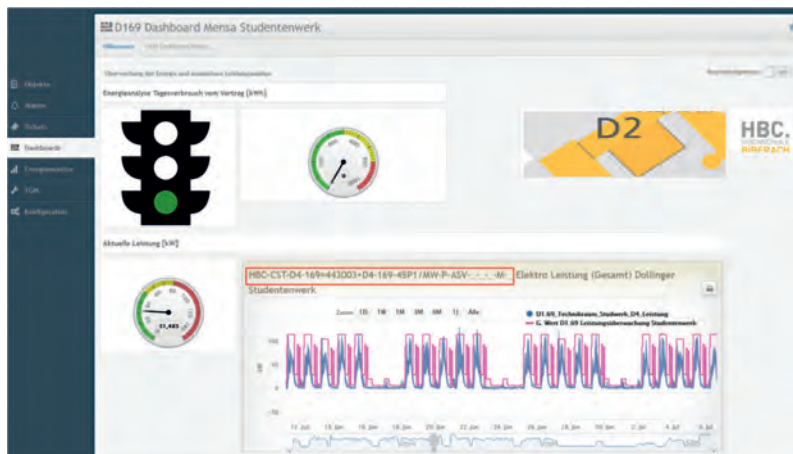


Bild 8: Beispiel für Dash-Board zu EDL- Portal mit Bezug zum Kennzeichnungssystem (KS)

5. Ausblick

Die im Pilotprojekt EnMa-HAW gewonnenen Erfahrungen bestätigen die Anwendbarkeit des Kennzeichnungssystems nach der Richtlinie VDI 3814, Teil 4.1 für die Praxis. Für die Weiterentwicklung gilt es im Kontext einer zunehmenden Digitalisierung zu beachten, dass zukünftig die Kennzeichnungssysteme dahingehend weiterentwickelt werden sollten, dass vererbaren Informationen verwendet werden können. Diese sind dann nicht mehr mit statischen Information, wie z.B. einer fest zugeordneten Raumnummer, aufgebaut, sondern können dynamisch bei Flächennutzungsänderungen (z.B. bei flexiblen Open Space Bereichen) angepasst werden. Die automatisierte Zählererfassung stellt hierbei die Grundlage für ein systematisches Flächen- und Gebäudemanagement inklusive eines agilen Energiemanagements dar. Wichtig hierbei ist auch, dass für jedes Bauvorhaben bereits in einer frühen Planungsphase (Bedarfsplanung) diese Aspekte analysiert und für die zukünftigen Betreiberprozesse passend berücksichtigt werden. Als Grundlage für ein nachhaltiges Planen, Bauen und Betreiben bietet die zunehmende Digitalisierung gute Möglichkeiten, um einen durchgängigen Planungs-, Ausführungs- und Betreiberprozess auf Basis von cloudbasierten Softwarelösungen passend zu unterstützen. Die umgesetzten Lösungen sollten dann auch kontinuierlich gepflegt und weiterentwickelt werden. Hierzu lässt sich aufgrund der Komplexität und erforderlichen IT-Kompetenz auch ein neues Dienstleistungs- und Geschäftsmodell für den Pflegedienst von Daten und Informationen ableiten.

Literatur

- [1] Becker, M.: Gebäudeautomation –Quo Vadis? Editorial HLH 11/12-2019, S. 2
- [2] Becker, M.; Knoll, P.: Weißbuch (White Paper): Nachhaltiges Planen, Bauen und Betreiben von Gebäuden, 2. Auflage, März 2019, 32 Seiten, Hochschule Biberach
- [3] Knoll, P.; Becker, M.: Kennzeichnungssystem als Basis für BIM-fähige Gebäudeautomation, 3. VDI-Fachkonferenz Gebäudeautomation bei GMA-Kongress, 02./04. Juli 2018, Baden-Baden
- [4] Knoll, P.; Straßer, M.; Becker, M.: Konzept und Umsetzung eines automationsgestützten Energiemanagements für komplexe Liegenschaften, 2. Innovationskongress Ulm, 2021, ISBN 978-3-9820843-3-6, DOI: <http://dx.doi.org/10.18725/OPARU-38503>
- [5] Becker, M.: Die neue VDI 3814 Richtlinienreihe Gebäudeautomation HLH Nr. 01-02 2021, S. 22-26
- [6] „Hochschulfinanzierungsvereinbarung Baden-Württemberg 2021 - 2025, 2020.
<https://mwk.baden-wuerttemberg.de/de/hochschulen-studium/hochschulfinanzierung-2021-bis-2025/>
- [7] VDI 3814, Blatt 4.1. Gebäudeautomaiton (GA) – Methoden und Arbeitsmittel für Planung, Ausführung und Übergabe - Kennzeichnung, Adressierung und Listen, Januar 2019
- [8] DIN 276 - Kosten im Bauwesen, Beuth, Berlin, 2018.

Prototypische Implementierung von Capability- und Feasibility-Checks in I4.0-Verwaltungsschalen der technischen Gebäudeausrüstung

Nicolai Maisch, Ralf Ulmer, Prof. Dr. Jochen Müller,

Institut für Technische Gebäudeausrüstung,

Technische Hochschule Köln;

Siwara Schmitt, Denis Uecker, Frank Schnicke,

Maximilian Conradi,

Fraunhofer IESE, Kaiserslautern

1. Kurzfassung

Der aktuelle Stand der Technik in der Gebäudeautomation ist durch starre Automatisierungslösungen geprägt. So ist z. B. die Aktivierungsreihenfolge verschiedener Wärmeerzeuger und Verbraucher vorgegeben – Auswertungen von Zustandsinformationen, wirtschaftliche und klimatische Randbedingungen o. ä. werden in deren Automatisierung nicht berücksichtigt. Neue Entwicklungen im Umfeld von Industrie 4.0 [1] ermöglichen konfigurationsfreie Interaktionen von Komponenten der Technischen Gebäudeausrüstung, die auf standardisierten Beschreibungen von Fähigkeiten (Capabilities) und Merkmalen von Assets in deren Verwaltungsschalen basieren [2]. Verwaltungsschalen als virtuelle Vertreter automatisierungstechnischer Assets erkunden in zukünftigen Automationslösungen selbstständig Merkmale und Capabilities anderer Verwaltungsschalen in ihrer Umgebung und erfüllen zustandsorientiert und flexibel Aufgaben der Gesamtautomatisierung [3]. Der händische Aufwand zur Programmierung und Parametrierung der Automation wird bei einer Erhöhung der Anlagenflexibilität deutlich reduziert. Das Paper beschreibt die prototypische Implementierung einer Automatisierungslösung am Beispiel der Wärmeversorgung eines Gebäudes mit unterschiedlichen Wärmeerzeugern (Brennwertkessel, Wärmepumpe, Solarmodul) und Verbrauchern durch konfigurationsfreie Interaktionen aktiver Verwaltungsschalen auf Basis von Capability- und Feasibility-Checks [3]. Durch den Einsatz der Anwendung und das Umsetzen des Checker-Frameworks wird die Möglichkeit flexibilisierter Interaktionen zwischen autonom handelnden Assets in Gebäudeenergienetzen nachgewiesen.

1. Einleitung

In Gebäuden werden Steuer- und Regelungsaufgaben von zentralen Automationseinrichtungen übernommen, die mit Feldgeräten der technischen Gebäudeausrüstung (TGA) verbunden sind [4], (Bild 1). Zur Realisierung der Automationsaufgabe werden die Automationseinrichtungen unter hohem manuellen Aufwand programmiert und parametrieren, wofür ein Pool an Steuer- und Regelfunktionalitäten in Form von Funktionsbausteinen zur Verfügung steht [5]. Nach der Inbetriebnahme ist die Automation starr eingestellt. Nachträgliche Umbaumaßnahmen erfordern einen hohen manuellen Aufwand beim erneuten Engineering.



Bild 1: Ebenen der Automation in Gebäuden

Die zunehmende Nutzung erneuerbarer Energien im Gebäudesektor und gestiegene Effizienzvorgaben bewirken eine Verbreitung hybrider Energieerzeugungssysteme mit mehreren Erzeugern und Energieübertragungssystemen [6]. Deren Automatisierung erfordert eine größere Flexibilität, die mit starr eingestellten Automationseinrichtungen nicht realisierbar ist. Die Automation eines Heizungssystems mit verschiedenen Energieerzeugern (z. B. Brennkessel, Wärmepumpe, Solarmodul) beinhaltet beispielsweise Schaltreihenfolgen, die Umwelt- und Nutzungsbedingungen in voreingestellte Ansteuerungen übersetzen.

Industrie 4.0 (I4.0) eröffnet andere Möglichkeiten zur Automatisierung. Digitale Zwillinge automatisierungstechnischer Assets erkunden selbstständig Eigenschaften und Fähigkeiten anderer Assets in ihrer Umgebung und erfüllen zustandsorientiert und flexibel Aufgaben der Gesamtautomatisierung. Dies setzt die Abbildung der Eigenschaften und Fähigkeiten der Assets in standardisierten virtuellen Repräsentanten voraus – sogenannten I4.0-Verwaltungsschalen. Teilmodelle der Verwaltungsschalen repräsentieren unterschiedliche Aspekte der Assets und enthalten Teilmodellelemente, in denen Asset-Merkmale abgebildet sind [2]. Durch eine Erweiterung der Verwaltungsschalen um einen aktiven Teil und die Nutzung standardisierter Kommunikationstechnologien kann ein Netzwerk flexibel interagieren-

der Verwaltungsschalen implementiert werden [7], (Bild 2). Grundlage der Interaktionen liefert das gegenseitige Abfragen von Fähigkeiten (Capability-Check) und Zuständen (Feasibility-Check) [3].

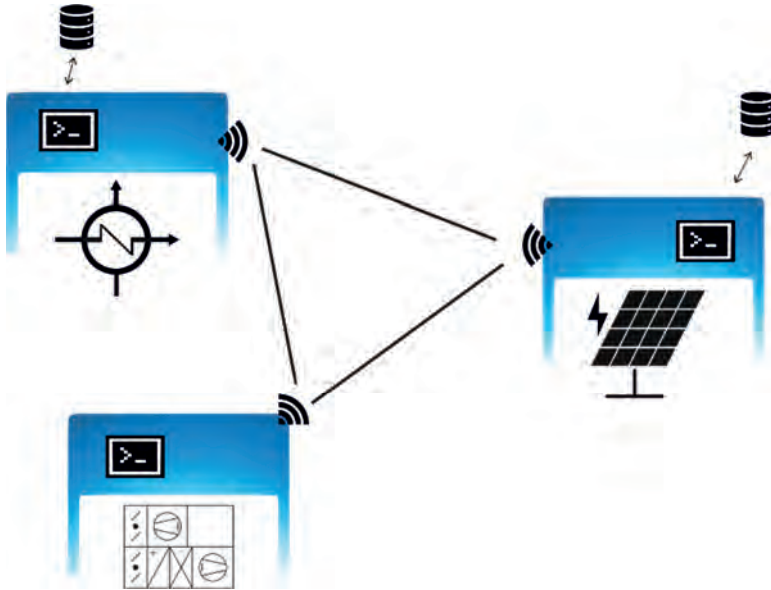


Bild 2: Interagierende TGA-Komponenten mit aktiven Verwaltungsschalen ohne externe Steuerlogik

Bestehende Konzepte zur Nutzung von Capability- und Feasibility-Checks in der Automatisierungstechnik mithilfe von I4.0 zeigen die Potentiale zur Vereinfachung des Engineerings, basieren jedoch auf einer Orchestrierung der Verwaltungsschalen [8]. Die Integration von Checker-Funktionalitäten in selbsthandelnde aktive Verwaltungsschalen fehlt dabei.

2. Fähigkeitsbasiertes Engineering als Grundlage für flexible Interaktionen selbstorganisierender Assets

In der Industrieautomatisierung besteht der Bedarf nach mehr Anlagenflexibilität mit dem Ziel einer individuellen Massenproduktion und der Herstellung verschiedenster Produktvarianten und -typen auf einer Fertigungslinie [9]. Ein Schlüssel hierzu ist fähigkeitsbasiertes Engineering, bei dem Produktionssysteme (Assets) ihre Funktionalitäten nach außen gekapselt als

Service anbieten [3]. Das Engineering basiert auf einer Abstraktion der Fähigkeitsbeschreibungen (Capabilities) der Assets losgelöst von ihrer Implementierung. Erwartete Vorteile des fähigkeitsbasierten Engineerings in der Industrie 4.0 sind neben der Flexibilität auch die Reduzierung des Engineeringaufwands [10, 11].

Das Rezept (Recipe) zur Produktion eines Produkts beinhaltet eine technologie-agnostische Reihenfolge zu erbringender Leistungen in Form von benötigten Capabilities, z. B. „Bohren, Drehen, Lackieren“. Die Auswahl von Assets zur Erbringung eines Service beginnt mit dem Capability Check, in dem angebotene Fähigkeiten existierender Assets mit den im Rezept beschriebenen abgeglichen werden, (Bild 3). [1]

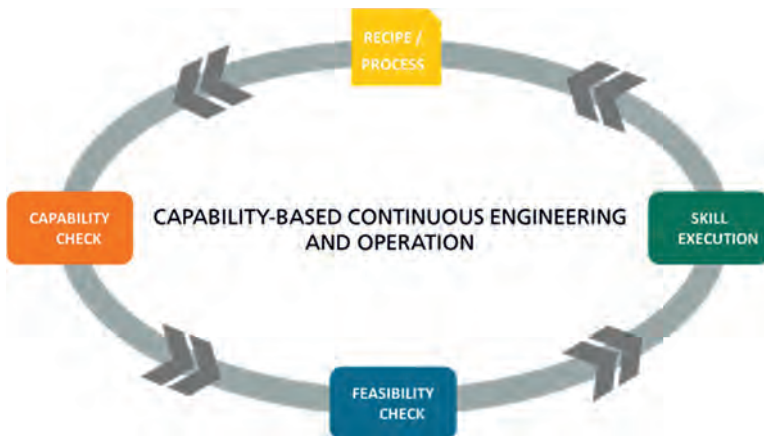


Bild 3: Zyklus des fähigkeitsbasierten Engineerings

Im Feasibility-Check erfolgt eine Überprüfung der angebotenen Services auf ihre Machbarkeit in der realen Welt, bspw. mittels einer Simulation oder mathematisch-physikalischen Berechnungen. In der Skill Execution wird eine ermittelte validierte Lösung in Form eines parametrisierten Prozessablaufs ausgeführt, (Bild 3). Der Begriff *Skill* repräsentiert die technische Realisierung einer Capability, z. B. Services, Operationen oder auch Function Blocks. [1]

3. Ontologien als Wissensbasen für fähigkeitsbasiertes Engineering

Eine Grundvoraussetzung des fähigkeitsbasierten Engineerings ist die semantisch eindeutige Beschreibung der angebotenen Fähigkeiten eines Assets. Dies lässt sich durch die Ver-

wendung von Ontologien sicherstellen, in denen Fähigkeiten formal und eindeutig beschrieben werden, sowie Relationen zu übergeordneten oder abgeleiteten Fähigkeiten spezifiziert sind [12]. Werden innerhalb einer Domäne verschiedene Ontologien als Wissensbasen genutzt, werden semantische Verbindungen untereinander ebenfalls ontologisch formalisiert.

In der TGA existieren Ontologien, die als Basis für Capability-Checks für Versorgungsaufgaben innerhalb von Gebäuden dienen [13]. Mithilfe von Referenzierungen zwischen diesen können durch semantisch ausgezeichnete Verwaltungsschalen mit ihren jeweiligen Fähigkeiten nach [14] Rückschlüsse auf die Funktionalität innerhalb eines Gebäudeenergiesystems getroffen werden [15]. Gaskessel mit der Fähigkeit „Combustion_Heating“ und Solarmodule mit der Fähigkeit „Solar_to_thermal_energy_Generating“ können dadurch beispielsweise gleichermaßen als Kandidaten zur Erfüllung einer Heizanfrage identifiziert werden.

4. Konzept einer aktiven Verwaltungsschale mithilfe eines Checker-Frameworks

Verwaltungsschalen dienen als virtuelle Repräsentanten von Assets mit ihren Merkmalen und Fähigkeiten. Sie können um einen aktiven Teil ergänzt werden, in dem Algorithmen ausgeführt werden und durch den Interaktion zu anderen Verwaltungsschalen ermöglicht wird [7]. Im Folgenden wird die Architektur eines aktiven Teils einer Verwaltungsschale auf Basis des fähigkeitsbasierten Engineerings vorgestellt. Eine aktive Verwaltungsschale ist in der Lage, Anfragen (ServiceRequests) zu verarbeiten, Checks an anderen Verwaltungsschalen durchzuführen und entsprechende Aktionen (ServiceOffering) auszuführen. Die Architektur des aktiven Teils der Verwaltungsschale besteht aus folgenden Kernkomponenten bzw. -systemen (Bild 4):

- Service-Facade: Orchestriert den Ablauf und stellt Schnittstelle nach Außen dar,
- Capability-Checker: Führt den Capability-Check durch,
- Feasibility-Checker: Führt den Feasibility-Check durch,
- Plant AAS System: Betrieb von Verwaltungsschalen im Gesamtsystem,
- Ontology: Stellt die Ontologie bereit, die für den Capability-Abgleich notwendig ist.

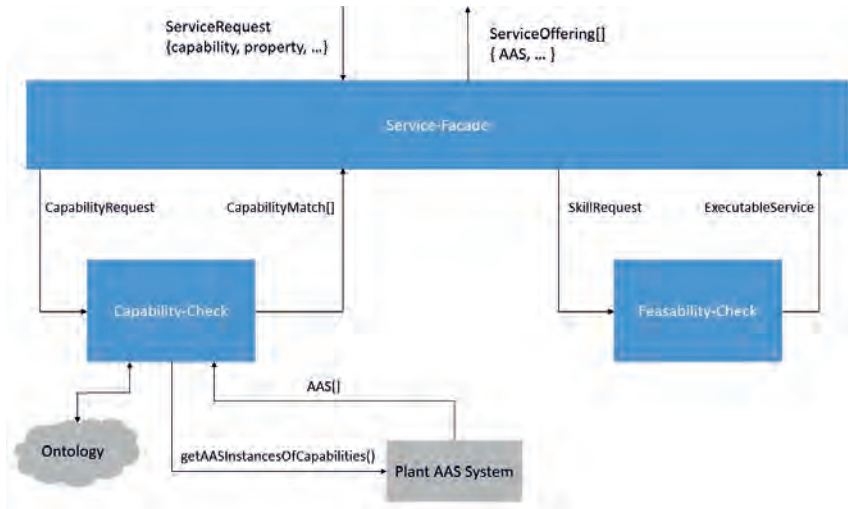


Bild 4: Integration der Capability- und Feasibility-Checks in die Verwaltungsschale

Die *Service-Facade* stellt die Schnittstelle der Capability- und Feasibility-Checks nach außen dar, indem ServiceRequests entgegengenommen und mit einem passenden ServiceOffering beantwortet werden. Die Ausführung der Checks und die Beschaffung der benötigten Informationen aus der Verwaltungsschale werden hier verwaltet. Die Service-Facade beinhaltet somit das Rezept im Sinne des fähigkeitsbasierten Engineerings und setzt die Skill Execution um.

Die beiden Komponenten *Capability-Check* und *Feasibility-Check* implementieren die Funktionalitäten zur Überprüfung der Umsetzbarkeit von Services. Durch die Bereitstellung der dafür nötigen Informationen und Operationen über das Verwaltungsschalen-Metamodell gliedern sich die Komponenten nahtlos in das bestehende Verwaltungsschalen-Ökosystem ein und es findet kein Bruch hinsichtlich Schnittstellen und Datenmodellen statt. Die Architektur bietet somit die Grundlage für die Eingliederung des aktiven Teils in Verwaltungsschalen und realisiert programmatisch das fähigkeitsbasierte Engineering.

5. Umsetzung Checker-Framework mit Eclipse BaSyx

Auf Basis der Referenzimplementierung von Verwaltungsschalen *Eclipse BaSyx* [16] wurde die Anwendung *AasApp*¹ entwickelt, mit der das vorgestellte Checker Framework implementiert wird.

Beim Starten der *AasApp* wird eine Verwaltungsschale als Datei nach [17] eingelesen. Anschließend wird ein http-Server mit den Informationen der Verwaltungsschale nach dem Standard [2] erstellt. Merkmale und Fähigkeiten der Assets, die semantisch durch Referenzen auf Ontologien definiert sind, werden dadurch innerhalb eines Netzwerks für andere Verwaltungsschalen erkundbar und können mittels einer standardisierten API abgefragt oder geändert werden [18].

Neben der passiven Bereitstellung der Asset-Informationen ermöglicht es die *AasApp* dem erstellten Server, Aktionen durch dessen REST-Schnittstelle auszuführen. Die Funktionalitäten der Verwaltungsschale werden durch das Teilmodellelement *Operation* [2] repräsentiert, in dem deren Inputs und Outputs definiert sind. Beim Ausführen einer Operation werden die Inputs eingelesen und eine vordefinierte Funktion ausgeführt.

Damit die *AasApp* für alle Assets generisch nutzbar ist, wurden die ausführbaren Funktionalitäten der Verwaltungsschale in die Verwaltungsschalen selbst integriert. Dafür wurde das Teilmodellelement *Blob* (Binary Large Object) genutzt. Dieses ermöglicht es, den Quelltext einer Datei in binär kodierter Form in die Verwaltungsschale zu integrieren [2]. Wird der ausführbare Code einer Operation in Form eines Blobs kodiert, kann dieser durch die *AasApp* eingelesen werden, indem die in dem Blob bereitgestellte Java-Klasse zur Laufzeit kompiliert und geladen wird. Dadurch wird die Anwendung um weitere Funktionalitäten erweitert, die im Blob definiert sind (Bild 5), z. B. ein Capability-Check.

¹ <https://github.com/thcologne-gart/activeAas>

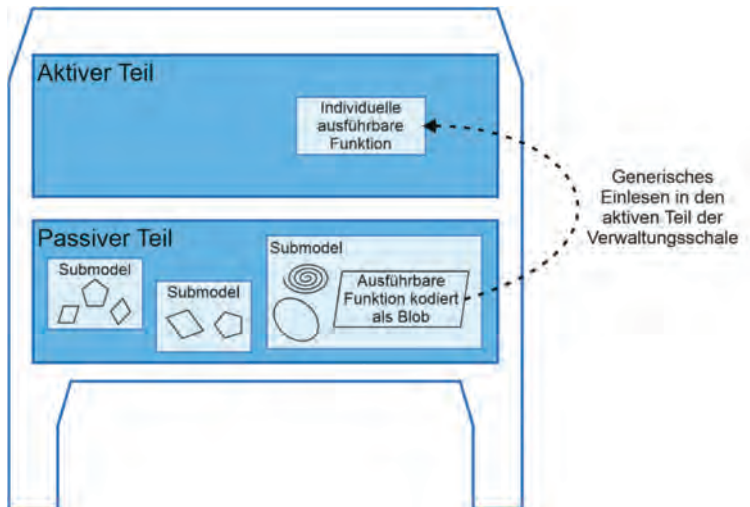


Bild 5: Generisches Einlesen einer individuellen Funktion als Teilmodellelement Blob

Die vorgestellte Service-Facade wird durch eine Main-Operation umgesetzt, die ebenfalls als Blob eingelesen wird. Diese steuert die Aktionen der Verwaltungsschale, indem sie Triggerpunkte überprüft, Capability- und Feasibility-Checks ausführt oder auf Aktionen anderer Verwaltungsschalen zugreift.

Ändern sich die Funktionalitäten der Verwaltungsschalen, müssen innerhalb der AasApp keine Re-Parametrierungen stattfinden. Ein Update der Blobs und ggf. der Input- und Output-Referenzen innerhalb der Verwaltungsschale genügt. Sämtliche Informationen über das Asset und die Funktionserfüllungen durch die AasApp liegen also stets in der Verwaltungsschale selbst. Mit der AasApp wurde eine Anwendung entwickelt, die es ermöglicht, aktive Verwaltungsschalen allein auf Basis der Informationen innerhalb der statischen Verwaltungsschalen zu erstellen und auszuführen. Es benötigt kein zusätzliches händisches Engineering zum Orchestrieren der Verwaltungsschalen, da die Funktionalitäten der Assets innerhalb der Verwaltungsschale liegen und die erzeugten Server selbstständig miteinander interagieren können.

6. Prototypische Umsetzung im Bereich der TGA

Zur Realisierung flexibler und autonomer Interaktionen gebäudetechnischer Assets ohne manuelles Engineering wurde eine Reihe von Funktionen und Diensten von Verwaltungs-

schalen prototypisch implementiert. Als verbreiteter Use Case der Gebäudetechnik wurde die Wärmeversorgung eines verbrauchenden Assets durch einen oder mehrere Wärmeerzeuger umgesetzt.

Verwaltungsschalen verbrauchender Assets beinhalten eine Trigger-Funktion, die überprüft, ob ein Bedarf an Wärme vorliegt und einen dadurch ausgelösten Capability-Checker, der überprüft, ob Assets in der Umgebung die abstrakte Fähigkeit „Heizen“ haben. Dafür werden die Semantic-IDs der angebotenen Capabilities mithilfe der referenzierten Ontologien abgeglichen. Im Anschluss werden ein oder mehrere Feasibility-Checker ausgelöst, die überprüfen, ob die möglichen Wärmeerzeuger die Bedingungen der spezifischen Wärmeanforderung erfüllen und welcher Erzeuger die besten Bedingungen bereitstellt, (Bild 6). Informationen über den Skill zur Wärmeerzeugung und dessen Attribute (z. B. Nennleistung) sind innerhalb der Verwaltungsschale referenziert und mit der Capability verlinkt.

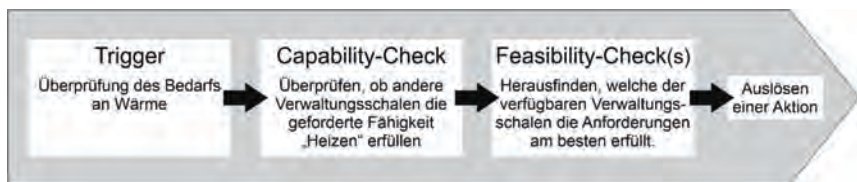


Bild 6: Schema der internen Dienste für den Funktionsablauf einer Wärmeanforderung

In den Feasibility-Checks werden die Properties der Wärmeanforderung mit den Informationen der Wärmeerzeuger für eine Machbarkeitsanalyse abgeglichen. Beispiele für Properties einer Wärmeanforderung sind notwendige Properties wie der *Versorgungsbereich* oder hinreichende Properties wie die benötigte *Wärmeleistung* oder die *Mindestvorlauftemperatur*. In mathematisch-physikalischen Auswertungen wird ermittelt, welche Assets die geforderten Bedingungen erfüllen können oder ob Assets eine Teilmenge der geforderten Bedingungen erfüllen können. Bei erfolgreichen Checks löst der Wärmeverbraucher die Erzeugung von Wärme beim Wärmeerzeuger mit den besten Bedingungen aus.

7. Zusammenfassung und Ausblick

Die entwickelte AasApp ermöglicht das generische Einlesen und das Ausführen von Funktionalitäten autonom handelnder Assets ohne individuelle Anpassungen und zusätzliches Engineering zur Systemintegration. Die Konzepte des fähigkeitsbasierten Engineerings werden genutzt, um ein Checker-Framework als Basis von Interaktionen aktiver Verwaltungsschalen

zu entwickeln. Anstatt mit starr vorgegebenen Automationsregeln eingestellt zu sein, interagieren die aktiven Verwaltungsschalen flexibel und können auf sich ändernde Umwelt- und Nutzungsbedingungen reagieren. Bei nachträglichen Änderungen im Gebäudeenergiesystem entfällt eine aufwändige Re-Parametrierung der Automationslösung. Der Aufwand des Engineerings wird deutlich reduziert.

Die Entwicklungen können Grundlage zukünftiger Automationslösungen in der TGA sein. Insbesondere ergeben sich Vorteile bei Gebäudeenergiesystemen mit mehr als einem Wärmeerzeuger oder bei der Integration verschiedener Gebäude und Speichersysteme innerhalb eines Smart Grids. Die Flexibilität von TGA-Anlagen ist jedoch auch hardwareseitig bedingt, z. B. durch installierte Rohrverläufe. Gebäudeenergiesysteme ändern sich in Bestandsgebäuden wegen des hardware- und softwareseitigen Aufwands selten. Durch die Nutzung selbsthandelnder Assets wird der softwareseitige Aufwand reduziert.

In anknüpfenden Forschungsarbeiten werden weitere Feasibility-Checks für den Betrieb von Assets der TGA entwickelt, insbesondere Checks, die wirtschaftliche Aspekte der Prozesse stärker berücksichtigen, um eine Grundlage für Anwendungen as a Service zu legen [19].

Das entwickelte Konzept einlesbarer Funktionen in Verwaltungsschalen lässt sich in allen industriellen Bereichen einsetzen, z. B. in der industriellen Fertigung. Es wird deshalb bei der Weiterentwicklung von Eclipse BaSyx berücksichtigt. In einem weiteren Schritt können die Ergebnisse der Arbeit auch als Grundlage dafür dienen, die aktiven Verwaltungsschalen in andere I4.0-Netze wie bspw. [20] einzugliedern.

8. Literaturangaben

- [1] D-2.11 Capability- und Feasibility-Interaktion. Diedrich, C., Miny, T. u. Schmitt, S., 2021
- [2] Details of the Administration Shell. Part 1 - The exchange of information between partners in the value chain of Industrie 4.0, Version 3.0. Bundesministerium für Wirtschaft und Energie (Hrsg.), Berlin 2020
- [3] Describing Capabilities of Industrie 4.0 Components. White Paper. Bayha, A., Bock, J., Boss, B., Diedrich, C. u. Malakuti, S. Bundesministerium für Wirtschaft und Energie (Hrsg.), Berlin 2020
- [4] VDI 3814:2019-01. *Gebäudeautomation (GA). Blatt 1: Grundlagen*
- [5] VDI 3814:2019-01. *Gebäudeautomation (GA). Blatt 2.1: Planung - Bedarfsplanung, Betreiberkonzept und Lastenheft*
- [6] Shell BDH Hauswärme-Studie. Klimaschutz im Wohnungssektor - Wie heizen wir morgen?, Fakten, Trends und Perspektiven für Heiztechniken bis 2030. Adolf, J., Schabla, U., Lücke, A., Breidenbach, L., Bräuninger, M., Leschus, L., Ehrlich, L., Otto, A., Oschatz, B. u. Mailach, B. Shell Deutschland Oil GmbH u. BDH Bundesindustrieverband Deutschland Haus-, Energie- und Umwelttechnik e.V. (Hrsg.), Hamburg, Köln 2013
- [7] Belyaev, A. u. Diedrich, C.: Aktive Verwaltungsschale von I4.0-Komponenten. Erscheinungsformen von Verwaltungsschalen. Automation 2019 - Autonomous Systems and 5G in Connected Industries (20. Leitkongress der Mess- und Automatisierungstechnik), S. 517–530
- [8] Motsch, W., Dorofeev, K., Gerber, K., Knoch, S., David, A. u. Ruskowski, M.: Concept for Modeling and Usage of Functionally Described Capabilities and Skills. 26th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA) (2021), S. 1–8
- [9] Industrie 4.0 Index. Deutscher Industrie 4.0 Index 2019, Eine Studie der Staufen AG und der Staufen Digital Neonex GmbH. STAUFEN.AG (Hrsg.), Köngen 2019
- [10] Hoang, X.-L. u. Fay, A.: A Capability Model for the Adaptation of Manufacturing Systems. 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA) (2019), S. 1053–1060
- [11] Herzog, J., Ropke, H. u. Luder, A.: Allocation of PPRS for the plant planning in the final automotive assembly. 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA) (2020), S. 813–820
- [12] Gruber, T.: Ontology. In: Liu, L. u. Özsu, M. T. (Hrsg.): Encyclopedia of Database Systems. New York, NY: Springer New York 2016, S. 1–3

- [13] Maisch, N.: CapOnto. Describing the development and usage of an ontology for I4.0-Capabilities of HVAC-Assets, Cologne 2021. <https://github.com/thcologne-gart/CapOnto>, abgerufen am: 12.01.2022
- [14] DIN EN IEC 81346-2:2020-10. *Industrielle Systeme, Anlagen und Ausrüstungen und Industrieprodukte - Strukturierungsprinzipien und Referenzkennzeichnung - Teil 2: Klassifizierung von Objekten und Kennbuchstaben für Klassen (IEC 81346-2:2019); Deutsche Fassung EN IEC 81346-2:2019*
- [15] Maisch, N., Both, M. u. Müller, J.: Interoperable Interaktionen semantisch heterogener Industrie 4.0 Komponenten auf Basis einer Capability Ontologie für die Technische Gebäudeausrüstung. Automation 2021 - Navigating towards resilient Production (22. Leitkongress der Mess- und Automatisierungstechnik) (2021), S. 691–704
- [16] Fraunhofer-Institut für Experimentelles Software Engineering IESE: BaSyx. <https://www.eclipse.org/basyx/>, abgerufen am: 25.04.2022
- [17] Hoffmeister, M., Orzelski, A., Mendes, M., Grote, J., Göllner, D. u. Bischof, S.: AASX Package Explorer. Festo AG & Co. KG; PHOENIX CONTACT GmbH & Co. KG; Fraunhofer IOSB-INA Lemgo; Schneider Electric Automation GmbH; SICK AG; KEB Automation KG; Lenze SE 2021
- [18] Ziesche, C.: BaSyx Asset Administration Shell Repository HTTP REST-API. v1. The full description of the generic BaSyx Asset Administration Shell Repository HTTP REST-API. https://app.swaggerhub.com/apis/BaSyx/basyx_asset_administration_shell_repository_http_rest_api/v1#/, abgerufen am: 18.03.2022
- [19] Bundesministerium für Wirtschaft und Klimaschutz: Everything as a Service (XaaS), Berlin. <https://www.plattform-i40.de/IP/Redaktion/DE/Standardartikel/Themen-und-Technologiekatalog/everything-as-a-service.html>, abgerufen am: 03.05.2022
- [20] Diedrich, C.: VWS VERNETZT. Projekt "Verwaltungsschale vernetzt", Magdeburg. <https://vwsvernetzt.de/>, abgerufen am: 20.04.2022



VDE

Mess- und
Automatisierungstechnik



AUTOMATION

ISBN 978-3-18-092399-4, E-ISBN 978-3-18-102399-0

<https://doi.org/10.51203/9783181023990> - Generiert durch IP 216.73.216.36, am 20.01.2026, 12:19:01. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts in Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.