Der Beitrag untersucht, inwieweit etablierte maschinelle Verfahren bei der inhaltlichen Erschließung von Forschungsdaten sinnvoll zur Anwendung kommen können. Dabei wird zunächst ein Blick auf die unterschiedliche Erschließungspraxis der Deutschen Nationalbibliothek (DNB), des DIPF | Leibniz-Instituts für Bildungsforschung und Bildungsinformation und der Leibniz-Informationszentrum Technik und Naturwissenschaften Universitätsbibliothek (TIB) geworfen. Daran schließt sich ein Bericht über Tests an, bei denen Forschungsdaten von DIPF und TIB mit den bei der DNB angewandten Methoden maschinell erschlossen wurden. Die Bewertung dieser Tests führt vor Augen, dass die Anwendung maschineller Erschließungsverfahren auf Forschungsdaten im Bereich des Möglichen liegt, aber mit noch größeren Schwierigkeiten als bei der Anwendung auf konventionelle Medienwerke verbunden ist.

This article examines the extent to which established machine-based techniques can usefully be applied to the content indexing of research data. It first considers the different cataloguing practices of the German National Library (DNB), the DIPF | Leibniz Institute for Research and Information in Education and the Leibniz Information Centre for Science and Technology and University Library (TIB). This is followed by a report on tests in which research data from the DIPF and TIB were machine-indexed using the methods deployed by the DNB. The evaluation of these tests demonstrates that it is fundamentally possible to apply machine indexing methods to research data. However, their use for this purpose is more problematical than when they are used on conventional media works.

Maschinelle Erschließung von Forschungsdaten in Bibliotheken und Forschungsdatenzentren – Stand und Perspektiven

In den vergangenen Jahren ist die Notwendigkeit von gut katalogisierten und kuratierten Forschungsdaten sowohl von der wissenschaftlichen Community als auch von der politischen Entscheidungsebene erkannt worden. Davon zeugen nicht nur Netzwerke wie die Nationale Forschungsdateninfrastruktur (NFDI) oder die seit kurzem vom BMBF geförderten Datenkompetenzzentren,[1] sondern auch die verstärkte Aufbereitung und Bereitstellung von Forschungsdaten durch Bibliotheken und Forschungsdatenzentren. Aufgrund der enormen Menge an Forschungsdaten, die von der Wissenschaft produziert und nachgenutzt wird, ist diese Aufgabe jedoch mit einigen Schwierigkeiten verbunden, sodass eine Automatisierung durch maschinelle bzw. KI-gestützte Verfahren wünschenswert erscheint. Im Folgenden wird zunächst die traditionelle Medienerschließung an der Deutschen Nationalbibliothek (DNB) skizziert, um dann anhand zweier Beispiele aufzuzeigen, wie sich insbesondere die hier entwickelten maschinellen Verfahren möglicherweise auf Forschungsdaten übertragen ließen. Außerdem werden Einblicke in die aktuellen Erschließungspraktiken des DIPF | Leibniz-Instituts für Bildungsforschung und Bildungsinformation sowie des Leibniz-Informationszentrums Technik und Naturwissenschaften Universitätsbibliothek (TIB) gegeben und gezeigt, welche Chancen und Schwierigkeiten sich dort mit der Automatisierung bestimmter Erschließungspraktiken verbinden.

Die Erschließungspraxis der Deutschen Nationalbibliothek und ihre Konsequenzen für den Umgang mit Forschungsdaten

Zu den zentralen Dienstleistungen, die von Bibliotheken angeboten werden, zählt die inhaltliche Erschließung von Dokumenten, die Nutzer*innen auf vielfältige Weise bei der Literatursuche unterstützt. Dabei stehen drei Funktionen im Vordergrund: Die Inhaltserschließung sorgt zunächst für eine Förderung des Information Retrievals durch die Optimierung der Auffindbarkeit der zu einem bestimmten Thema vorhandenen Literatur. Darüber hinaus erleichtert sie den Umgang mit einer konkreten Ressource durch die Anfügung einer Zusammenfassung bzw. thematischen Einordnung des Dokumenteninhalts am Titeldatensatz. Ein weiterer wichtiger, oft übersehener Effekt der Inhaltserschließung ist die Wissensexploration, die Ermöglichung der Entdeckung von Nützlichem durch Navigieren und Browsen ohne vorher festgelegte Zielsetzung.[2]

Die inhaltliche Erschließung von Dokumenten kann auf unterschiedliche Weise erfolgen und schließt eine Vielzahl von Methoden ein. Hierzu zählen nicht nur die etablierten Formen der klassifikatorischen und verbalen

Erschließung, sondern auch aktuell weniger verbreitete Verfahren wie das Anfügen von Annotationen und Abstracts.[3] Auch die Verknüpfung von Inhaltsverzeichnissen mit den Metadaten eines Dokuments kann zur Beförderung der von der Inhaltserschließung anvisierten Ziele beitragen. Die verbreitetsten Formen der Inhaltserschließung im deutschsprachigen Bibliothekswesen sind jedoch die systematische Erschließung durch die Anwendung von Klassifikationssystemen sowie die verbale Erschließung mit Schlagwörtern.

Besonders die systematische bzw. klassifikatorische Erschließung, bei welcher Dokumente in ein meist hierarchisch gegliedertes System der Wissensorganisation eingeordnet werden, hat eine lange Tradition und prägte das deutsche Bibliothekswesen als dominantes Verfahren bis in die 1970er-Jahre. Neben ihrer Nützlichkeit im Hinblick auf die Bildung von Signaturen ist es besonders die durch Klassifikationen gegebene Möglichkeit, beim Retrieval größere Dokumentenmengen zusammenzufassen, die ihren Einsatz gerade im Hinblick auf das Ziel der Wissensexploration sinnvoll erscheinen lässt. Charakteristisch für den Einsatz von Klassifikationen in Bibliotheken des deutschen Sprachraums ist allerdings ein hohes Maß an Heterogenität. Seit den 1960er-Jahren machen sich jedoch auch hier Standardisierungstendenzen bemerkbar, und seit etwa zwanzig Jahren nehmen hausübergreifende Systeme wie die Dewey-Dezimalklassifikation (DDC), die Regensburger Verbundklassifikation (RVK) und die Basisklassifikation (BK) eine immer wichtigere Rolle ein.[4]

Seit den 1970er-Jahren hat zusätzlich die verbale Erschließung immer mehr an Bedeutung gewonnen. Anders als beim Einsatz von Klassifikationen steht bei dieser Form der Erschließung nicht die Generierung großer Treffermengen, sondern die möglichst präzise, passgenaue Bestimmung von Dokumenten im Vordergrund des Interesses. Im Gegensatz zur klassifikatorischen Erschließung wurde hier bereits ein sehr hohes Maß an Standardisierung erreicht. Die Erschließung mittels Schlagwörtern wird im ganzen deutschen Sprachgebiet praktiziert. Sie erfolgt heute durchgehend mit den Schlagwörtern der Gemeinsamen Normdatei (GND) und unter Rückgriff auf zwei etablierte Regelwerke, die Regeln für den Schlagwortkatalog (RSWK) und Resource Description and Access (RDA).[5]

Die DNB wendet seit vielen Jahren Verfahren der verbalen und der klassifikatorischen Erschließung an und publiziert die dabei entstandenen Erschließungsdaten in der Deutschen Nationalbibliografie. Sie beteiligt sich außerdem intensiv an der Weiterentwicklung der entsprechenden Regelwerke. Die verbale Erschließung wird mit dem GND-Vokabular durchgeführt; als Klassifikation findet die DDC Verwendung.

Im Laufe der Jahre hat die DNB ihre Erschließungskonzepte mehrfach den Erfordernissen der sich wandelnden Zeitumstände anpassen müssen. Einen Einschnitt stellte hier die Novellierung des Gesetzes über die Deutsche Nationalbibliothek im Jahr 2006 dar, die eine Ausweitung des gesetzlichen Sammelauftrags auf unkörperliche Medienwerke, d. h. online im Internet zugängliche Veröffentlichungen, beinhaltete.[6]

Die sehr dynamische Entwicklung des digitalen Publikationsmarkts und die daraus resultierende Vielzahl an digitalen Objekten erzwangen in der Folge eine Veränderung der Erschließungspraxis. Eine intellektuelle Erschließung aller von der DNB gesammelter Ressourcen erschien mit den vorhandenen Kapazitäten nicht mehr vorstellbar. Um dennoch weiterhin eine umfassende Inhaltserschließung sicherstellen zu können, setzte die DNB fortan auf den verstärkten Einsatz maschineller Verfahren. Primär betraf dies die unkörperlichen Medienwerke, deren intellektuelle Erschließung ab 2010 eingestellt wurde.[7] Die Katalogisate für diese Publikationen wurden fortan maschinell erstellt. Auch die Erschließung der körperlichen Medienwerke erfuhr seitdem einige Modifikationen. In der Konsequenz kam es zu einer Reduktion der intellektuellen Erschließungsleistung, die erforderlich war, um Ressourcen in die Entwicklung maschineller Erschließungsverfahren umzulenken.[8]

Forschungsdaten sind derzeit nicht Bestandteil des Sammelauftrags der Deutschen Nationalbibliothek.[9] Sollte es zukünftig hier zu einer Veränderung kommen, kann von einem Einsatz maschineller Erschließungsverfahren auch für diese Datenart ausgegangen werden. Deshalb ist der Austausch mit anderen Institutionen, die bereits solche Verfahren einsetzen, umso wichtiger.

Forschungsdaten Bildung am DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation

Die Erschließung und Archivierung von Forschungsdaten der Bildungsforschung gehört zu den Kernaufgaben des Teams Forschungsdaten Bildung am DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation. Zentrale Bedeutung kommt dabei dem Verbund Forschungsdaten Bildung (VerbundFDB) und dem FDZ Bildung zu. Der VerbundFDB stellt Angebote und Services für die empirische Bildungsforschung in den Bereichen Beratung, Datenarchivierung und Datenbereitstellung zur Verfügung. Damit trägt er dazu bei, das Potenzial der in der Bildungsforschung generierten Forschungsdaten möglichst umfassend auszuschöpfen und eine Kultur des Data Sharing zu etablieren. Der Workflow der verteilten Archivierung ermöglicht es, Forschungsdaten zentral an den VerbundFDB zu melden und zu übermitteln, der dies dann wiederum an eines der Forschungsdatenzentren (FDZ) mit der entsprechenden Fachexpertise übergibt, um sie dort zu erschließen und zu veröffentlichen.[10] Mit dem FDZ Bildung ist ein Partner-FDZ des VerbundFDB ebenfalls am DIPF verortet. Es stellt unter anderem gering standardisierte Forschungsdaten, wie etwa Unterrichtsvideos, Transkripte und Beobachtungsprotokolle, bereit. Im Zusammenspiel

mit dem VerbundFDB bietet das FDZ Bildung somit Services, Infrastruktur, Informationen und Beratung zu forschungs- und forschungsdatenrelevanten Themen an, die von der Projekt- und Antragsplanung über rechtliche Themen bis hin zur Nachnutzung von Forschungsdaten reichen.

Gering standardisierte Forschungsdaten stellen im Erschließungsprozess eine besondere Herausforderung dar. »Dies liegt einerseits an der Verwobenheit der Erhebungs-, Aufbereitungs-, Analyse- und Interpretationsschritte qualitativer Forschung und andererseits an dem iterativen und alternierenden Erkenntnisprozess (…), der mit einer fortlaufenden Kontextualisierung der erhobenen Daten einhergeht. Zudem muss den vielfältigen Datensorten (z. B. Bilder, Videos, Interviews, Beobachtungen, Dokumente) und Vorgehensweisen der Datenerhebung sowie Auswertung in den verschiedenen Feldern Rechnung getragen werden (…).«[11]

Aufgrund dieser Komplexität und der besonderen datenschutzrechtlichen und forschungsethischen Anforderungen der Forschungsdaten erfolgt ihre Erschließung am DIPF durchweg intellektuell. Der damit verbundene hohe Arbeitsaufwand konkurriert mit dem Wunsch, Forschungsdaten ohne lange Verzögerungen bereitzustellen. An diese Ausgangslage schließt sich die Frage an, ob Methoden der maschinellen Erschließung den Prozess der Bereitstellung der Forschungsdaten unterstützen können. Hierzu werden beispielhaft drei mögliche Interventionspunkte des Forschungsdaten-Workflows betrachtet:

- die Anonymisierung textueller Daten,
- das Abstracting von Transkripten,
- die Erschließung von Publikationen, entstanden aus Forschungsdaten.

Dabei sollen sowohl Möglichkeiten als auch Schwierigkeiten und Hürden, die mit einer solchen Automatisierung verbunden sind, betrachtet werden.

Die Einhaltung von Datenschutzbestimmungen und die Wahrung der Persönlichkeitsrechte von Untersuchungspersonen sind wesentliche Voraussetzungen für die Archivierung und Nachnutzung personenbezogener Forschungsdaten. Insbesondere textuelle Forschungsdaten wie Interviewtransskripte erfordern daher einen gewissen Grad der Anonymisierung, um sie bereitstellen zu können.

Der Ausdruck »Anonymisierung« bezieht sich auf den Vorgang, bei dem aus Forschungsdaten Informationen entfernt werden, die eine Identifizierung einer Person ermöglichen. Hierbei gibt es unterschiedliche Grade der Anonymisierung. Die formale Anonymisierung beinhaltet das Entfernen unmittelbarer Identifizierungsmerkmale (z. B. Personen- und Ortsnamen, Bilder, Stimmen) aus den Daten. Die faktische Anonymisierung besteht darin, die Daten so zu modifizieren, dass »die Person nur mit einem völlig unverhältnismäßigen Aufwand reidentifiziert werden kann«.[12] Absolut anonymisierte Daten weisen eine solche Veränderung auf, dass es unmöglich ist, eine Person zu reidentifizieren.[13]

Eine klassische Methode der Verfremdung von Textdaten stellt die Pseudonymisierung dar. Anstatt Informationen vollständig zu entfernen, können diese durch pseudonyme Daten ersetzt werden. Dies bedeutet, dass persönliche Daten durch fiktive, aber konsistente Ersatzwerte substituiert werden, die keine direkte Identifizierung der Person ermöglichen, aber eine Nachverfolgbarkeit innerhalb des Datensatzes erhalten.[14] Dieser Prozess könnte mit Named Entity Recognition (NER)-Modellen automatisiert werden. Ein Beispiel hierfür ist das Tool NETANOS (Named Entity-based Text Anonymization for Open Science). Diese Software kann benannte Entitäten wie Namen, Orte, Daten und Organisationen in unstrukturierten Texteingaben identifizieren und ändern.[15]

Es besteht dabei jedoch die Möglichkeit, dass kontextabhängige Zusammenhänge nicht vollständig erfasst werden, was zu ungewolltem und unnötigem Informationsverlust führen kann. Um zu gewährleisten, dass die Anonymisierung die Nützlichkeit der Daten für analytische Zwecke beibehält, könnte NER mit Large-Language-Modellen wie GPT (Generative Pre-trained Transformer) kombiniert werden, die den Kontext analysieren und konsistente semantisch äquivalente Ersatztexte generieren.

Die Effektivität einer KI-gestützten Anonymisierung hängt aber von der Qualität und Vielfalt der Trainingsdaten ab. Die Modelle sollten zudem nicht voreingenommen sein und die Privatsphäre aller Individuen gleichermaßen schützen. Deshalb sind auch robuste Methoden zur Evaluierung der Anonymisierung notwendig, um sicherzustellen, dass die Daten adäquat anonymisiert sind und das Risiko der Reidentifikation minimiert wird, denn schon ein einziges Wort könnte ausreichen, um ein ganzes Transkript zu anonymisieren oder umgekehrt die betreffenden Personen zu identifizieren.

Das FDZ Bildung sieht bei der Bereitstellung von Daten der Bildungsforschung Abstracts auf allen Ebenen seines Metadatenschemas vor, angefangen auf der Ebene des Gesamtprojekts über die (Teil-)Studien und Erhebungswellen bis hin zu den einzelnen Dokumenten des Korpus. Da für die höheren Ebenen auf Projektwebseiten oder in Publikationen zumeist Abstracts aus der Hand der Forscher*innen selbst vorliegen, soll hier das Abstracting des einzelnen Forschungsdatums erörtert werden. Auf dieser sehr tiefen Erschließungsebene werden Abstracts im engeren Sinne nur für audiovisuelle Forschungsdaten wie etwa Unterrichtsvideos verfasst. Aufgrund ihrer Größe, der Übertragungszeit und des erforderten Speicherplatzes ist es für Forscher*innen besonders hilfreich, schon vor dem Download Einblick in die Videos zu erhalten, um einschätzen zu können, ob sie für die eigene Forschung infrage kommen. Für textförmige Daten hingegen gilt dies nicht. Diese werden

wegen des hohen Aufwands nur mit kurzen allgemeinen Beschreibungen zur Situation und zu den Beteiligten angereichert; zudem können sie auf der Webseite des FDZ Bildung mit einer Volltextsuche durchsucht werden.[16] Dennoch wäre es äußerst gewinnbringend, den gesamten Bestand textförmiger Forschungsdaten mit maschinell generierten Abstracts zu versehen, ohne diese nach Maßgabe des erforderten intellektuellen Aufwands gestalten zu müssen – sowohl um die Suchmöglichkeiten zu verbessern, als auch um Forscher*innen die Einschätzung des Materials zu erleichtern.

Wie genau die KI-generierte inhaltliche Verdichtung der Abstracts idealerweise aussehen sollte, ist zunächst bedingt durch ihre pragmatische Position im Ablauf von Forschungsprozessen: Die Abstracts sollen nicht etwa die Lektüre eines Textes ersetzen. Die Zusammenfassungen sollen vielmehr Evidenzen dafür bieten, ob die genauere Analyse des empirischen Materials im Licht der eigenen Forschungsfrage aufschlussreich zu sein verspricht oder nicht. Das Ziel der maschinellen Erschließung müssten deshalb indikative Abstracts sein, die einen »möglichst maßstabsgetreu[en]« und »globalen thematischen Überblick«[17] über Texte bieten, die, wie etwa ein Unterrichtstransskript, die verschiedensten Formen sprachlicher Handlungen zahlreicher Akteur*innen ebenso umfassen können wie lange Erzählungen nur eines Sprechers oder einer Sprecherin, der oder die ein narrativ-biografisches Interview gibt. Damit im Zusammenhang steht ein weiteres wichtiges Kriterium: die Objektivität der Abstracts. Anders als die Texte selbst sollen ihre Abstracts von Werturteilen nur in der dritten Person handeln, selbst aber keine fällen.

Die Heterogenität der Texte ist ein weiterer Aspekt, der bei der maschinellen Generierung von Abstracts berücksichtigt werden muss: Die vom FDZ Bildung bereitgestellten Texte protokollieren nicht nur ein breites Spektrum von Kommunikationsformen und -situationen, sie variieren auch hinsichtlich ihres Umfangs und des verwendeten Transkriptionssystems teilweise beträchtlich. Während manche Texte nur vier Seiten umfassen, sind andere 50 Seiten lang; und während manche der Transkripte eine Angleichung an die Schriftsprache vornehmen, protokollieren andere mit verschiedenen Notationskonventionen und Präzisionsgraden mehr oder weniger exakt das gesprochene Wort sowie außer- oder parasprachliche Handlungen.

Für ein Training des Modells könnten deshalb verschiedene Klassen von Texten gebildet werden, wie etwa Interaktionsprotokolle, Beobachtungsprotokolle und Interviewtransskripte. Die Uneinheitlichkeit der Notationskonventionen und des Präzisionsgrades kann hierdurch aber nicht behoben werden. Auch könnten für das Training all jene Texte exkludiert werden, die einen bestimmten Umfang überschreiten. Der Richtwert für die Länge der zu generierenden Abstracts liegt gemäß den Konventionen des FDZ Bildung unabhängig vom Input bei ca. 10–20 Zeilen. Wie schon im Falle der Anonymisierung ist hierfür der Einsatz eines Large-Language-Modells wie ChatGPT denkbar, dessen Einsatz bei der Erstellung wissenschaftlicher Abstracts schon evaluiert wurde.[18]

Die Bereitstellung von Projektliteratur ist eine Aufgabe des VerbundFDB, die in enger Zusammenarbeit mit dem FDZ Bildung wahrgenommen wird. Sie ist von herausragender Bedeutung, da viele Forscher*innen ihre Suche nach nachnutzbaren Forschungsdaten mit einer Literaturrecherche beginnen.[19] Darüber hinaus können sie mit ihrer Hilfe besser beurteilen, ob ein Datensatz für die eigene Forschungsfrage geeignet ist (data fit).[20]

Als Projektliteratur werden Publikationen bezeichnet, die im Rahmen eines Forschungsprojektes auf Grundlage der erhobenen Forschungsdaten entstanden sind.[21] Forschungsprojekte sollen bei Übergabe der Forschungsdaten auch die daraus entstandenen Publikationen an den VerbundFDB melden. Aufgabe des VerbundFDB ist es dann, diese Publikationen mit den Forschungsdaten zu verknüpfen. Dies geschieht mithilfe einer Publikationsmeldung des Forschungsdatenteams an das Fachportal Pädagogik am DIPF.[22] Die Verknüpfung zwischen Forschungsdaten und Projektpublikationen erfolgt anhand einer ID, welche der Service FIS Bildung[23] als Teil des Fachportals Pädagogik jeder einzelnen Publikation zuweist, die dort vom Forschungsdatenteam neu gemeldet wird. Somit sind die Forschungsdaten und Projektpublikationen mit der FIS Bildung Literaturdatenbank und dem Webangebot des VerbundFDB reziprok verknüpft und für die Nutzenden langfristig auffindbar.

Der etablierte Workflow sieht gerade für die Meldung der Publikationen verschiedene Arbeitsschritte vor, die ein hohes Maß an intellektueller Arbeit erfordern. Der größte Teil dieses Workflows besteht aus der Literaturrecherche zur Verifikation der gemeldeten Projektpublikationen. Die gemeldeten Publikationen werden anhand zuverlässiger und valider Quellen nachrecherchiert und mit den nötigen bibliografischen Angaben angereichert, damit sie in die FIS Bildung-Datenbank aufgenommen werden können. Dabei wird überprüft, ob die Literaturangaben von den Projekten vollständig zitiert wurden und ob die gemeldeten Publikationen tatsächlich innerhalb der Projekte oder mit einem Bezug zu den Forschungsdaten entstanden sind.

Als valide Quellen werden zum einen Verlagswebseiten und zum anderen unterschiedliche Online-Datenbanken und -kataloge herangezogen. Je nach Dokumententyp wird in den Online-Katalogen der Zeitschriftendatenbank,[24] der Deutschen Nationalbibliothek[25] und der OGND (Gemeinsame Normdatei online)[26] recherchiert. Die auf diesem Wege ermittelten bibliografischen Angaben müssen an eine von FIS Bildung vorgegebene Syntax angepasst werden, damit sie dem Qualitätsanspruch zur Aufnahme in deren Datenbank entsprechen und ohne er-

neute Überprüfung in diese Datenbank eingepflegt werden können. Die benötigten Metadaten werden mittels des Literaturverwaltungsprogramms Citavi erfasst und dann quartalsweise als angepasste BibTEX-Datei in FIS Bildung exportiert. In der Komplexität des Metadatensets sowie den unterschiedlichen Quellen für die Nachrecherche der Publikationsmeldungen liegen die beiden großen Herausforderungen für die Erschließung der Projektpublikationen. Denn zum einen sind die bibliografischen Angaben, die von den Projekten gemeldet werden, mitunter unvollständig, unstrukturiert und in verschiedenen Zitationsstilen verfasst. Dies betrifft sowohl einzelne Meldungen als auch umfangreiche Publikationslisten. Zum anderen sind die gemeldeten Dokumententypen sehr heterogen, was die Nachrecherche in den entsprechenden Onlinekatalogen zeitaufwendig macht.

Um die Frage zu beantworten, ob ein maschinelles Verfahren die komplexen Arbeitsgänge vereinfachen könnte, soll ein Blick auf ein aktuelles Exportprojekt geworfen werden. Es umfasst insgesamt 183 gemeldete Publikationen.[27] Ein großer Teil davon zählt zur »grauen Literatur«, wie z. B. technische Reports und Konferenzschriften. Schon dieses Faktum erschwert den möglichen Einsatz von KI-gestützten Verfahren, denn mit den Technical Reports wird ein wichtiger Teilbereich der grauen Literatur, der für die Kontextualisierung von Forschungsdaten essentiell ist, von Bibliotheken nicht oder selten strukturiert erfasst und bereitgestellt. Für die Anwendung eines lernenden maschinellen Verfahrens sind damit nicht genügend Daten vorhanden, um die automatische Indexierung dieses Dokumententyps zu trainieren. Zudem müsste eine automatisierte Anwendung in der Lage sein, bei der Recherche und Verifikation der Projektpublikationen auf die oben genannten unterschiedlichen Quellen zurückzugreifen.

Da aufgrund der heterogenen Quellen sowie der niedrigen Wiederholungsrate der Recherchetätigkeit an den kritischen Stellen im Erschließungsworkflow kein oder ein nur sehr geringes Automatisierungspotenzial vorhanden ist, könnten in näherer Zukunft wohl eher Human-in-the-Loop-Prozesse bei der Optimierung des Workflows eingesetzt werden. Hierfür werden derzeit unter anderem Tools wie ANNIF,[28] Malibu[29] oder IRIS.AI[30] für den Erschließungsworkflow getestet. Eine intellektuelle Nachkontrolle der durch diese Tools erschlossenen Einträge ist allerdings derzeit noch unerlässlich.

An allen drei vorgestellten Interventionspunkten konnten Automatisierungspotenziale identifiziert werden. Dabei wurden auch die komplexen Anforderungen, die an entsprechende Lösungen zu stellen sind, beleuchtet.

Zudem müssen KI-Tools selbst den gesetzlichen Vorgaben (etwa der DSGVO) entsprechen – ebenso wie auch die mit diesen Tools durchgeführte Datenverarbeitung. Sorgfältige Validierungen und Evaluierungen sind demnach erforderlich, um sicherzustellen, dass Anonymisierung und Abstracting den Datenschutzanforderungen entsprechen und keine Risiken der Reidentifikation bestehen. Komplexere Sprachmodelle etwa, die mit solch sensiblen Daten arbeiten, müssen daher lokal ausgeführt werden, um eine Diffusion der datenschutzrelevanten Parameter in die Cloud zu verhindern.

Zugleich konnte gezeigt werden, wie komplex die Datenbasis, etwa bei der Erschließung von Projektpublikationen oder der Generierung von Abstracts, ist. Beispielsweise kann die Bearbeitung eines einzigen Begriffs die effektivste Form der Anonymisierung darstellen. Dies führt zu der Einschätzung, dass eine kontext- und komplexitätssensitive Automatisierung der exemplarisch genannten Arbeitsschritte durch Sprachmodelle wie GPT, wenn überhaupt, nur mit einem hohen Trainingsaufwand erreicht werden kann. Hierfür aber fehlen in der Bildungsforschung derzeit noch die Trainingsdaten, die zudem aufgrund der komplexen und verschachtelten Strukturen (Projekt – Studie – Datenkollektion bis hinunter zu einer einzelnen Erhebungseinheit wie einem Interviewtransskript) nicht mit dem Datenpool einer Bibliothek zu vergleichen sind.

Eine Alternative kann der Einsatz bereits existierender Tools an Einzelpunkten des Workflows sein, wie etwa NETANOS. Deren Ergebnisse – oder besser Vorschläge – sollten dann aber immer wieder in den schon genannten Human-in-the-Loop-Prozessen durch Fachleute überprüft und akzeptiert werden, um die effektive und rechtlich sichere Erschließung von Forschungsdaten sicherzustellen. Es bleibt die Frage, ob eine maschinelle Methode in der Lage ist, eine solche Effektivität zu erreichen.

Status und Herausforderungen der inhaltlichen Erschließung von MINT-Daten an der TIB

Naturwissenschaftliche und technische Forschungsdaten[31] liegen in einer Vielzahl von Formen vor: Bilddaten wie z. B. Satellitenbilder in den Erdsystemwissenschaften, Spektren, Plots, Herbarien, Kartierungen, Forschungssoftware, wissenschaftliche Artbeschreibungen, chemische Strukturen wie beispielsweise Röntgenstrukturen, Kataloge astronomischer Objekte u. v. m. Eine häufige Form sind numerische (Fakten-)Daten. Eine Gemeinsamkeit dieser Forschungsdaten ist, dass sich längere Fließtexte eher in den Publikationen und gegebenenfalls weiteren Beschreibungen der Daten finden. Bei einer solchen Beschreibung handelt es sich allerdings um ein neues, von den eigentlichen Daten verschiedenes Objekt, aus welchem ggf. Schlagwörter extrahiert werden können oder welches anderweitig erschlossen werden kann. Während für Texte eine nachträgliche Erschließung diesen Kontext herstellen kann, muss er für MINT-Daten[32] näher am Forschungsprozess und mit nachvollziehbarer Provenienz dokumentiert werden.

Metadaten für MINT-Forschungsdaten gliedern sich grob in drei Schichten: bibliografische Metadaten, inhaltliche Metadaten (z. B. Schlagwörter oder Klassifikationen) und formatspezifische technische Metadaten. Um

beispielsweise numerische Daten inhaltlich zu erschließen, müssen diese interpretierbar sein – ganz gleich, ob es sich bei der erschließenden Entität um einen Menschen oder um eine Software handelt. Ein beschreibender Text stellt eine Interpretationshilfe (einen »Beipackzettel«) für den Menschen dar.

Aktuell findet eine inhaltliche Erschließung der meisten MINT-Fächer typischerweise in Form einer Vergabe von formalisierten, zum Teil auch fachspezifischen Metadaten und weiterhin meist freien Schlagwörtern statt, welche zur Durchsuchbarkeit von Forschungsdatenrepositorien verwendet werden. Diese Erschließung ist nicht notwendigerweise deckungsgleich mit einer Erschließung im bibliothekarischen Sinne: In Bibliothekskatalogen sind diese Forschungsdaten bislang nur selten nachgewiesen.[33] Insbesondere herrscht eine große Heterogenität bezüglich der Datenerschließung mit kontrollierten, d. h. standardisierten Vokabularen und Metadatenschemata, je nach Fachgebiet, Engagement der Fachcommunity und Ausrichtung bzw. Betriebsmodell der jeweiligen Forschungsdatenrepositorien.

Die TIB betreut in ihrer Funktion als Universitätsbibliothek für die Leibniz-Universität Hannover (LUH) das institutionelle Repositorium für Textpublikationen der LUH. Ab August 2024 besteht die Möglichkeit, Forschungsdaten von Hochschulschriften als separate Publikation zu veröffentlichen. Für Datenpublikationen von LUH-Angehörigen steht das durch die Leibniz Universität IT Services betriebene Datenrepositorium[34] bereit. Zu den Aufgaben der TIB gehört dabei auch die Inhaltserschließung: Seit Ende 2023 werden Hochschulschriften durch Mitarbeiter*innen im Team »Hochschulschriften und Geschenke« im Rahmen der Titelaufnahmen mit einem Merkmal »enthält Forschungsdaten« ausgezeichnet, wenn die Forschungsdaten im Anhang der Dissertation enthalten sind oder wenn eine zusätzliche Datei hochgeladen wurde. Damit ist ein erster Schritt zur Inhaltserschließung von Forschungsdaten getan. Forschungsdaten, die im Fließtext der Arbeit integriert sind, können so aber nicht systematisch identifiziert werden.

Auf der Ebene der üblichen bibliografischen und inhaltlichen Metadaten können erprobte Werkzeuge der (halb-)automatischen Erschließung nachgenutzt werden: So erprobt die Universitätsbibliothek Stuttgart die Anwendung des »Digitalen Assistenten« auf Titelaufnahmen von Forschungsdaten, mit Schwerpunkt auf Natur- und Ingenieurwissenschaften. Dabei stellen die freien Schlagwörter eine Herausforderung dar.[35]

Anforderungen an die (automatisierte) Inhaltserschließung naturwissenschaftlicher und technischer Forschungsdaten ergeben sich aus den Anwendungsfällen für ihre Nachnutzung: Forscher*innen müssen abwägen, ob sie für ihre Forschungsfrage einen Datensatz nachnutzen können oder eine eigene Messung bzw. Datenerhebung anstrengen müssen. Einige Fachcommunitys der MINT-Disziplinen stellen hierfür sowohl Kriterien und Leitlinien zur Datenpublikation (z. B. ein fachspezifisches Metadatenschema[36]) als auch zur Nachnutzung (Fachspezifische Interpretation der FAIR Data-Kriterien[37]) bereit.

Eine hohe Qualität der Metadaten ermöglicht, die Qualität der Daten besser zu bewerten. Letztere kann sowohl Aspekte einer absoluten (z. B. Auflösung eines Spektrums) wie auch einer »subjektiven« Qualität (bezogen auf die Passung zwischen der ursprünglichen und der nachnutzenden Forschungsfrage) beinhalten. Für letztere muss der wissenschaftliche Kontext hinreichend genau dargestellt sein. Dies findet typischerweise im Abstract eines Artikels statt. Hier bergen maschinelle Methoden der Extraktion von Kontextinformationen aus Volltexten Chancen für die Weiterentwicklung der Inhaltserschließung. Solche Verfahren sind aber auf das Vorhandensein von Verknüpfungen durch persistente Identifikatoren zwischen Forschungsdaten und wissenschaftlichen Artikeln angewiesen, welche in den Repositorien aber nicht in allen Fällen vorliegen. Die zunehmende Verfügbarkeit kontextueller Metadaten, z. B. in Wissensgraphen wie dem ORKG,[38] ermöglicht eine Bewertung eines Datensatzes auch als alleinstehend. Zunehmend werden Datensätze aus Naturwissenschaft und Technik in Datenjournalen[39] bzw. Softwarejournalen[40] für Forschungssoftware veröffentlicht.[41] Dieser Publikationsweg ermöglicht eine weitergehende Inhaltserschließung.

Einige fachspezifische Datenformate können prinzipiell bereits technische Metadaten mitliefern, z. B. netCDF (Geowissenschaften), HDF5 (verschiedene physikalische Wissenschaften) und FITS (Astrophysik). Da diese Formate nur in bestimmten Teildisziplinen zum Einsatz kommen, stellen sie keine generische Lösung[42] dar, können aber wertvolle Ansätze zum Definieren von Metadaten-Workflows liefern. Für größere numerische Datensätze sollten einerseits die statistische Analyse und Qualitätssicherung und andererseits die Inhaltserschließung und Gewinnung von Metadaten verschmelzen. FAIR-kompatible Forschungssoftware sollte beides leisten können.

Durch die NFDI entstehen in den Disziplinen, aber auch disziplinübergreifend durch die Sektionen der NFDI, neue Ansatzpunkte für die Erschließung. Hier werden disziplinspezifische Metadatenschemata entwickelt, die kontrollierte Vokabulare aus Terminologien einbinden und so die Bereitstellung maschinenlesbarer Metadaten vorantreiben.[43] Kontrolliertes Vokabular kann entweder direkt zur Verschlagwortung oder aber durch eine Verknüpfung der Repositorien mit Terminologiediensten wie z. B. den *TerminologyServices4NFDI*[44] oder dem *TIB Terminology Service*[45] und den dort gehosteten Terminologien genutzt werden. Diese Services sind allerdings noch recht neu und entsprechend nur punktuell genutzt.[46] Des Weiteren stellt sich ebenso die Frage des Mappings mit bibliothekarischen Normdaten wie der GND.

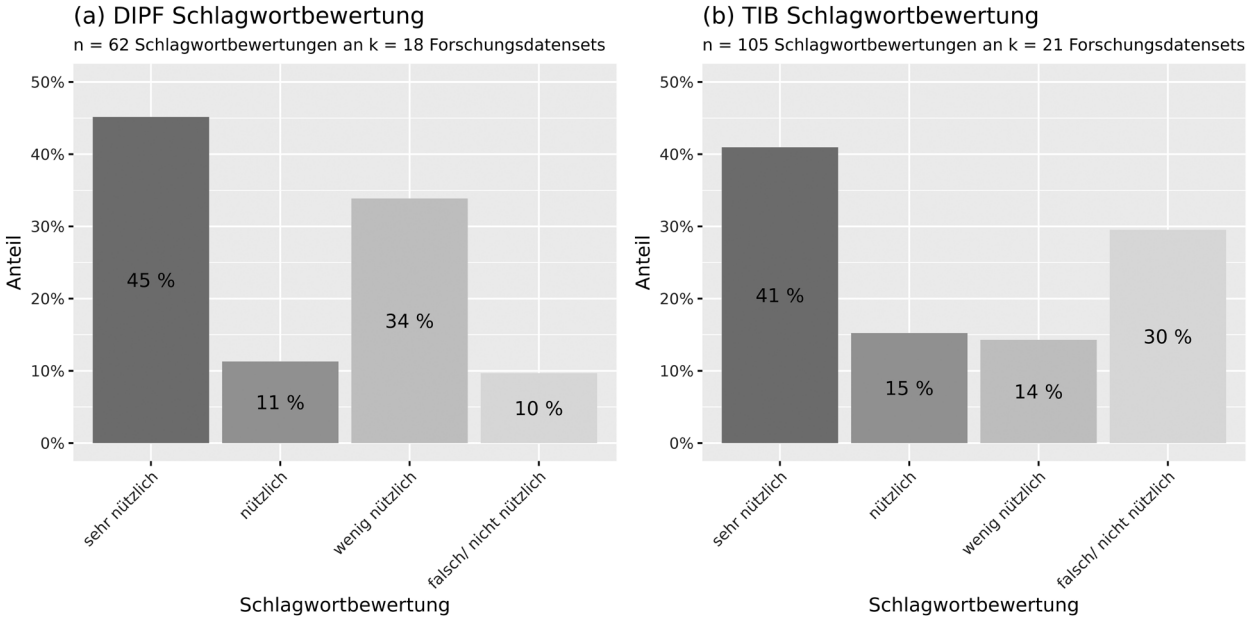
Über die technische Seite hinaus sollten die verschiedenen Akteure (Forscher*innen, Research Software Engineers, Betreiber von Repositorien, NFDI-Konsortien, Fachinformationsdienste, Bibliotheksverbünde), welche unterschiedliche Perspektiven, Kompetenzen und Interessen einbringen und jeweils nur Teilaspekte überblicken, Übereinkünfte treffen, wie eine hochqualitative Erschließung von MINT-Forschungsdaten gelingen kann. Fachreferent*innen an wissenschaftlichen Bibliotheken können zu dieser Aufgabe beitragen und den Austausch befördern, indem sie erstens die inhaltliche Erschließung weiterentwickeln, insbesondere in Hinsicht auf ihre Automatisierung, und zweitens kontrollierte Vokabulare kuratierend betreuen. Als Forum für den Diskurs aller Stakeholder bietet sich insbesondere die o. g. Sektion »(Meta-)Daten, Terminologien, Provenienz« der NFDI an.

Maschinelle Erschließung von Forschungsdaten im Test

Die oben für DIPF und TIB beschriebenen Verfahren der intellektuellen Erschließung von Forschungsdaten lassen sich nicht auf rein maschinelle Workflows abbilden. Das betrifft etwa die sehr wünschenswerte und nützliche Erschließung von Forschungsdaten durch Abstracts. Die intellektuelle verbale Erschließung auf der Basis von RSWK bot hier insofern eine Lösung, als die durch die Verknüpfung von Schlagwörtern gebildeten Schlagwortketten dezidiert dazu dienen sollten, »Kurz-Abstracts des Dokumenteninhalts anzuzeigen«[47] und den Nutzenden eine Hilfe zur Einschätzung der Relevanz und zur Selektion bei größeren Titelmengen zu geben. Die maschinelle Erschließung muss vorläufig hiervon Abstand nehmen und sich auf die Anreicherung mit thematischen Zugriffspunkten beschränken, bei denen keine syntaktische Verknüpfung im Sinne eines kohärenten Ganzen vorgenommen wird. Die Bereitstellung maschineller Äquivalente zu Kurz-Abstracts kann aktuell nicht geleistet werden, sollte aber langfristig Ziel der Entwicklungen sein. Die folgenden Überlegungen beziehen sich daher alle auf die klassische verbale Erschließung mittels Schlagwörtern.

Um die Möglichkeiten der maschinellen verbalen Erschließung von Forschungsdaten aufzuzeigen, wurde das Modell der GND-Beschlagwortung von Online-Publikationen aus der Erschließungsmaschine EMa,[48] das derzeit in der DNB eingesetzt wird, an einer kleinen Menge ausgewählter Forschungsdaten des DIPF und der TIB getestet. Es handelt sich dabei jedoch nur um ein Planspiel, das im Zusammenhang mit diesem Beitrag durchgeführt wurde. Die hier präsentierten Ergebnisse erheben daher keinen Anspruch auf Repräsentativität und sind nur als erste Eindrücke und Anregungen zu verstehen. Für repräsentative Tests müssten Verfahren und Modelle speziell für den Zweck der Erschließung von Forschungsdaten ausgewählt und angepasst werden. Die EMa ist eine modular aufgebaute Software, die in der IT-Infrastruktur der DNB betrieben wird und flexibel erweiterbar ist. So kann die Software mit geringem technologischem Aufwand an aktuelle Entwicklungen angepasst werden.

Die Klassifikation und Beschlagwortung erfolgen mithilfe unterschiedlicher KI-basierter Verfahren, die aus ANNIF, einem von der finnischen Nationalbiblio-



1 Bewertungsergebnisse maschineller Beschlagwortung in zwei Anwendungsfällen (a) Beschlagwortung deutschsprachiger Forschungsdatenabstracts aus der Bildungsforschung (DIPF) (b) Beschlagwortung englischsprachiger Forschungsdatenabstracts aus den Natur- und Ingenieurwissenschaften (TIB)

thek entwickelten Open-Source-Toolkit, in die EMa übernommen wurden. Für die Beschlagwortung mit der GND werden zwei Methoden aus ANNIF, Omikuji-Bonsai und MLLM (Maui-like Lexical Matching), ersteres ein lernendes, letzteres ein lexikalisches Verfahren, zu einem Ensemble-Modell kombiniert.

Im Rahmen eines Tests wurde eine geringe Anzahl an Forschungsmetadaten der TIB und des DIPF durch das Ensemble-Modell mit GND-Schlagwörtern erschlossen. Bei den Daten der TIB handelt es sich um 21 ausgewählte englischsprachige Abstracts und Titel von Forschungsdatensets. Die Daten des DIPF umfassen 18 ausgewählte deutschsprachige Titel und Abstracts von Forschungsdatensets (s. Abb. 1). Exemplarisch werden jeweils zwei maschinelle Erschließungsergebnisse von Forschungsdatensets mit ihren Bewertungen aus dem DIPF und der TIB in den Abbildungen 2 und 3 dargestellt.

Im Rahmen der Bewertung durch die Kolleg*innen des DIPF wurde festgestellt, dass eine positive Bewertung eines Schlagworts nicht zwangsläufig eine gute Gesamtindexierung eines Forschungsdatensets bedeutet. Im Rahmen des hier angewandten Bewertungsverfahrens konnte aber lediglich die Vergabe der einzelnen Schlagwörter berücksichtigt werden, nicht jedoch der Kontext. Es hat sich allerdings gezeigt, dass die Verwendung von vier bis fünf Schlagwörtern in der Regel nicht ausreichend ist, um ein Forschungsdatenset adäquat zu beschreiben. Außerdem werden auch im DIPF Schlagwörter wie »Deutschland« vergeben, die ohne Kontextualisierung mit anderen Schlagwörtern keine Aussagekraft besitzen. Des Weiteren sind die eine Methode bezeichnenden Schlagwörter, wie beispielsweise »Interview«, die für die Indexierung von Forschungsdaten von besonderer Relevanz sind, nicht berücksichtigt worden.

Die TIB hat festgestellt, dass die Qualität der maschinellen Erschließung in erheblichem Maße variiert. Dabei reicht das Spektrum von einer nahezu perfekten und vollständigen Erschließung bis zu einem totalen Ausfall. In der kleinen Stichprobe sind alle Varianten vertreten. In einem Fall konnte kein Schlagwort identifiziert werden. Stattdessen wurde ein Code aufgezeichnet, der keinem Schlagwort zugeordnet werden kann (s. Abb 2).

Die festgestellten Qualitätsunterschiede lassen sich auf verschiedene Faktoren zurückführen. Ein möglicher Grund könnte in der Eignung der jeweiligen Terminologie für die Beschreibung von Forschungsdaten liegen. Es besteht die Möglichkeit, dass Konzepte, die beim DIPF oder bei der TIB verwendet werden, in der GND fehlen. Des Weiteren ist zu berücksichtigen, dass die Verfügbarkeit von Trainingsmaterialien für jedes Konzept variiert, was sich unterschiedlich auf die einzelnen Forschungsdatensätze auswirkt. Ebenfalls von Bedeutung ist die Aussagekraft von Abstract und Titel, insbesondere im Hinblick auf die Vollständigkeit der enthaltenen Informationen. Eine weitere mögliche Ursache für Qualitätsverluste ist eine nicht optimierte Einstellung der EMa für Forschungsdaten.

MIKS 2 - Mehrsprachigkeit als Handlungsfeld interkultureller Schulentwicklung

GND-Vorschläge EMa:

04126892X	Schulentwicklung	*sehr nützlich*	
040223493	Grundschule	*sehr nützlich*	
040384039	Mehrsprachigkeit	*sehr nützlich*	
040473767	Professionalisierung	*sehr nützlich*	
040350932	Lehrerbildung	*sehr nützlich*	
1175990566	MINT-Fächer	*Falsch/Nicht Nützlich*	Bewertung Gesamtindexat: *Wenig Nützlich*

Goldstandard DIPF:

Mehrsprachigkeit; Sprachkompetenz; Sprachfertigkeit; Interkulturalität; Handlungsstrategie; Schulentwicklung; Multiplikatorenfortbildung; Lehrerfortbildung; Qualifizierungsmaßnahme; Professionalisierung; Grundschule; Migrationshintergrund; Intervention; Wirksamkeit; Beobachtung; Befragung; Interview; Deutschland

Audiovisuelle Aufzeichnungen von Schulunterricht in der Bundesrepublik Deutschland

GND-Vorschläge EMa:

040118827	Deutschland	*Falsch/Nicht Nützlich*	Bewertung Gesamtindexat: *Falsch/Nicht Nützlich*

Goldstandard DIPF:

Lehrerausbildung; Unterrichtsanalyse; 70er Jahre; 80er Jahre; 90er Jahre; 20. Jahrhundert; 21. Jahrhundert; Politische Bildung; Politische Didaktik; Politische Weltkunde; Sozialkunde; Videoaufzeichnung; Historische Bildungsforschung; Unterrichtsforschung; Deutschland; Deutschland-BRD

2 Zwei Erschließungsergebnisse deutschsprachiger Forschungsdatensets aus der Bildungsforschung (DIPF) mit den Bewertungen

3 Zwei Erschließungsergebnisse englischsprachiger Forschungsdatensets aus den Natur- und Ingenieurswissenschaften (TIB) mit den Bewertungen

Fazit

Anhand der hier betrachteten Fälle ist sichtbar geworden, dass die bekannten Probleme maschineller Erschließung von Medienwerken bei der Anwendung auf Forschungsdaten bestehen bleiben bzw. sogar verstärkt werden. Die maschinelle Erschließung von Forschungsdaten ist mit einer Reihe spezifischer Herausforderungen konfrontiert, die es zu bewältigen gilt. Es ist sicherzustellen, dass die Vorgaben der Datenschutzgrundverordnung (DSGVO) nicht nur in Bezug auf die Forschungsdaten selbst, sondern auch in Bezug auf die eingesetzten Automatisierungstools erfüllt werden. Im Gegensatz zu Medienwerken oder Publikationen weisen Forschungsdaten eine deutlich komplexere Struktur auf, die sich durch Verschachtelungen und unterschiedliche methodische Ansätze je nach Disziplin auszeichnet. In zahlreichen Disziplinen fehlen standardisierte Vokabulare, was die maschinelle Erschließung erschwert. Im Rahmen der Erschließung von Forschungsdaten wird der Fokus auf die Kriterien Reliabilität und Nachnutzbarkeit sowie auf Forschungsfragen und Datenqualität gelegt. Hier unterscheiden sich die Forschungsbedarfe von der primären Suche nach relevanter Literatur in Bibliotheken. In einigen Disziplinen sind darüber hinaus zusätzliche Erschließungsschritte wie Anonymisierung und Abstraktion notwendig, um den Forschenden Kontextinformationen zur Verfügung zu stellen. Ein weiteres Problem stellt die im Vergleich zu Publikationen begrenzte Verfügbarkeit von Trainingsdaten dar, was die Anwendung maschineller Lernverfahren erschwert. Angesichts der steigenden Menge an Forschungsdaten und begrenzter personeller Ressourcen wird die Entwicklung effizienter automatisierter Verfahren zunehmend wichtiger, um möglichst viele Forschungsdaten nachvollziehbar und für die Nachnutzung durch andere Forschende zugänglich zu machen. Die von Institutionen wie dem DIPF und der TIB angeregten Verfahren zeigen bereits vielversprechende Ansätze in diesem Bereich. Bei der Weiterentwicklung dieser Verfahren gilt es, eine Balance zwischen Quantitätssteigerung durch Erschließung mittels KI-Verfahren und der Qualität der Erschließung zu finden und die verschiedenen Funktionen der Forschungsdatenbereitstellung zu berücksichtigen. Maschinelle Erschließung muss dabei durch kontinuierliche Qualitätssicherung begleitet werden.

Die erfolgreiche Integration von maschinellen Erschließungsverfahren wird letztlich dazu beitragen, das Potenzial von Forschungsdaten besser auszuschöpfen und ihre Nachnutzbarkeit zu erhöhen. Allerdings verläuft die Entwicklung der notwendigen Technologien aufgrund der genannten Herausforderungen asynchron, sodass der Einsatz intellektueller Erschließung vorerst unverzichtbar bleibt.

Anmerkungen

1 Vgl. Bundesministerium für Bildung und Forschung. *Aufbau von Datenkompetenzzentren an Hochschulen und Forschungseinrichtungen*. [Zugriff am: 24.07.2024]. Verfügbar unter: https://www.bildung-forschung.digital/digitalezukunft/de/wissen/Datenkompetenzen/datenkompetenzzentren_für_die_wissenschaft_ordner/datenkompetenzzentren_fuer_die_wissenschaft_node.htm

2 Vgl. WIESENMÜLLER, Heidrun. Verbale Erschließungen in Katalogen und Discovery-Systemen – Überlegungen zur Qualität. In: FRANKE-MAIER, Michael; KASPRZIK, Anna; LEDL, Andreas; SCHÜRMANN, Hans, Hrsg. *Qualität in der Inhaltserschließung*. Berlin/Boston: De Gruyter, 2021, S. 279–301, hier S. 281 f.

3 Vgl. BERTRAM, Jutta. *Einführung in die inhaltliche Erschließung: Grundlagen – Methoden – Instrumente*. Würzburg: Ergon, 2005, S. 24 ff.

4 Vgl. BEE, Guido. Universalklassifikationen in Bibliotheken des deutschen Sprachraums. In: ALEX, Heidrun; BEE, Guido; JUNGER, Ulrike, Hrsg. *Klassifikationen in Bibliotheken: Theorie – Anwendung – Nutzen*. Berlin/Boston: De Gruyter, 2018, S. 23–63, hier besonders S. 56 f.

5 Vgl. BEE, Guido; HORSTKOTTE, Martin; JAHNS, Yvonne; KARG, Helga; NADJ-GUTTANDIN, Julijana. Wissen, was verbindet: GND-Arbeit zwischen Terminologie und Redaktion. *o-bib*. 2023, 20(1). [Zugriff am: 24.07.2024]. Verfügbar unter: https://doi.org/10.5282/o-bib/5900

6 Gesetz über die Deutsche Nationalbibliothek vom 22.06.2006. [Zugriff am: 24.07.2024]. Verfügbar unter: https://www.gesetze-im-internet.de/dnbg/BJNR133800006.html

7 Vgl. GÖMPEL, Renate; JUNGER, Ulrike; NIGGEMANN, Elisabeth. Veränderungen im Erschließungskonzept der Deutschen Nationalbibliothek. *Dialog mit Bibliotheken*. 2010, 22, S. 19–22.

8 Vgl. JUNGER, Ulrike und SCHOLZE, Frank. Neue Wege und Qualitäten – Die Inhaltserschließungspolitik der Deutschen Nationalbibliothek. In: FRANKE-MAIER, Michael; KASPRZIK, Anna; LEDL, Andreas; SCHÜRMANN, Hans, Hrsg. *Qualität in der Inhaltserschließung*. Berlin/Boston: De Gruyter, 2021, S. 55–70, hier S. 57 ff.

9 Vgl. Verordnung über die Pflichtablieferungsverordnung von Medienwerken an die Deutsche Nationalbibliothek, §9, num. 10. [Zugriff am: 29.7.2024]. Verfügbar unter: https://www.gesetze-im-internet.de/pflav/__9.html. Gesammelt werden lediglich öffentlich zugängliche Forschungsdaten, die für eine ablieferungspflichtige Dissertation und die Nachprüfbarkeit der darin enthaltenen Forschungsergebnisse unverzichtbar sind.

10 Derzeit besteht der Kreis der am VerbundFDB beteiligten FDZ aus dem FDZ am IQB (Institut zur Qualitätsentwicklung im Bildungswesen) für Daten der Kompetenz- und Leistungsmessung, dem GESIS (Leibniz-Institut für Sozialwissenschaften e.V.) für Umfrage- und Aggregatdaten, dem FDZ Bildung am DIPF (Leibniz-Institut für Bildungsforschung und Bildungsinformation) für qualitative Daten, dem FDZ des DZHW (Deutsches Zentrum für Hochschul- und Wissenschaftsforschung) für Daten aus dem Bereich der Hochschulforschung und dem FD-LEX (Forschungsdatenbank Lernertexte) für Lernertexte.

11 STEINHARDT, Isabel; FISCHER, Caroline; HEIMSTÄDT, Maximilian; HIRSBRUNNER, Simon David; İKIZ-AKINCI, Dilek; KRESSIN, Lisa; KRETZER, Susanne; MÖLLENKAMP, Andreas; PORZELT, Maike; RAHAL, Rima-Maria; SCHIMMLER, Sonja; WILKE, René; WÜNSCHE, Hannes. *Das Öffnen und Teilen von Daten qualitativer Forschung*. RatSWD Working Paper 273/2020. Berlin: Rat für Sozial- und Wirtschaftsdaten (RatSWD), 2020, S. 7. [Zugriff am: 14. Juni 2024]. Verfügbar unter: https://doi.org/10.17620/02671.55

12 METSCHKE, Rainer und WELLBROCK, Rita. *Datenschutz in Wissenschaft und Forschung*. Berlin, 2002, S. 21.

13 Vgl. MEYERMANN, Alexia und PORZELT, Maike. *Hinweise zur Anonymisierung qualitativer Daten*. Frankfurt am Main: DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation, 2014. [Zugriff am: 14. Juni 2024]. Verfügbar unter: https://doi.org/10.25656/01:21968

14 Vgl. MOZYGEMBA, Kati und HOLLSTEIN, Betina. (2023). *Anonymisierung und Pseudonymisierung qualitativer textbasierter Forschungsdaten – eine Handreichung*. 2023. [Zugriff am: 14. Juni 2024]. Verfügbar unter https://doi.org/10.26092/elib/2525

15 Der Text wird zunächst vom Stanford Named Entity Recognizer analysiert, der Organisationen, Personen und Daten extrahiert. Danach werden alle identifizierten Entitäten im Text ersetzt. Um die Qualität der Datenanonymisierung weiter zu verbessern, kann anschließend das Named Entity Recognition Tool von NLP Compromise ausgeführt werden, um verbleibende nicht erkannte Entitäten zu identifizieren. Dabei handelt es sich um Tools, die probalistische und lexikonbasierte Ansätze kombinieren. Vgl. KLEINBERG, Bennett; MOZES, Maximilian; VAN DER TOOLEN, Yaloe; VERSCHUERE, Bruno. *NETANOS – Named entity-based Text Anonymization for Open Science*. 2017, S. 10. [Zugriff am: 14. Juni 2024]. Verfügbar unter: https://doi.org/10.31219/osf.io/w9nhb sowie KLEINBERG, Bennett und MOZES, Maximilian. *Web-based text anonymization with Node.js: Introducing NETANOS (Named entity-based Text Anonymization for Open Science)*. The Journal of Open Source Software. 2017, 2(14), S. 293. Verfügbar unter: https://doi.org/10.21105/joss.00293

16 https://www.fdz-bildung.de

17 HAHN, Udo. B 12 Automatisches Abstracting, In: Rainer KUHLEN; Wolfgang SEMAR und Dietmar STRAUCH, Hrsg. *Grundlagen der praktischen Information und Dokumentation: Handbuch zur Einführung in die Informationswissenschaft und -praxis*. 6. Aufl. Berlin/Boston: De Gruyter Saur, 2013, S. 286–301, S. 286. Verfügbar unter: https://doi.org/10.1515/9783110258264.286

18 Vgl. HWANG, Taesoon; AGGARWAL, Nishant; ZARAK KHAN, Pir; ROBERTS, Thomas; MAHMOOD, Amir; GRIFFITHS, Madlen M.; PARSONS, Nick; KHAN, Saboor. Can ChatGPT assist authors with abstract writing in medical journals? Evaluating the quality of scientific abstracts generated by ChatGPT and original abstracts. *PLoS ONE*. 2024. 19(2): e0297701. Verfügbar unter: https://doi.org/10.1371/journal.pone.0297701

19 DANIEL, Andreas; JAKOWATZ, Stefan; JUNG, Nadeshda; KLEINE, Lydia; KOCAJ, Aleksander; MEYERMANN. Alexia; MOZYGEMBA, Kati; SCHUSTER, Alexander. *Die Erfassung von Publikationen aus der Datennutzung: Verfahren, Herausforderungen und Nutzen. Ein Erfahrungsbericht von Forschungsdatenzentren.* RatSWD Working Paper 281/2023. Berlin: Rat für Sozial- und Wirtschaftsdaten (RatSWD), 2023, S. 4. [Zugriff am: 14. Juni 2024]. Verfügbar unter: https://doi.org/10.17620/02671.80

20 Ebd.

21 Projektliteratur umfasst Monografien, Konferenzschriften, Dissertationen, Abschlussberichte, Technical Reports, Zeitschriftenaufsätze, graue Literatur, Sammelwerke, Sammelwerksbeiträge. Damit nicht gemeint sind Projektposter, Vortragsfolien oder ähnliches.

22 https://www.fachportal-paedagogik.de/

23 https://www.fachportal-paedagogik.de/literatur/produkte/fis_bildung/fis_bildung.html

24 https://zdb-katalog.de/index.xhtml

25 https://www.dnb.de/DE/Home/home_node.html

26 https://swb.bsz-bw.de
27 Dieser Korpus setzt sich aus 104 Monografien, 27 Sammelwerken, 28 Sammelwerksbeiträgen, 23 Zeitschriftenaufsätzen sowie einer Hochschulschrift zusammen.
28 ANNIF – Tool for automated subject indexing and classification; National Library of Finland unter: https://annif.org/
29 Malibu – Mannheim library utilities; Universität Mannheim unter: https://data.bib.uni-mannheim.de/malibu/isbn/suche.html
30 IRIS.AI The Researcher Workspace; Iris AI AS unter: https://iris.ai/
31 Einen Sonderfall stellen den Forschungsprozess dokumentierende Materialien wie Fotos von Versuchsaufbauten o. ä. dar: Sie enthalten Forschungsdaten und stützen ihre Interpretation, sind aber i. d. R. nicht die Essenz der Forschungsergebnisse. Auf letztere Daten beziehen sich die Verfasser*innen im Folgenden.
32 MINT = Mathematik, Informatik, Naturwissenschaften und Technik.
33 Die Zahlen aus dem TIB-Portal (https://www.tib.eu/de/) geben einen ungefähren Anhaltspunkt: So sind für die Physik, auf die 72 % aller nachgewiesenen Forschungsdaten entfallen, 286.000 Forschungsdatensätze nachgewiesen, denen 15,9 Millionen Zeitschriftenartikel gegenüberstehen. Es ist eine vernünftige Annahme, als erste, grobe Näherung von einem 1:1–Verhältnis überhaupt existierender Artikel und Forschungsdatensätze auszugehen.
34 https://www.fdm.uni-hannover.de/de/forschungsdatenrepositorium
35 WEINSPACH, Karoline. *Digitaler Assistent für das Forschungsdatenmanagement (DA-FDM) – Erweiterung des DA-3 zur Sacherschließung von Forschungsdaten*. Folien zu einem Vortrag bei der 112. BiblioCon. [Zugriff am: 15.07.2024]. Verfügbar unter https://nbn-resolving.org/urn:nbn:de:0290-opus4-190634
36 Z. B.: LOUYS, Mireille; RICHARDS, Anita; BONNAREL, François; MICOL, Alberto; CHILINGARIAN, Igor; McDOWELL, Jonathan; International Virtual Observatory Alliance Data Model Working Group. *Data Model for Astronomical DataSet Characterisation, Version 1.13*. IVOA Recommendation, 2008. [Zugriff am 26.07.2024]. Verfügbar unter https://doi.org/10.25504/FAIRsharing.yYjZWb
37 Z. B.: CONRAD, Tim; FERRER, Eloi; MIETCHEN, Daniel; PUSCH, Larissa; STEGMÜLLER, Johannes; SCHUBOTZ, Moritz. Making Mathematical Research Data FAIR: A Technology Overview. [Zugriff am: 26.07.2024]. Verfügbar unter https://doi.org/10.48550/arXiv.2309.11829; siehe auch z. B. https://www.nfdi4chem.de/de/wie-man-mit-fair-daten-in-der-chemie-beginnt/
38 Open Research Knowledge Graph; vgl. https://orkg.org
39 VAN EDIG, Xenia; DELLMANN, Sarah; RENZIEHAUSEN, Anna-Karina; ZIEDORN, Frauke. Data Papers – Eine Ode an die Daten. Blogeintrag im TIB-Blog, 15.02.2022. [Zugriff am: 15.07.2024]. Verfügbar unter: https://blog.tib.eu/2022/02/15/data-papers-eine-ode-an-die-daten/
40 LU, Linna, VAN EDIG, Xenia; RENZIEHAUSEN, Anna-Karina. Software Journals: der Einfluss von Software auf die Forschung. Blogeintrag im TIB-Blog, 29.02.2024. [Zugriff am: 15.07.2024]. Verfügbar unter: https://blog.tib.eu/2024/02/29/software-journals-der-einfluss-von-software-auf-die-forschung/
41 Z. B. https://www.inggrid.org/
42 Ein generischer Ansatz sind FAIR Digital Objects, siehe https://fairdo.org/
43 STRÖMERT, Philip; LIMBACHIA, Vatsal; OLADAZIMI, Pooya; HUNOLD, Johannes; KOEPLER, Oliver. *Towards a Versatile Terminology Service for Empowering FAIR Research Data: Enabling Ontology Discovery, Design, Curation, and Utilization Across Scientific Communities.* Studies on the Semantic Web. Volume 56: Knowledge Graphs, Semantics, Machine Learning, and Languages. [Zugriff am: 29.07.2024]. Verfügbar unter: https://ebooks.iospress.nl/doi/10.3233/SSW230005
44 https://base4nfdi.de/projects/ts4nfdi
45 https://terminology.tib.eu/ts
46 Die semantische Annotation von Forschungsdaten in der Chemie schreitet durch die Vernetzung von ELN-TS-Repos schnell voran. Das Mapping zur GND wird mit der zukünftigen Integration von Mapping Services angegangen.
47 RSWK 13,1; vgl. SCHEVEN, Esther und NADJ-GUTTANDIN, Julijana, Bearb. *Regeln für die Schlagwortkatalogisierung: RSWK*. 4. Aufl. 2017. [Zugriff am: 25.07.2024]. Verfügbar unter: https://d-nb.info/1126513032/34
48 MÖDDEN, Elisabeth. *Maschinelle Beschlagwortung mit Algorithmen.* Ein Blick in die Werkstatt des KI-Projektes der Deutschen Nationalbibliothek. *b.i.t. online* 27 (3), 2024, S. 242. [Zugriff am: 28.07.2024]. Verfügbar unter: 2024-03-fachbeitrag-moedden.pdf (b-i-t-online.de)

Verfasser*Innen

Tristan Bauder, Arbeitsbereich Forschungsdaten Bildung, DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation, Rostocker Straße 6, 60323 Frankfurt am Main, Telefon +49 69 24708-493, t.bauder@dipf.de
Foto: DIPF

Dr. Guido Bee, Inhaltserschließung, Deutsche Nationalbibliothek, Adickesallee 1, 60322 Frankfurt am Main, Telefon +49 69 1525-1511, g.bee@dnb.de, https://orcid.org/0000-0002-2896-5790
Foto: DNB, Josephine Kreutzer

Mathias Begoin, Fachreferat Elektrotechnik / Verkehrstechnik, Abteilung Bestandsentwicklung und Metadaten, TIB – Leibniz Informationszentrum Technik und Naturwissenschaften und Universitätsbibliothek, Welfengarten 1 B, 30167 Hannover, Telefon +49 511 762-14140, mathias.begoin@tib.eu
Foto: Christian Bierwagen

Dr. Sarah Dellmann, Bereich Publikationsdienste – Stellv. Bereichsleitung Publikationsdienste, Programmbereich Benutzungs- und Informationsdienste, TIB – Leibniz Informationszentrum Technik und Naturwissenschaften und Universitätsbibliothek, Welfengarten 1 B, 30167 Hannover, Telefon +49 511 762-4005, sarah.dellmann@tib.eu
Foto: Christian Bierwagen

Dr. Linus Hartmann-Enke, Wissenschaftlicher Dienst, Fachreferent für Musik & Geschichte, Deutsche Nationalbibliothek, Deutscher Platz 1, 04103 Leipzig, Telefon +49 341 2271-518, l.hartmann-enke@dnb.de, https://orcid.org/0000-0003-3792-5907
Foto: Nadja Enke

Dr. Holger Israel, Fachreferat Mathematik, Koordinierung der maschinellen Inhaltserschließung im wissenschaftlichen Dienst, Abteilung Bestandsentwicklung und Metadaten, TIB – Leibniz Informationszentrum Technik und Naturwissenschaften und Universitätsbibliothek, Welfengarten 1 B, 30167 Hannover, Telefon +49 511 762-3979, holger.israel@tib.eu
Foto: Christian Bierwagen

Nadeshda Jung, Arbeitsbereich Forschungsdaten Bildung, DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation, Rostocker Straße 6, 60323 Frankfurt am Main, Telefon +49 69 24708-309, n.jung@dipf.de
Foto: privat

Maximilian Kähler, Automatische Erschließungsverfahren, Netzpublikationen, Deutsche Nationalbibliothek, Deutscher Platz 1, 04103 Leipzig, Telefon +49 341 2271-133, m.kaehler@dnb.de, https://orcid.org/0000-0003-4695-0565
Foto: DNB, Josephine Kreutzer

Dr. rer. nat. Oliver Koepler, Bereichsleitung Lab Linked Scientific Knowledge, Forschung und Entwicklung, TIB – Leibniz Informationszentrum Technik und Naturwissenschaften und Universitätsbibliothek, Welfengarten 1 B, 30167 Hannover, Telefon +49 511 762-3449, oliver.koepler@tib.eu
Foto: Christian Bierwagen

Dr. Angelina Kraft, Lead Lab Research Data Services, TIB – Leibniz Informationszentrum Technik und Naturwissenschaften und Universitätsbibliothek, Welfengarten 1 B, 30167 Hannover, Telefon +49 163 762-5961, angelina.kraft@tib.eu
Foto: Christian Bierwagen

Elisabeth Mödden, Leitung Referat Automatische Erschließungsverfahren, Netzpublikationen, Deutsche Nationalbibliothek, Adickesallee 1, 60322 Frankfurt am Main, Telefon +49 69 1525-1533, e.moedden@dnb.de, https://orcid.org/0000-0001-6809-3926
Foto: privat

Stefanie Psczolla, Arbeitsbereich Forschungsdaten Bildung, DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation, Rostocker Straße 6, 60323 Frankfurt am Main, Telefon +49 69 24708-324, s.psczolla@dipf.de
Foto: DIPF

Dr. Dirk Weisbrod, Verbund Forschungsdaten Bildung, DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation, Rostocker Straße 6, 60323 Frankfurt am Main, Telefon +49 69 24708-391, d.weisbrod@dipf.de
Foto: DIPF