Inhalte explorativ durchsuchen

Geisteswissenschaftliche Forschungsdaten und Facettendrilldown

Roman Bleier, Sebastian Stoff

Abstract: *Die stetig zunehmende Komplexität von geisteswissenschaftlichen Forschungsdaten verlangt nach neuen programmatischen Zugängen, zu denen auch das im Artikel vorgestellte explorative Durchsuchen bzw. Untersuchen zählt. Explorative Zugänge – wie der in diesem Beitrag besprochene Facettendrilldown – ermöglichen einen niederschwelligen Einstieg bei der Arbeit mit Forschungsdaten und gleichzeitig können Expert:innen neue Zusammenhänge erkennen. Die Entwicklung von geeigneter Software – wie dem Facettendrilldown – verlangt nach umfassenden Mitteln, die von einzelnen Forschungsprojekten kaum aufgebracht werden können. Eine breite Nachnutzung in anderen Projekten ist daher erstrebenswert, besonders in einer Infrastruktur wie dem Geisteswissenschaftlichen Asset Management System (GAMS) der Universität Graz, in dem standardisierte Daten und Abläufe bereits die Grundlagen für projektübergreifende Applikationen liefern.*

Keywords: *Informationssystem; Digitales Repositorium; Digitale Edition; Digitale Geisteswissenschaften; Forschungsdaten; Suche; Datenanalyse*

Einleitung

Die Digitalen Geisteswissenschaften sehen sich mit einer stetig zunehmenden Datenmenge und -komplexität konfrontiert. Dabei stellt nicht nur nachhaltige Aufbewahrung, Verwaltung und Sicherung der digitalen Inhalte die Wissenschaft vor eine besondere Herausforderung, sondern bereits die Zurverfügungstellung eines verständlichen Zugangs (oder verständlicher Zugänge) zu den erschlossenen Forschungsdaten fordert Aufmerksamkeit.[1] Dieser Beitrag beschäftigt sich primär mit

1 Vgl. Heike Neuroth; Bibliothek, Archiv, Museum; in: Fotis Jannidis/Hubertus Kohle/Malte Rehbein (Hg.), Digital Humanities, Stuttgart 2017, 214, DOI: 10.1007/978-3-476-05446-3_23 (abgerufen 8.9.2022); Malte Rehbein, Informationsvisualisierung, in: Fotis Jannidis/Hubertus Kohle/Malte Rehbein (Hg.), Digital Humanities, Stuttgart 2017, 328–342, DOI: 10.1007/978-3-

digitalen Editionen, die im digitalen Repository des Zentrums für Informationsmodellierung (ZIM) erstellt wurden und werden, und wofür es standardisierte Abläufe im Geisteswissenschaftliches Asset Management System (GAMS) gibt. Editionen, auch digitale Editionen, sind von zentraler Bedeutung für die Bereitstellung von Texten und Daten für die geisteswissenschaftliche Forschung allgemein und die historische Forschung im Speziellen.

Digitale Editionen[2] begegnen dem Problem der zunehmenden Datenmenge und -komplexität (unter anderem) mit der Zurverfügungstellung unterschiedlicher webbasierter Such- und Filterverfahren wie zum Beispiel einer Volltextsuche integriert in die allgemein zugängliche Weboberfläche.[3] Such- und Filteransätze unterscheiden sich bisweilen stark in Bezug auf den Implementierungsaufwand. Verkürzt gesagt gilt die Regel: Je strukturierter der Zugang zu den Daten sein soll, desto mehr Aufwand erzeugt die nötige Strukturierung des Forschungsmaterials und desto aufwendiger gestaltet sich die maschinelle Auswertung (der angelegten Strukturen).[4] Richtig angewandt sind es jedoch gerade Suchstrategien basierend auf hochstrukturierte Daten, die ein produktives Arbeiten mit komplexen Forschungsdaten erlauben und darüberhinaus Forscher:innen – womöglich andernfalls unerkannte – Einsichten in gesammeltes Quellenmaterial ermöglichen. Tools für exploratives Durchsuchen können gerade den letzten Punkt fördern.

In den Digitalen Geisteswissenschaften gilt als allgemein akzeptierte Tatsache, dass es umfangreiche Kooperationen unter anderem in Form von gemeinsamen Infrastrukturen (wie zum Beispiel Datenzentren) braucht, um den zuvor ausschnitthaft beschriebenen Anforderungen digitaler Forschung gerecht werden zu können.[5] Gerade im Umfeld digitaler Editionen und der Text Encoding Initiative (TEI) gibt

476-05446-3_23 (abgerufen 8.9.2022); vgl. auch Johanna Drucker, Visualization and Interpretation: Humanistic Approaches to Display Visualization and Interpretation, Cambridge, Massachusetts 2020.

2 Vgl. auch Patrick Sahle, 2. What is a Scholarly Digital Edition?, in: Matthew J. Driscoll/Elena Pierazzo (Hg.), Digital Scholarly Editing: Theories and Practices, Cambridge 2016, 19–39.

3 Wie zum Beispiel die Digitale Edition »Hugo Schuchardt Archiv« am GAMS https://gams.uni-graz.at/hsa (abgerufen 8.9.2022), oder »Stefan Zweig Digital« ebenfalls am GAMS https://gams.uni-graz.at/szd (abgerufen 8.9.2022).

4 Vgl. u.a. Jeffrey Beall, The Weaknesses of Full-Text Searching, in: The Journal of Academic Librarianship 34 (2008) 438–444; Barbara H. Kwasnick, The role of classification in knowledge representation and discovery, in: Library Trends 48 (1999) 22–47.

5 Vgl. Johannes Stigler, Digitale Nachhaltigkeit, in: Helmut W. Klug (Hg. unter Mitarbeit von Selina Galka und Elisabeth Steiner), KONDE Weißbuch, 2021, Handle: hdl.handle.net/11471/562.50.6 (abgerufen 8.9.2022). Im Projekt Kompetenznetzwerk Digitale Edition (KONDE) wurde versucht, gemeinsame Infrastrukturen für digitale Editionen in Österreich zu schaffen. Vgl. Georg Vogeler, Einleitung: Gibt es eine österreichische Editionskultur, in: Roman Bleier/Helmut W. Klug, Digitale Edition in Österreich. Norderstedt 2023, III–X.

es immer wieder die Forderung nach generischen Tools und übergreifenden Infrastrukturen.[6] Wir wollen in diesem Beitrag am Beispiel der Applikation Facettendrilldown, die exploratives Durchsuchen von Editionsdaten ermöglicht, die Forderung nach generischen Tools wiederholen und darlegen, inwiefern nur die Kooperation zwischen verschiedenen Forschungsprojekten und -partnern eine erfolgreiche Umsetzung ermöglicht.

Der Beitrag ist in vier Abschnitte unterteilt: Im ersten Abschnitt wird exploratives Untersuchen in der geisteswissenschaftlichen Forschung und grundlegende Terminologie zum Facettendrilldown besprochen. Die standardisierten Abläufe zur Datenmodellierung, -verarbeitung und -analyse sind Thema des nächsten Abschnitts und diese werden exemplarisch an der Edition der Regensburger Reichstagsakten von 1576 (RTA 1576) im dritten Abschnitt besprochen. Im letzten Teil wird eine facettierte Drilldown-Anwendung, die gegenwärtig am ZIM entwickelt und im Projekt RTA 1576 implementiert wird, vorgestellt. Die Ausrichtung des Beitrags und die Beispiele sind zwar aus der digitalen Reichstagsedition gewählt, bei ihrer Beschreibung ist uns aber explizit die Übertragbarkeit auf andere Projekte aus dem Umfeld der digitalen Geschichtswissenschaft und auf das Projekt Digitale Erinnerungslandschaft Österreichs (DERLA) im Speziellen wichtig. Im GAMS sind über 60 Projekte nach dem Prinzip der Standardisierung und Übertragbarkeit von Datenmodellierung, -verarbeitung und -analyse veröffentlicht und das ZIM versucht Anwendungen wie das Facettendrilldown so zu entwickeln, dass die Applikation leicht von anderen Projekten nachgenutzt werden kann.

Exploratives Erforschen

Quantitative Datenanalyse hat zwar eine lange Tradition in den Geschichtswissenschaften, besonders in der Sozial- und Wirtschaftsgeschichte, aber die Entwicklung in den vergangenen zwei Jahrzehnten hat quantitative Methoden vermehrt in den Fokus der Forschung gerückt. Durch die verstärkte Verwendung des Internets

6 Im Bereich des digitalen Edierens gibt es seit Jahren Bestrebungen von der TEI Stylesheets und Tools für die Verarbeitung und Veröffentlichung von TEI Dokumenten bereitzustellen. Tools, in: TEI <text encoding initiative>, https://tei-c.org/tools/(abgerufen 8.9.2022). Andere Tools kommen aus konkreten Projekten, aber wurden für eine breitere Nachnutzung überarbeitet: Beispiele sind die Anwendung Edition Visualization Technology (EVT, http://evt.labcd.unipi.it/) oder die Versioning Machine (VM, http://v-machine.org).
Als ein Beispiel für eine projektübergreifende Infrastruktur möchten wir in diesem Zusammenhang das Projekt correspSearch anführen, in dem Daten aus Briefeditionen verwendet werden, um eine editionsübergreifende Datenbank zu befüllen (correspSearch: Search scholarly editions of letters, https://correspsearch.net/en/home.html).

durch Historiker:innen und die beginnende Massendigitalisierung von geisteswissenschaftlichen Quellenmaterialien sind jetzt viel mehr Texte und Daten digital verfügbar als jemals zuvor. Man kann daher in manchen Bereichen bereits von »Big Data« (auch in den Geisteswissenschaften) sprechen und dadurch werden quantitative Methoden aus der Informatik und Statistik verstärkt relevant für Datenauswertung und geschichtswissenschaftliche Forschung.[7]

Am bekanntesten ist der aus dem Bereich der digitalen Literaturgeschichte stammende »Distant Reading«-Ansatz.[8] Distant Reading, ein Begriff der von Franco Moretti eingeführt wurde[9], bezeichnet eine quantitative Herangehensweise bei der Analyse von großen Textmengen. Bei der Analyse spielen auch Visualisierungen eine zentrale Rolle, da diese es ermöglichen, Zusammenhänge, Muster oder Ausreißer in einem Datensatz besser zu erkennen. Der Name Distant Reading wurde analog zum »Close Reading«, der traditionellen Methode der Textanalyse in der literaturwissenschaftlichen Forschung, benannt und wird meist in Kombination mit diesem eingesetzt. Distant Reading ermöglicht es, losgelöst von Details Zusammenhänge, Muster oder Ausreißer zu erkennen, aber für die Interpretation und Detailanalyse ist das Close Reading durch Expert:innen notwendig. In Editionsprojekten werden Werkzeuge geschaffen, mit denen Nutzer:innen je nach Bedarf Distant Reading und Close Reading betreiben können, was auch als »Scalable Reading« bezeichnet wird.[10] Scalable Reading ist ein wichtiger Zugang für Fachleute, da das Distant Reading ohne ein Close Reading nicht sinnvoll ist und umgekehrt ist es auch ein Vorteil, wenn man bei der Detailstudie von Texten durch Musteranalyse größere Zusammenhänge nicht aus den Augen verliert.

Einem zentralen Aspekt des digitalen Paradigmas folgend, werden digitale Editionen nicht nur als eine Sammlung von edierten Texten und Paratexten betrachtet, sondern zeichnen sich durch eine verstärkte Datenorientierung aus.[11] Diese Datenorientierung kommt bereits durch die semantische Anreicherung in den TEI/

7 Vgl. Jonathan Blankey/Sarah Milligan/Marty Steer/Jane Winters, Doing Digital History, Manchester 2021, 5–25. Mareike König, Digitale Methoden in der Geschichtswissenschaft: Definitionen, Anwendungen, Herausforderungen, in: BIOS – Zeitschrift für Biographieforschung, Oral History und Lebensverlaufsanalysen 1–2 (2017) 7–21, DOI: 10.3224/bios.v30i1-2.02 (abgerufen 8.9.2022).

8 Vgl. Ted Underwood, A Genealogy of Distant Reading, in: Digital Humanities Quarterly 11 (2017) 1, https://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html (abgerufen 8.9.2022).

9 Vgl. Franco Moretti, Conjectures on World Literature, in: New Left review 1 (2000) 54–68.

10 Vgl. Martin Mueller, Scalable Reading dedicated to DATA: digitally assisted text analysis. Vgl. auch: Katharina Zeppezauer-Wachauer, Distant Reading, Close Reading, Scalable Reading, in: Helmut W. Klug (Hg., unter Mitarbeit von Selina Galka und Elisabeth Steiner), KONDE Weißbuch, 2021, Handle: hdl.handle.net/11471/562.50.71 (abgerufen 26.09.2022).

11 Vgl. Patrick Sahle, 2. What is a Scholarly Digital Edition?, in: Matthew J. Driscoll/Elena Pierazzo (Hg.), Digital Scholarly Editing: Theories and Practices, Cambridge 2016, 32.

XML Dateien zum Ausdruck, aber auch bei der Schaffung von Zugängen für Nutzer:innen spielt die Datenorientierung eine zentrale Rolle, da Suchen, quantitative Auswertungen und (interaktive) Visualisierungen auf den Editionsdaten basieren. Besonders relevant ist dies bei Editionen in den Geschichtswissenschaften, da hier durch das allgemeine Interesse der Historiker:innen an Inhalten und »Fakten« eine inhalts- und datenzentrierte Editionsform bevorzugt verwendet wird. Georg Vogeler schlägt für solche Editionen den Begriff »assertive editions« vor.[12] »Assertive editions« enthalten visuelle Zugänge zu den Editionsdaten, die das Arbeiten für Fachleute vereinfachen sollen. Am Zentrum für Informationsmodellierung der Universität Graz (ZIM) wurde diese Art des Edierens etwa bei der Edition von Rechnungsbüchern[13] oder Dokumenten zur Herdsteuer im England des 17 Jh. erprobt.[14] Die in diesem Beitrag vorgestellte Edition der Reichstagsakten wurde auch als eine solche Edition konzipiert. Je nach Art der Daten können dabei Zeitstrahl (Timeline), über Graphvisualisierungen bis hin zu Karten zum Einsatz kommen. Diese visuellen Darstellungen bieten den Nutzer:innen unterschiedliche Perspektiven auf die Editionsdaten.

Durch Interaktionsmöglichkeiten erhalten diese visuellen Zugänge einen zusätzlichen Mehrwert für die Nutzer:innen: zum Beispiel durch das Aus- bzw. Einblenden von zusätzlichen Daten, Zoom-Funktionalität oder durch die Möglichkeit, einen Teildatensatz für die weitere Analyse auszuwählen. Abb. 1 zeigt als Beispiel den Netzwerkgraph im Weißbuch des Kompetenznetzwerks Digitale Edition (KONDE), der es Nutzer:innen ermöglicht, explorativ die Vielzahl an Themen, die das Weißbuch enthält, und Zusammenhänge zu erkennen. Ein Interface, das Interaktion mit den Forschungsdaten ermöglicht, fördert einen explorativen Ansatz bei der Analyse, womit gemeint ist, dass Nutzer:innen dabei unterstützt werden, den jeweiligen Datensatz kennenzulernen und relevante Information zu finden. Dieses ex-

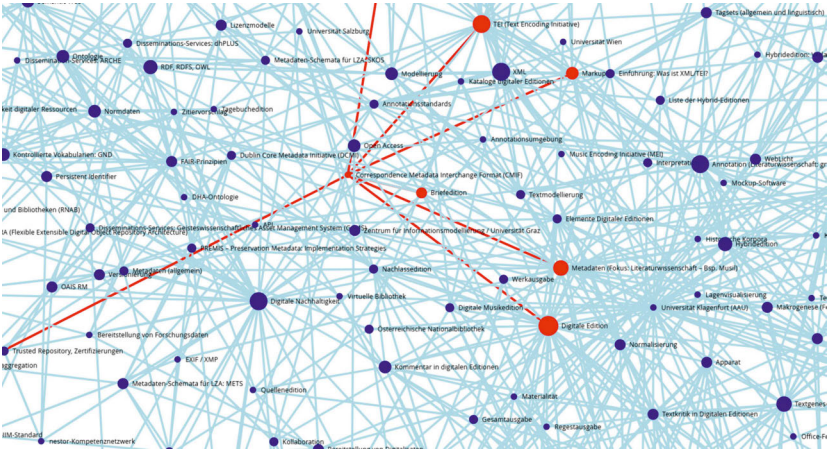
12 Vgl. Georg Vogeler, The ›assertive edition‹. On the consequences of digital methods in scholarly editing for historians, in: International Journal of Digital Humanities 1 (2019) 309–322; auch Georg Vogeler/Christopher Pollin/Roman Bleier, ›Ich glaube Fakt ist…‹. Der geschichtswissenschaftliche Zugang zum digitalen Edieren In: Karoline Döring/Stefan Haas/Mareike König/Jörg Wettlaufer (Hg.), Digital History. Konzepte, Methoden und Kritiken Digitaler Geschichtswissenschaft, Berlin 2022,171– 190.

13 Vgl. Georg Vogeler, Warum werden mittelalterliche und frühneuzeitliche Rechnungsbücher eigentlich nicht digital ediert?, in: Constanze Baum/Thomas Stäcker (Hg.), Grenzen und Möglichkeiten der Digital Humanities (Sonderband der Zeitschrift für digitale Geisteswissenschaften, 1), 2015, DOI: 10.17175/sb001_007. Susanna Burghartz (Hg.), Jahresrechnungen der Stadt Basel, http://gams.uni-graz.at/context:srbas?mode=projekt (abgerufen 8.9.2022).

14 Vgl. Andrew Wareham, et al., The ›Confronting the Digital‹ Debate and an Assertive Digital Edition: British History and Hearth Tax Records, in: Sanita Reinsone, et al., Post-Proceedings of the 5th Conference Digital Humanities in the Nordic Countries (DHN 2020), CEUR Workshop Proceedings, Volume 2865, 2021, 39–50, http://ceur-ws.org/Vol-2865/paper4.pdf; https://gams.uni-graz.at/context:htx (abgerufen 8.9.2022).

plorative Untersuchen der Forschungsdaten[15] kann sowohl für Nutzer:innen, die die Inhalte der Daten nicht kennen und diese schrittweise kennenlernen möchten, als auch für Expert:innen, die durch Kombination unterschiedlicher Kategorien neue Zusammenhänge und Erkenntnisse erzielen möchten, interessant sein.

Abb. 1: Explorative Suche mit dem Netzwerkgraph im KONDE-Kompetenznetzwerk Digitale Edition – Weißbuch, (Quelle: Eigener Screenshot – KONDE Weißbuch).



Eine Form des explorativen Durchsuchens von Forschungsdaten, die wir in diesem Beitrag vorstellen möchten, ist die Möglichkeit über Facetten eine Einschränkung, oder »Drilldown«, von Suchergebnissen zu bewirken. Der Facettendrilldown verbindet traditionelle Such- und Filtermöglichkeiten, die für Datenbanken schon lange Tradition sind, mit quantitativer Auswertung, wodurch das explorative Potential gefördert wird. Im Kontext von Facettensuche und -drilldown sind Facetten Kategorien und Unterkategorien eines hierarchischen Ordnungssystems. Basierend auf diesem System können Filtermöglichkeiten für die verfügbaren Daten ausgeführt werden. Es können auch mehrere Kategorien miteinander kombiniert werden, wodurch eine weitere Einschränkung (Drilldown) ermöglicht wird.

Suchen mit der Möglichkeit, die Ergebnisse über Facetten einzuschränken, sind besonders aus dem Onlinehandel und Bibliothekskatalogen bekannt (Abb. 2). Zentrale Bestandteile im Facettendrilldown Menu sind: Die individuellen Facetten, die Einschränkungswerte jeder Facette und der Zähler (Abb. 3). Durch auswählen bestimmter Kategorien wird das Suchergebnis eingeschränkt, es ändern sich auch

15 Im Projekt Letters 1916–1923: Ordinary Lives–Extraordinary Times wird es auch »visual expiration« genannt: https://letters1916.ie/wp-post/visual-exploration (abgerufen 8.9.2022).

die verfügbaren Kategorien und auch der Zähler wird entsprechend angepasst. Der Zähler gibt an, wie viele Ergebnisse auf eine bestimmte Facette zutreffen und es entsteht dadurch eine Art Vorschau auf den Umfang des Suchergebnisses unter Berücksichtigung des Facettenfilters. Die Kombination aus Kategorien, nach denen gefiltert werden kann, und wie stark die Kategorien im Suchergebnis vertreten sind, ermöglichen es Nutzer:innen, ähnlich wie beim Distant Reading, bereits grundlegende Muster und Abweichungen leicht zu erkennen. Die Möglichkeit Facetten beliebig auszuwählen und zu kombinieren, wodurch sich der Zähler und das Suchergebnis ändern, lädt Nutzer:innen ein zu experimentieren und explorativ mit diesem Werkzeug zu arbeiten und gleichzeitig Erfahrungen über die Inhalte und Zusammensetzung der Daten zu sammeln.

Abb. 2: Online Bibliothekskatalog der Universität Graz. Das Suchergebnis kann über diverse Facetten (links im Bild) gefiltert werden. (Quelle: Eigener Screenshot).

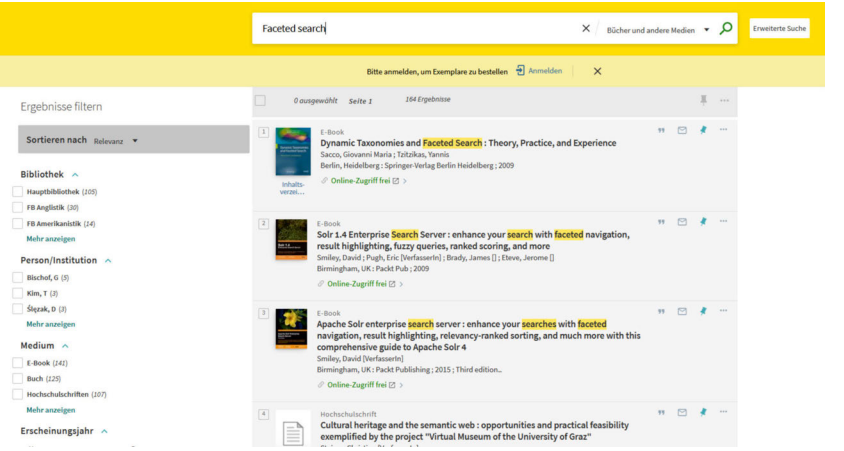
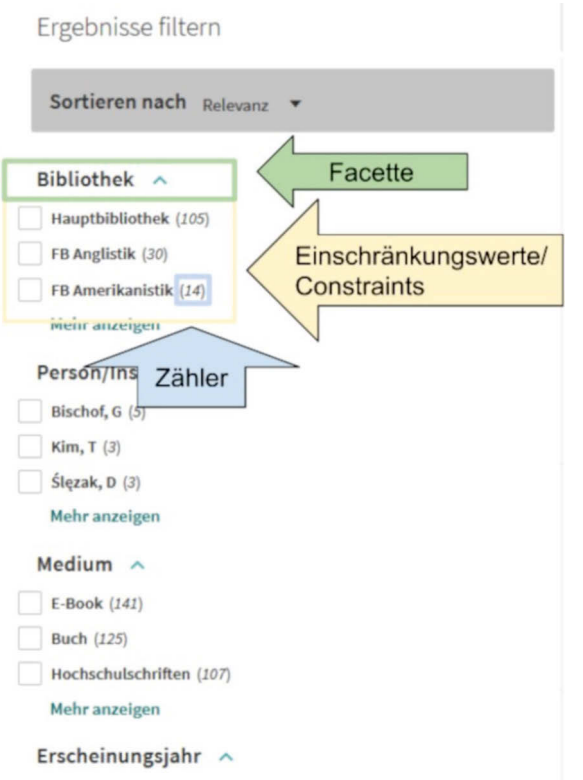


Abb. 3: Vergrößerung des Facettenfilters von Abb. 1 (Quelle: Eigener Screenshot).



GAMS: Standardisierte Abläufe der Digitalen Geisteswissenschaften

Wie bereits zuvor erwähnt, verlangt die computergestützte Analyse im Rahmen der Digitalen Geisteswissenschaften maschinell verarbeitbare Forschungsdaten. Bei der Erstellung eines hochwertigen, maschinen-lesbaren Datensatzes handelt es sich keinesfalls um eine triviale Aufgabe, sondern um einen anspruchsvollen und langwierigen Forschungsprozess, der verschiedenen Stufen der Qualitätssicherung unterliegen muss.[16] Es soll zum Beispiel verhindert werden, dass nach mehrjähriger

16 Vgl. Selina Galka, Modellierung, in: Helmut W. Klug (Hg., unter Mitarbeit von Selina Galka und Elisabeth Steiner), KONDE Weißbuch, 2021, Handle: hdl.handle.net/11471/562.50.137 (abgerufen 8.9.2022). Vgl. auch Martina Bürgermeister, Informationsarchitektur, in: Helmut W. Klug (Hg., unter Mitarbeit von Selina Galka und Elisabeth Steiner), KONDE Weißbuch, 2021, Handle: hdl.handle.net/11471/562.50.97 (abgerufen 8.9.2022).

Forschungsarbeit festgestellt wird, dass sich erhobenes Datenmaterial nicht für die vorgesehene maschinelle Analyse eignet. Die Qualitätssicherung eines IT-Projektes ist nicht selten der aufwändigste und teuerste Teil des Unterfangens (zusammen mit Erhebung der Anforderungen).[17] Selbiges gilt somit auch für Forschungsprojekte im Rahmen der Digitalen Geisteswissenschaften, die auf computergestützte Methoden zurückgreifen.

Um angemessene Qualität zu erreichen, ist das Einhalten von breit etablierten standardisierten Abläufen und Techniken, die – unter anderem – den Austausch von Expertise über stark unterschiedliche Forschungsunternehmungen sicherstellen, notwendig.[18] Diese Gütekriterien werden unter anderem durch das Erstellen von akademisch-hochwertigen »nachhaltigen« Digitalen Editionen eingehalten und umgesetzt.[19] So unterscheiden sich RTA 1576 und DERLA inhaltlich zwar stark, die Projekte teilen sich jedoch dieselben standardisierten Verfahren (und denselben Hintergrund der Geschichtswissenschaft). Folglich könnten generalisierte Analysewerkzeuge wie der Facettendrilldown aus RTA 1576, auch für DERLA wiederverwendet werden.[20]

Am ZIM Graz spielt das digitale Langzeitarchiv GAMS[21] insofern eine zentrale Rolle, da es genannte standardisierte Abläufe für die abgewickelten Editionsprojekte forciert. Viele Aspekte der digitalen Langzeitarchivierung sind für einzelne Forschungsprojekte äußerst schwer zu realisieren. Das geht von kleineren Dingen – wie der Erzeugung von stabilen Webadressen – bis hin zur Anstellung von dauerhaftem Personal, welches der Pflege des abgelegten Datenmaterials verpflichtet ist.[22] Das GAMS liefert diese Funktionalitäten »out-of-the-box«.

---

17 Vgl. Sebastian Stoff, Testen als Qualitätssicherung. In: KONDE Weißbuch. Hg. v. Helmut W. Klug unter Mitarbeit von Selina Galka und Elisabeth Steiner im HRSM Projekt »Kompetenznetzwerk Digitale Edition«,2021, (abgerufen 22.9.2022. Handle: hdl.handle.net/11471/562.50.182. PID: o:konde.182)

18 Wie zum Beispiel die Bereitstellung von Forschungsdaten, vgl. Helmut W. Klug, Bereitstellung von Forschungsdaten, in: Helmut W. Klug (Hg., unter Mitarbeit von Selina Galka und Elisabeth Steiner), KONDE Weißbuch, 2021, Handle: hdl.handle.net/11471/562.50.87 (abgerufen 8.9.2022).

19 Vgl. Lisa Rieger, Digitale Edition, in: Helmut W. Klug (Hg., unter Mitarbeit von Selina Galka und Elisabeth Steiner), KONDE Weißbuch, 2021, Handle: hdl.handle.net/11471/562.50.59 (abgerufen 8.9.2022).

20 Vgl. dieser Artikel, Abschnitt »Facettierter Drilldown«.

21 Startseite des GAMS; Elisabeth Steiner, Johannes Stigler, Zentrum für Informationsmodellierung – Universität Graz, https://gams.uni-graz.at/(abgerufen am 22.09.2022)

22 Online Dokumentation des GAMS, Vgl. Elisabeth Steiner/Johannes Stigler, GAMS and Cirilo Client. Policies, Documentation and Tutorial, 2014 (letztes Update 28.7.2022), https://gams.uni-graz.at/o:gams.doku (abgerufen 8.9.2022) und Nähere Erläuterungen zum GAMS als digitales Repositorium finden sich im Artikel von Selina Galka und Sebastian Stoff im selben Buch.

Folgend werden die zuvor erwähnten einheitlichen Forschungsabläufe der DH und deren Umsetzung am GAMS beschrieben. Diese lauten grob:[23]

1. Datenmodellierung
2. Datenprozessierung
3. Datenanalyse[24]

Datenmodellierung

Jegliches Forschungsprojekt am GAMS braucht ein sogenanntes Datenmodell. Vereinfacht versteht man darunter eine Art einheitliche Schablone für die Erstellung von maschinen-lesbaren Forschungsdaten. Ein simplifiziertes Beispiel hierfür wären zum Beispiel projektübergreifende, einheitliche Personentabellen in denen jeweils die gleichen Spalten (Vorname, Nachname, Alter etc.) für individuelle Personeneinträge definiert sind. Dies dient dazu, um sicherzustellen, dass die maschinelle Weiterverarbeitung für jede Person auch tatsächlich das gleiche (wohl dokumentierte) Resultat liefert.

Die Erarbeitung eines solchen Datenmodells bedarf einer intensiven Zusammenarbeit zwischen der Expertise der Digitalen Geisteswissenschaften und der jeweiligen Domäne, da es die Bedingungen der »maschinellen Welt« mit dem eigentlichen Forschungsinteresse für die spätere Analyse zu verbinden gilt. Beispielsweise könnte als wissenschaftliches Interesse die Frage nach dem Durchschnittsalter der genannten Personentabelle an den Computer gestellt werden. Ist nun für alle Personen in der »Alter-Spalte« eine gültige Zahl eingetragen, würde nun auch ein gültiges Ergebnis produziert. Was passiert aber im Falle von unklaren bzw. ungenauen Einträgen, zum Beispiel wenn das Alter einfach nicht bekannt sein sollte? Sollten nun für die Berechnung alle Personen mit fehlendem Alter ignoriert werden? Welche Aussage besitzt diese Rechnung noch, wenn für dreißig Prozent aller erfassten

23 In Anlehnung an die »Methoden und die Realisierung von Digitalen Editionen« und der konkreten Entwicklung am GAMS – Vgl. Patrick Sahle, Digitale Edition, in: Fotis Jannidis/Hubertus Kohle/Malte Rehbein (Hg.), Digital Humanities, Stuttgart 2017, 241–245, DOI: 10.1007/978-3-476-05446-3_23 (abgerufen 8.9.2022); Johanna Drucker, Visualization and Interpretation: Humanistic Approaches to Display Visualization and Interpretation. Cambridge, Massachusetts 2020.

24 Vgl. Sean Winslow, Gerlinde Schneider, Roman Bleier, Christian Steiner, Christopher Pollin, Georg Vogeler; Ontologies in the Digital Repository: Metadata Integration, Knowledge Management and Ontology-Driven Applications, in: CEUR Workshop Proceedings, JOWO 2019, Joint Ontology Workshops, Graz, 2019, Vol. 2518, 1–8, urn:nbn:de:0074-2518-1, http://ceur-ws.org/Vol-2518/paper-WODHSA11.pdf (abgerufen 29.09.2022)

Personen das Alter nicht recherchiert werden konnte? Solche Entscheidungen verlangen nach einem vertieften Domänenwissen einerseits und Kenntnis über die Bedingungen der maschinellen Auswertung auf der anderen Seite.[25]

Datenprozessierung

Im Schritt der Datenprozessierung verlangt das GAMS neben dem Einspielen der eigentlichen Forschungsdaten (die dem Datenmodell folgen) die Definition von Weiterverarbeitungsschritten für die folgenden Analyseverfahren. Üblicherweise müssen die erhobenen Rohdaten des Projektes von etwaigen Abweichungen bereinigt werden (nach Vorlage des Datenmodells). Hierzu werden eigene Programme geschrieben, die einerseits in der Lage sind, diese Abweichungen auszubessern und die Projekt-Rohdaten in die von GAMS vorgeschriebenen Archivformate zu bringen. Anschließend erlaubt GAMS die Ablage des Datenmaterials im Archivsystem. In einem weiteren Schritt der Datenprozessierung gilt es, das nun langzeitarchivierte Datenmaterial für die Analyseverfahren vorzubereiten. Wieder dem Beispiel der Personendaten folgend, stellt sich die Frage welche Teile der erhobenen Information für welche Analyseverfahren geeignet sind und inwiefern diese weiterverarbeitet werden müssen. Im Falle des Durchschnittsalters müssen die einzelnen Altersangaben pro Person aus den Daten extrahiert werden (und gegebenenfalls ungültige Werte – wie Personen mit fehlendem Alter – übersprungen werden).[26] So wird die für die Analyse notwendige Information für die vom GAMS bereitgestellten Datenbanken zur Verfügung gestellt und für den Schritt der Datenanalyse vorbereitet.[27]

25 Vgl. Fotis Jannidis, Grundlagen der Datenmodellierung, in: Fotis Jannidis/Hubertus Kohle/Malte Rehbein (Hg.), Digital Humanities, Stuttgart 2017,99-108, DOI: 10.1007/978-3-476-05446-3_23 (abgerufen 8.9.2022); Johanna Drucker, Visualization and Interpretation: Humanistic Approaches to Display Visualization and Interpretation. Cambridge, Massachusetts 2020; Selina Galka, Modellierung, in: Helmut W. Klug (Hg., unter Mitarbeit von Selina Galka und Elisabeth Steiner), KONDE Weißbuch, 2021, Handle: hdl.handle.net/11471/562.50.137 (abgerufen 8.9.2022); Vgl. Steiner/Stigler, GAMS-Dokumentation, Abschnitt »Content models in cirilo«

26 Vgl. Harald Klinke, Datenbanken, in: Fotis Jannidis/Hubertus Kohle/Malte Rehbein (Hg.), Digital Humanities, Stuttgart 2017,109-127, DOI: 10.1007/978-3-476-05446-3_23 (abgerufen 8.9.2022); Johanna Drucker, Visualization and Interpretation: Humanistic Approaches to Display Visualization and Interpretation. Cambridge, Massachusetts 2020;

27 Vgl. Steiner/Stigler, GAMS-Dokumentation, Abschnitt »Content models in cirilo«.

Datenanalyse

Die Datenanalyse am GAMS erfolgt primär über die Abfrage der prozessierten Archivdaten in den am System anliegenden Datenbanken. Je nach Anwendungsfall bzw. Fragestellung werden verschiedene Datenbanken angesteuert und anschließend Analysen vollzogen. Je nach angesteuerter Datenbank werden eigene Abfragen formuliert, die die eigentlichen Datenanalysen vollziehen. Im Falle der Personendaten würde beispielsweise eine Abfrage gegen die verwendete semantische Datenbank natürlichsprachlich lauten: »*Zeige alle Personen. Filtere alle Personen, die kein gültiges Alter besitzen. Berechne den Durchschnitt*«. Dieses Ergebnis wird in weiterer Folge einer Routine übergeben, die die grafische Visualisierung steuert – zum Beispiel in Form eines Diagramms.[28]

Case Study: Datenmodell und Abläufe in der Edition Regensburger Reichstag von 1576

An der digitalen Edition des Regensburger Reichstags von 1576 (RTA 1576) zeigen wir exemplarisch, wie die standardisierten Archivierungs- und Publikationsabläufe am GAMS in einem konkreten Projekt angewandt werden. Diese Edition wird seit Mai 2018 in einem DACH Projekt (DFG: 386773508; FWF: I 3446) umgesetzt.[29] Dabei handelt es sich um eine Kollaboration zwischen der Historischen Kommission bei der Bayerischen Akademie der Wissenschaften (HiKo) und dem ZIM. Die Edition ist Teil eines größeren Editionsvorhabens, in dem seit 1986 die deutschen Reichsversammlungen zwischen 1556 und 1662 ediert werden, jedoch entsteht zum ersten mal eine born-digital Reichstagsedition und es werden Erfahrungen im digitalen Edieren frühneuzeitlicher Quellen gesammelt. Ein Schwerpunkt der digitalen Edition ist es, dem Forschungsinteresse an den Prozessen der Entscheidungsfindung[30] in frühmodernen Ständeversammlungen gerecht zu werden. Web of Data Technologien spielen bei der Entwicklung der Edition eine wichtige Rolle und für die Datenmodellierung und -analyse wurde eine domänenspezifische Ontologie geschaffen.[31]

Das Web of Data, manchmal auch als Semantic Web bezeichnet, ist eine Erweiterung des Word Wide Webs. Während es in der ersten Phase des WWW das Ziel

28 Vgl. Steiner/Stigler, GAMS-Dokumentation, Abschnitt »cirilo:Query« oder »cirilo:Context«; Helmut W. Klug, Analysemethoden, in: Helmut W. Klug (Hg., unter Mitarbeit von Selina Galka und Elisabeth Steiner), KONDE Weißbuch, 2021, Handle: hdl.handle.net/11471/562.50.16 (abgerufen 8.9.2022).

29 RTA RV 1576, URL: https://gams.uni-graz.at/context:rta1576 (abgerufen 8.9.2022).

30 Vgl. Gabriele Haug-Moritz, Deliberieren: Zur ständisch-parlamentarischen Beratungskultur im Lateineuropa des 16. Jahrhunderts, in: Historisches Jahrbuch 141 (2021) 114–155.

31 Vgl. Georg Vogeler, The »Assertive Edition«, in: International Journal of Digital Humanities 1.2 (1 July 2019) 309–322, DOI: 10.1007/s42803-019-00025-5 (abgerufen 8.9.2022).

war, Dokumente (z.B. HTML Seiten) miteinander zu verknüpfen, versucht das Web of Data nun Inhalte und Informationen zu verknüpfen mit dem Ziel, diese für komplexere Suchen fruchtbar zu machen. Die Grundlage dafür ist Linked Open Data. Tim Berners-Lee hat ein »5 Star Linked Open Data« Modell beschrieben[32], das besagt, dass nur Daten, die gewisse Kriterien erfüllen, als »true« Linked Open Data (LOD) bezeichnet werden können. Die Kriterien sind: Verfügbarkeit durch eine freie Lizenz, Verfügbarkeit in einem nicht proprietären, maschinenlesbaren Format (am besten als RDF), die Daten müssen über persistente URIs adressierbar sein und durch Links auf andere Daten muss Kontext hergestellt werden. Zum Beispiel können Personendaten in digitalen Editionen durch eine Verlinkung auf Normdateien kontextualisiert und die jeweilige Person identifiziert werden. Im deutschsprachigen Raum ist dabei die Gemeinsame Normdatei (GND) eine wichtige Referenzquelle, die von vielen Projekten genutzt wird. Die Übersetzung der Inhalte von edierten Texten in RDF und die Verknüpfung mit externen Datenquellen bildet einen weiteren Schritt in Richtung Datenorientierung digitaler Editionen und sind auch, nach Vogeler, ein wichtiges Instrument zum Bau von »assertive editions«.

RTA 1576 basiert auf unterschiedlichen Datenquellen: den Kern der Edition bildet die Archivdokumentation, die edierten Texte und Bildobjekte. Die Archivdokumentation stellt ein »virtuelles Reichstagsarchiv« dar,[33] in dem die in 32 Archiven gesammelten Metadaten über relevante Bestände und überlieferte Dokumente zum Reichstag veröffentlicht werden. Die TEI/XML Struktur der AD, in Anlehnung an den Standard Encoding Archival Description (EAD 3.0)[34], die versucht die relevanten Bestände zu beschreiben, ist eine strukturierte Liste von tei:div, für übergeordnete Archivstrukturen, und tei:msDesc, für individuelle Texte. Die Hauptdatenstruktur bleibt TEI/XML und Verweise auf andere Datenmodelle, wie EAD 3.0, werden über das TEI Attribut @ana angeführt.

Die edierten Texte[35] wurden der Tradition der Reichstagseditionen folgend ausgewählt und ediert. Es wurde aber auch kritisch evaluiert, ob gewisse Praktiken für eine digitale Edition überhaupt relevant sind: z.B. wurde die in Druck übliche Praxis, dass Teile von längeren Texten zusammengefasst wurden, nicht beibehalten. Da in der digitalen Edition kein Platzmangel besteht, können Transkriptionen in voller Länge wiedergegeben werden.

32 Vgl. Tim Berners-Lee, Linked Data, 2006, https://www.w3.org/DesignIssues/LinkedData.html (abgerufen 8.9.2022).

33 Archivdokumentation, in: RTA RV 1576, URL: http://gams.uni-graz.at/context:rta1576.ad (abgerufen 8.9.2022).

34 Encoded Archival Description, URL: https://www.loc.gov/ead (abgerufen 8.9.2022).

35 Edierte Texte, in: RTA RV 1576, URL: http://gams.uni-graz.at/context:rta1576.ed (abgerufen 8.9.2022).

Um die Inhalte genauer zu erschließen, wurden Listen für Personen, Orte, Gruppen und Körperschaften und Sachbegriffe angelegt, die sich an den Reichstagsregistern orientieren, aber im Umfang über diese hinausgehen.

Die Ontologien der Pre-modern Parliamentary Communication (PPAC) und erwartete Forschungsergebnisse im Bereich der frühneuzeitlichen Ständeversammlungen werden an anderer Stelle genauer besprochen.[36] Im Kontext dieses Beitrags ist jedoch die Funktion der Ontologie als zentrales Verbindungsglied zwischen den Editionsdaten, aus AD, den edierten Texten und Registern, und der Forschungsdatenbank wichtig.

Basierend auf der PPAC Ontologie werden bestimmte Informationen aus den TEI/XML Daten extrahiert und über eine XSL Transformation in RDF umgewandelt und damit die Forschungsdatenbank, den Triplestore im GAMS, befüllt. Von zentraler Bedeutung ist dabei das bereits erwähnte Attribut @ana, welches das Mapping zur PPAC Ontologie enthält. Dieses Verfahren wird auch bei Vogeler beschrieben und in anderen Projekten in der GAMS umgesetzt.[37] Zu erwähnen sind etwa Editionen von Rechnungsbüchern, für die eine eigene Book Keeping Ontology entwickelt wurde.[38]

Das Projekt RTA 1576 bietet gegenwärtig drei zentrale Möglichkeiten an, die archivierten Daten zu analysieren: Volltextsuche, Register und Facettendrilldown. Die Volltextsuche im RTA 1576 Projekt basiert primär auf einer Datenbankabfrage. Für jeden Textabschnitt (Absatz in edierten Texten oder Stückeintrag in der AD) wird ein Datensatz im RDF geschaffen, der auch den Volltext des jeweiligen Abschnittes enthält. Dabei ist es möglich, in der Datenverarbeitung zu definieren, dass Informationen hinzugefügt werden sollen bzw. Personennamen und Abkürzungen auf-

36 Roman Bleier/Florian Zeilinger/Georg Vogeler, From Early Modern Deliberation to the Semantic Web: Annotating Communications in the Records of the Imperial Diet of 1576, in: Matti La Mela/Fredrik Norén/Eero Hyvönen (Hg.), Proceedings of the Digital Parliamentary Data in Action (DiPaDA 2022) Workshop, co-located with 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), CEUR Workshop Proceedings, Uppsala, Sweden, 2022. 86–100. Ontologie, in: RTA RV 1576, URL: http://gams.uni-graz.at/o:rta1576.ontology (abgerufen 8.9.2022).

37 Georg Vogeler, Warum werden mittelalterliche und frühneuzeitliche Rechnungsbücher eigentlich nicht digital ediert?, in: Constanze Baum/Thomas Stäcker (Hg.), Grenzen und Möglichkeiten der Digital Humanities (Sonderband der Zeitschrift für digitale Geisteswissenschaften, 1), 2015. DOI: 10.17175/sb001_007 (abgerufen 8.9.2022).

38 Vgl. Christopher Pollin, Digital Edition Publishing Cooperative for Historical Accounts and the Bookkeeping
Ontology, in: Thomas Riechert/Francesco Beretta/George Bruseker (Hg.), RODBH 2019, Proceedings of the Doctoral Symposium on Research on Online Databases in History, 2019, 7–14. URL: http://ceur-ws.org/Vol-2532 (abgerufen 8.9.2022); and URL: https://gams.uni-graz.at/context:depcha (abgerufen 8.9.2022).

gelöst werden sollen. Das RDF wird dann in die Datenbank eingespielt und über GAMS Query Objekte können Suchen definiert werden. Für die Volltextsuche wird das bds:search Interface der Blazegraph Datenbank verwendet.[39]

Die Datenanalyse mithilfe des Registers soll kurz anhand des Personenregisters exemplarisch dargestellt werden. Die Grunddaten des Personenregisters sind als eine tei:listPerson modelliert, wobei jede Person einen eindeutigen Identifikator besitzt und neben dem Personennamen noch andere relevante Daten (z.B. Zugehörigkeit zu einer Herrschaft, Information über Anwesenheit auf dem Reichstag, Verweis auf die GND oder andere Normdaten) zugeordnet sind.[40] Basierend auf der zuvor besprochenen Ontologie werden diese Daten zusammen mit den Daten aus den edierten Texten und der AD in den Triplestore eingespielt. Die Verbindung zwischen den Einträgen im Register und der Erwähnung einer Person im edierten Text wird über den Identifikator hergestellt. Dadurch kann z.B. die genaue Anzahl der Erwähnungen einer Person aus der Datenbank ausgelesen werden. Dadurch bekommt ein Nutzer einen Überblick über die Erwähnungen in der Edition. Der Facettendrilldown, im RTA 1576 als Filter-Recherche bezeichnet, bietet eine detailliertere Abfragemöglichkeit, die weit über eine Registerfunktion hinausgeht.

Facettendrilldown

Die Nachfrage nach einem Facettendrilldown entstand in mehreren Projekten am ZIM. Es erschien attraktiv traditionelle Filtermöglichkeiten bei Suchen durch etwas zu ersetzen, das durch die breite Anwendung in Bibliothekskatalogen und Online-Portalen wie Willhaben.at bereits sehr bekannt ist. Als erstes Testprojekt wurde am ZIM ein Facettendrilldown für die RTA Edition entwickelt. Gerade im RTA Projekt gibt es eine große Anzahl an Filterkategorien, die beliebig kombiniert werden können. Grob kann unterschieden werden zwischen inhaltlichen Kategorien, z.B. erwähnte Personen oder Orte, Daten oder Verhandlungsthemen, und dokument-beschreibende Kategorien, z.B. handelt es sich um ein Protokoll oder um einen Bericht eines Gesandten. Das Projekt ist zwar vorwiegend für ein Fachpublikum ausgelegt, aber wie viele digitale Editionen sollen die Forschungsdaten auch interessierten Laien, besonders Schüler:innen und Student:innen, zugänglich sein.

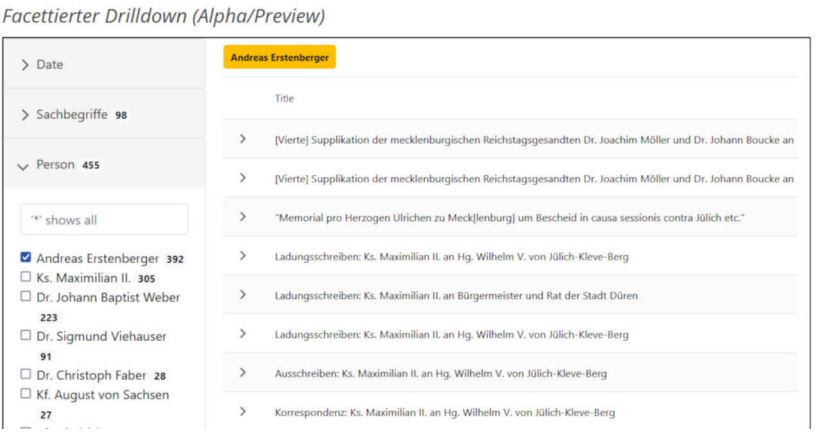
Die Filtermöglichkeit der im Projekt angebotenen Kategorien durch ein Facettendrilldown ermöglicht es beiden Nutzergruppen ein Werkzeug für exploratives Durchsuchen der Projektdaten in die Hand zu geben. Die erste Implementierung

39 Blazegraph Database Platform 2.1.5 API, URL: https://blazegraph.com/database/apidocs/com/bigdata/rdf/store/BDS.html (abgerufen 8.9.2022).

40 Recherche, in: RTA RV 1576, URL: http://gams.uni-graz.at/o:rta1576.bt1734r15t (abgerufen 8.9.2022).

(Abb. 4) macht Nutzer:innen primär inhaltliche Kategorien zugänglich. In Abbildung 4 kann man den Nutzen dieser Filtermethode sehen. Durch die Auswahl eines Personennamens, im Beispiel der kaiserliche Reichs- und Hofsekretär Andreas Erstenberger, erhält man sofort einen Überblick darüber, wie oft die Person mit anderen Personen zusammen in einem Dokument genannt wird. Man könnte durch das Aufklappen der anderen Facetten auch herausfinden, welche Orte und Themen in den Dokumenten erwähnt werden und durch die Auswahl von zusätzlichen Personen, Orten oder Themen das Suchergebnis entsprechend einschränken. Eine Datumseinschränkung ist ebenfalls Teil des Facettendrilldown-Interfaces und ermöglicht es etwa zu untersuchen, welche Themen, Orte oder Personen für bestimmte Verhandlungstage erwähnt werden. In der nächsten Überarbeitungsphase sollen auch dokument-beschreibende Kategorien über das Facettendrilldown Nutzer:innen zugänglich gemacht werden.

Abb. 4: Erste Implementierung des Facettendrilldowns in der Edition RTA 1576 (Quelle: Eigener Screenshot).



Nicht nur die Verwendung der gleichen Kategorien (wie Personen und Orte), sondern bereits die Tatsache, dass kontrollierte Kategorien verwendet werden, erlaubt die Entwicklung gemeinsamer Anwendungen im Rahmen von DH Projekten. Diese Nachnutzung über Projektgrenzen hinweg ist eine zwingende Voraussetzung für die Entwicklung komplexer Software wie dem Facettendrilldown, da nur so der nötige Entwicklungsaufwand – gerade für kleinere Projekte wie RTA und DERLA – bewältigt werden kann. Der Nachnutzungsgedanke findet Ausdruck in verschiedensten Arbeitsgruppen (z.B. Zu gemeinsamen Vokabularien) und gemeinsamen Arbeitsschritten am ZIM (wie projektübergreifende Anforderungsanalysen), in de-

nen der zuvor erwähnte Facettendrilldown als qualitätsvoller, wiederverwendbarer Service entwickelt wurde und betrieben wird.

Folgend würde sich auch DERLA als GAMS-Projekt am ZIM für die Anwendung des Facettendrilldown eignen, da bereits bei der Implementierung gemeinsame Charakteristika geisteswissenschaftlicher Forschungsprojekte berücksichtigt wurden und werden. Auch DERLA besitzt ein System kontrollierter Kategorien, mit einer Vielzahl an mehrfachen Zuweisungen zu einzelnen Entitäten (=Erinnerungsorte), die gerade für fachfremde Interessierte schwierig zu überblicken sind.[41] Der Facettendrilldown wird – wie zuvor erwähnt – durch das angebotene explorative Durchsuchen interessierten Laien einen verständlichen und interaktiven Zugang zu den komplexen Forschungsdaten DERLAs ermöglichen.

Conclusio

Die geisteswissenschaftliche Forschungslandschaft produziert zunehmend komplexe digitale Daten und es sind neue programmatische Zugänge notwendig, um den Nutzer:innen Werkzeuge in die Hand zu geben, mit denen diese Daten ausgewertet werden können. Such- und Filtermöglichkeiten stehen dabei an vorderster Stelle. Das in diesem Kapitel vorgestellte explorative Durchsuchen anhand eines Facettendrilldowns hat den zusätzlichen Mehrwert, dass es einen vergleichsweise niederschwelligen Zugang sowohl für Laien- und Expertennutzer:innen gleichzeitig bietet. Die Entwicklung des Facettendrilldowns, obwohl ursprünglich im Editionsprojekt RTA 1576 getestet, erfolgt am ZIM als ein generisches Service, das letztlich nicht nur von ZIM Projekten nachgenutzt werden kann. Folgend lässt sich gut die weit verbreitete Forderung der Digitalen Geisteswissenschaften nach gemeinsamen Infrastrukturen und umfassenden Kooperationen über Projektgrenzen hinweg wiederholen und anhand der (nötigen und geplanten) Übertragbarkeit des Facettendrilldowns vom RTA Projekt auf DERLA demonstrieren.

41 Vgl. auch Historischer Thesaurus des DERLA Projektes: http://gams.uni-graz.at/o:derla.thesaurus, CJS/ZIM Universität Graz (abgerufen am 22.09.2022)