Gubler, Kaspar

Forschungsdaten vernetzen, harmonisieren und auswerten. Methodik und Umsetzung am Beispiel einer prosopographischen Datenbank mit rund 200.000 Studenten europäischer Universitäten (1200–1800)

Oberdorf, Andreas [Hrsg.]: Digital Turn und Historische Bildungsforschung. Bestandsaufnahme und Forschungsperspektiven. Bad Heilbrunn : Verlag Julius Klinkhardt 2022, S. 127-145

Kaspar Gubler

Forschungsdaten vernetzen, harmonisieren und auswerten: Methodik und Umsetzung am Beispiel einer prosopographischen Datenbank mit rund 200.000 Studenten europäischer Universitäten (1200–1800)

Die Vernetzung von Forschungsdaten zählt seit längerem zu den beherrschenden Themen digitaler Forschung. Dies trifft auch für die Geisteswissenschaften und hier besonders für die historische Forschung zu. Ziel der Datenvernetzungen ist es, aus den Datensammlungen Einsichten und Erkenntnisse zu gewinnen, die erst durch die Verbindungen der Daten möglich werden. Allgemein wird davon ausgegangen, dass der Erkenntnisgewinn umso höher ausfallen wird, je umfangreicher und reichhaltiger die Datensammlungen sind. Ausgewertet werden sie nach gemeinsamen Merkmalen, um auffällige Muster, Zusammenhänge und Entwicklungen zu erkennen. Mittlerweile werden dazu auch in den Geisteswissenschaften verschiedene Analyseinstrumente eingesetzt, besonders für Datenvisualisierungen oder Netzwerkanalysen (Gubler u. a. 2022a, b).[1]

Im Folgenden sollen zunächst einige Prämissen zu den Schwierigkeiten der Vernetzung von Forschungsdaten in den Geisteswissenschaften dargelegt werden. Nach kurzer Vorstellung der herangezogenen zu vernetzenden Datenbankprojekte, wird ein vom Autor und seinen Partnern entwickelter Lösungsansatz vorgestellt, dem ein kleiner Exkurs zur Stellung der digitalen Geisteswissenschaften im Verhältnis zur Informatik folgt. Abschließend werden die Funktionsweisen des Lösungsansatzes an einem Anwendungsbeispiel erläutert.

Projekte zu vernetzten Datensammlungen, die vor allem auf frei verfügbaren Daten Linked Open Data, LOD beruhen, sind auch in den digitalen Geisteswissenschaften populär geworden. In der historischen Forschung etwa wurden solche Projekte bereits angegangen und eingehend die Schaffung von Standards für den Datenaustausch diskutiert (vgl. Beretta & Riechert 2016; Burrows 2017; Randeraad 2019; Dumont 2015; van den Heuvel & Alvarez Frances 2021, Koho u. a. 2021). Vieles bleibt hierbei jeweils auf dem Papier und falls LOD-Projekte

1 Für die kritische Durchsicht des vorliegenden Beitrags und verschiedene Anregungen bedanke ich mich herzlich bei meiner Kollegin Lotte Kosthorst, Padua. Zur historischen Netzwerkanalyse siehe auch: https://historicalnetworkresearch.org (Zugriff: 18.02.2022).

dennoch umgesetzt werden, ist die Strukturtiefe der Vernetzung allgemein gering.[2] Dies liegt aber nicht an der technischen Umsetzung. Schon seit Jahrzehnten ist es möglich, solche Daten zu verknüpfen. Die große Herausforderung ist der inhaltliche Abgleich der Daten aufgrund ihrer Uneinheitlichkeit und Unvollständigkeit, was Auswirkungen auf die Strukturierung der Datenmodelle und die Kodierung der Daten zu ihrer Erhebung hat. Daten in den Geisteswissenschaften sind grundsätzlich wenig einheitlich, da sie auf heterogenen und oft lückenhaften Quellen beruhen, beispielsweise auf Texten des Mittelalters. Als Folge davon erstellten Forschungsprojekte für die Kodierung individuelle begriffliche Ordnungen oder Kategoriensystemen, die keinem Standard angepasst waren. Zwar gibt es mittlerweile Bestrebungen zur Normierung von Forschungsdaten durch die Schaffung gemeinsamer Standards, doch ist die Anwendung solcher noch keine Selbstverständlichkeit.[3] Als ein wichtiges Projekt zur Normierung von Daten und zu ihrer Vernetzung kann heute Wikidata genannt werden, ein Schwesterprojekt von Wikipedia. Wikidata bietet Standards, die als Kategorien oder Klassifikationen angesehen und auf einfache Weise für eine Vernetzung verwendet werden können. Eine naheliegende und praktische Option für ein Forschungsprojekt ist es demzufolge, seine Daten auf Wikidata zu hinterlegen und damit gleichzeitig zu standardisieren und für einen Austausch bereitzustellen. Von Wikidata aus können dann andere Projekte, Archive oder Bibliotheken die Daten abfragen und so auch nachnutzen, unter Umständen die Daten gleich im Sinne einer digitalen Langzeitarchivierung bei sich im lokalen Archiv abspeichern.

Bedingt werden Uneinheitlichkeiten in geisteswissenschaftlichen Daten aber nicht nur durch strukturelle Kriterien, sondern in erster Linie durch ihre Qualität als Informationen, was mit dem Verfahren der Datenerhebung zusammenhängt. Es ist hier entscheidend, inwiefern Daten mit menschlicher oder künstlicher Intelligenz zusammengestellt werden. Daten, die mit menschlicher Intelligenz gezielt erhoben werden, beinhalten alleine dadurch einen höheren Informationswert. Werden die Daten zusätzlich kontextualisiert, wird dieser Wert gesteigert. Daten, die nur mit künstlicher Intelligenz erhoben werden, beinhalten vergleichsweise einen geringen Informationswert (vgl. Kuhlen 2004, 12). Die Daten werden bei der Erhebung mit menschlicher Intelligenz somit erst durch das qualitative Beob-

2 Vergleichsweise würden wir bei einer digitalen Edition von einer geringen ‚Auszeichnungstiefe' der Textstrukturen sprechen.

3 Das Festlegen auf eine Ontologie für die Datenerhebung kann erfahrungsgemäß zu langen Diskussionen führen, besonders wenn sich Projekte an bereits bestehenden Ontologien orientieren. Effizienter kann es deshalb sein, zuerst die Forschungsdaten möglichst nahe an den Quellen und einheitlich zu erheben und sie erst später einem Standard anzupassen. Damit lassen sich Diskussionen über die ‚richtige' Ontologie vermeiden, die ein Projekt verzögern können. Dies basiert auch aus Erfahrungen des Autors im Rahmen seiner Beratungen zur digitalen Umsetzung von Projekten an der Universität Bern, unter anderem Rahmen des Supports für die Datenmodellierung in der virtuellen Forschungsumgebung nodegoat.

doi.org/10.35468/5952-09

achten und Beschreiben zu Informationen von Wert. Die künstliche Intelligenz kann eine vergleichbar präzise Beobachtungen, eine abwägende Gewichtung oder Anreicherung der Daten, nicht erreichen, etwa durch Techniken des *Data Mining* oder *Machine Learning*. Beispielsweise generiert die mit Techniken des *Machine Learning* automatisierte Handschriftenerkennung durch die maschinelle Verarbeitung im Grunde minderwertige Informationen. Erst durch die fachkundige Bearbeitung dieser Textdaten erlangen diese die Qualität von Informationen. Dabei werden die automatisiert transkribierten Texte etwa von Hand kategorisiert oder ausgezeichnet (markiert). Dies bedeutet, dass auch in diesem Bereich, so wie beim Einsatz von künstlicher Intelligenz allgemein, immer noch viel Hand- und Kopfarbeit gefragt ist, um die Datenqualität zu heben.[4]

Für das vorliegende Fallbeispiel wurden vier Datenbankprojekte aus dem Bereich der digitalen Universitäts- und Gelehrtengeschichte gewählt, die sich unter anderem im Zuge der Digitalisierung in Richtung einer kontextualisierten Prosopographie und Wissensgeschichte geöffnet hat. Die Daten dieser Projekte, die bislang noch nicht vernetzt worden sind, beinhalten unterschiedlich dichte Informationen zu Studenten, zu ihrer geographischen Herkunft, zu Studienrichtungen und den Abschlüssen an den Universitäten in Bologna, Padua, Paris sowie an den Universitäten im Alten Reich (1200–1800). Die einzelnen Datenbanken decken unterschiedliche Zeiträume ab, schwerpunktmäßig das Mittelalter und die Frühe Neuzeit. Die Daten werden in den vier Projekten grundsätzlich von Hand erhoben und Automatisierungen nur für Ergänzungen zu den Datenbestände mit weiteren LOD-Daten genutzt. Entsprechend hoch ist die Qualität an Informationen dieser Projekte.[5] Ihre Datenbankmodelle sind dagegen, da sie zudem historisch gewachsen sind, deutlich verschieden. Hinsichtlich der Datenerhebung unterscheiden sich die Projekte von neueren Ansätzen, die hierfür auf künstliche Intelligenz setzen. Die Projekte nahmen als Ausgangspunkt die Quellen- und Forschungslage und erstellten Datenmodelle im Hinblick auf die Beantwortung von konkreten Forschungsfragen. Im Projektverlauf wurden die Datenmodelle, aufgrund der Erfahrungen der Quellenarbeit, fortlaufend angepasst. Die Konzeptionierung des Datenmodells so wie auch die Datenerhebung gründen somit auf

4 Dies verweist auf ein Grundproblem der digitalen Geisteswissenschaften. Wenn wir die künstliche Intelligenz (Informatik) gegenüber dem eigenen Fach zu stark gewichten, können wir zwangsläufig keine wirklich neuen Einsichten gewinnen, weil wir an der Oberfläche verharren müssen. Wir sammeln in einer solchen Konstellation zwar sehr viele Daten, die der Computer aber nicht gleichermaßen wie Forschende mit ihrer menschlichen Intelligenz und ihren Kenntnissen zur Quellen- und Forschungslage beobachten und beschreiben kann. Zielführender ist es, das geisteswissenschaftliche Fach an den Ausgangspunkt einer digitalen Untersuchung zu stellen, die wir mit profunden Kenntnissen zur Quellen- und Forschungslage beginnen und mit digitalen Ressourcen und Auswertungstools vollenden.

5 Die Daten stammen von Projekten zu den Universitäten von Bologna (ASFE), Padua (Padova 1222–2022), Paris (Studium Parisiense) sowie zu den Universitäten im Alten Reich, die das Repertorium Academicum Germanicum (RAG) erforscht.

doi.org/10.35468/5952-09

Annahmen und setzen damit ein profundes Wissen zur Quellen- und Forschungslage voraus. Damit unterscheidet sich ein solche Vorgehensweise von Projekten, die mit künstlicher Intelligenz (KI) eine Untersuchung beginnen oder ganz auf diese setzen.[6] Solch Projekte werden im Grundsatz von der Vorstellung geleitet, auch mit geringem Vorwissen zu Quellen- und Forschungslage, alleine durch das Durchforsten großer Quellenbestände zu Fragestellungen und Datenmodellen zu gelangen, also mehr in einem heuristischen Sinn. Dabei kann es durchaus eintreten, dass wir zum Beispiel in umfangreichen Textkorpora durch die algorithmische Suche auf unvorhersehbare Informationen stoßen und unter diesen Zusammenhänge erkennen. Der Erkenntnisgewinn wird allerdings auch bei diesem Vorgehen umso größer ausfallen, desto besser wir Datenbasis (Quellen) und Kontext (Forschung) zum Textkorpus kennen. Demzufolge muss das Vorgehen, um etwa in der historischen Forschung zu einer Fragestellung zu gelangen, durch die Anwendung digitaler Methoden nicht verändert werden. Es ist nach wie vor zielführend und zweckdienlich, mit solidem Vorwissen, ausgehend von einem Quellenbestand, Forschungsfragen zu ermitteln und aufzuarbeiten. Der Nutzen der digitalen Hilfsmittel liegt vor allem im Faktor Zeit. Wir können uns schneller Überblick über die Quellensituation verschaffen. Dies aber auch nur, wenn die Quellen in ausreichender Dichte und als möglichst fehlerfreie Volltexte vorliegen. Hinzu kommt, dass die mit künstlicher Intelligenz gefundenen Daten nur die Spitze des Eisbergs zeigen. Seinen großen Rest erkennen wir erst durch vertiefte Betrachtungen, die nur möglich sind mit unseren Kenntnissen zu Quellen- und Forschungsstand (vgl. Gubler 2022c). Solches Wissen muss zwingend vorhanden sein, um gerade bei umfangreichen Sammlungen vernetzter Daten den ‚Datenberg' in seiner Gesamtheit zu erkennen sowie Einzelheiten identifizieren und angemessen gewichten zu können. Auch für die Nachnutzung von Forschungsdaten muss besonders der Forschungskontext gut dokumentiert sein, zusätzlich zu den Regeln der Datenkodierung und einigen beispielhaften Datenauswertungen. Die Wichtigkeit der Kontextualisierung wurde auch im vorliegenden Fallbeispiel deutlich. Denn besonders erkenntnisreich sind die Interpretationen der Daten, wenn die beteiligten Projekte ihr Wissen nicht nur zu den Eigenheiten der Daten, sondern vor allem zum Forschungskontext einbringen. Damit sind wir beim Kernpunkt: Nicht die technische Vernetzung von Forschungsdaten ist die Herausforderung, sondern die Vereinheitlichung (Harmonisierung) für eine empirisch fundierte Vergleichbarkeit und kontextualisierte Interpretationen.

Einen Ansatz für die Datenharmonisierung erarbeitete der Autor zusammen mit dem Historiker Geert Kessels und dem Medienwissenschaftler Pim van Bree im Rahmen eines SPARK Projekts des Schweizerischen Nationalfonds (SNF) 2020–

6 Ein exemplarisches Projekt hierzu ist die Vernice Time Machine. Trotz großer Versprechungen zu Nutzen und Potential von KI für die Rekonstruktion der Geschichte Venedigs ist dieses Projekt letztlich an der Komplexität historischer Quellen gescheitert, vgl. Hafner 2019.

doi.org/10.35468/5952-09

2021, einem neu geschaffenen Fördergefäß für die rasche Umsetzung unkonventioneller Ideen. Das Projekt verfolgte zwei wesentliche Ziele. Erstens sollte erreicht werden, dass die an einer Datenvernetzung beteiligten Projekte ihre spezifischen Datenbankstrukturen möglichst nicht verändern müssen und zweitens, dass die Datenharmonisierung erst in einer zentralen Zieldatenbank realisiert wird dank eines dynamischen Imports der Daten. In der Zieldatenbank (Metadatenbank) werden sodann die Daten schrittweise qualitativ harmonisiert und vernetzt. Am Ende können die Daten der Zieldatenbank zudem als interoperable Forschungsdaten als LOD-Daten bereitgestellt werden.

Bevor wir auf das Fallbeispiel der vier Projekte näher eingehen, werden wir vorab allgemein die Methodik und Umsetzung der Datenvernetzung erläutern. Im Prinzip wurde, um Diskussionen über Ontologien vorzubeugen, das Vorgehen umgedreht und mit dem Import der Daten in die gemeinsame Zieldatenbank begonnen. Hierfür wurde ein spezielles Softwaremodul entwickelt, das verschiedene Zwecke erfüllt. Das Modul, als DDI-Modul bezeichnet (DDI für ‚Dynamic Data Ingestion'), erleichtert den direkten (dynamischen) Datenimport aus Datenbanken der beteiligten Projekte durch eine Benutzeroberfläche und macht zudem den Importvorgang für die Projekte transparent (John & Pamkaj 2017; Blasch et al. 2018). Das Modul wurde in die virtuelle Forschungsumgebung nodegoat eingebunden, die für die Verwaltung und Analyse von Forschungsdaten zahlreiche Funktionen enthält. Beispielsweise Visualisierungstechniken für Karten, Netzwerke und Zeitreihen.[7] Zusätzlich hat nodegoat den Vorteil, dass wir bei der Definition eines gemeinsamen Datenmodells für die zu importierenden Daten flexibel sind. Wir können auf Grundlage des Modells von nodegoat, das objektorientiert ist, ein auf unsere Anforderungen angepasstes Datenmodell in der Benutzeroberfläche erstellen und umgehend im Datenbereich mit der Datenerfassung oder dem Import beginnen (vgl. van Bree & Kessels 2015; Gubler 2022c). Für die vier Projekte wurde ein gemeinsames Modell einer Metadatenbank, der Zieldatenbank, erstellt mit separaten Tabellen für die Geodatensätze. Dies ist nur eine von vielen möglichen Umsetzungen. Wir könnten etwa auch vier an die Projekte angepasste Datenmodelle erstellen, ihre Daten importieren und dann harmonisieren oder auch zusammenführen. Eine solche Bearbeitung von Forschungsdaten wird vereinfacht, da mit nodegoat webbasiert und kollaborativ gearbeitet werden

7 Die Integration des Moduls in eine bestehende Umgebung war ein Ziel des SPARK Projekts, um ein möglichst nachhaltiges Modul zu schaffen, das in einen regelmässigen Updatezyklus eingebunden ist. Zu nodegoat vgl. van Bree & Kessels 2013; nodegoat wird seit 2011 von LAB 1100 entwickelt und mittlerweile weltweit von geisteswissenschaftlichen Forschungsprojekten eingesetzt, für beispielhafte Projekten s. https://nodegoat.net/usecases. Die Open-Source-Version von nodegoat findet sich auf GitHub, siehe https://github.com/nodegoat, die nodegoat-Dokumentation: https://nodegoat.net/documentation.

doi.org/10.35468/5952-09

kann, wobei die Forschenden Zugriff auf alle Daten, auch auf solche aus anderen Projekten derselben Umgebung, erhalten können.
Mit dem DDI-Modul können somit aktuelle Daten direkt aus den Datenbanken der Projekte (Quelldatenbanken) gleichsam einem Spinnen-Prinzip abgerufen und in der Zieldatenbank gespeichert werden.[8] Dies hat auch den methodischen Vorteil, dass bereits mit diesem Importvorgang Daten harmonisiert werden können. Dazu werden vor dem Import im DDI-Modul in einer Benutzeroberfläche die Datenfelder der Quelldatenbanken den gemeinsamen Datenfeldern der Zieldatenbank zugewiesen.

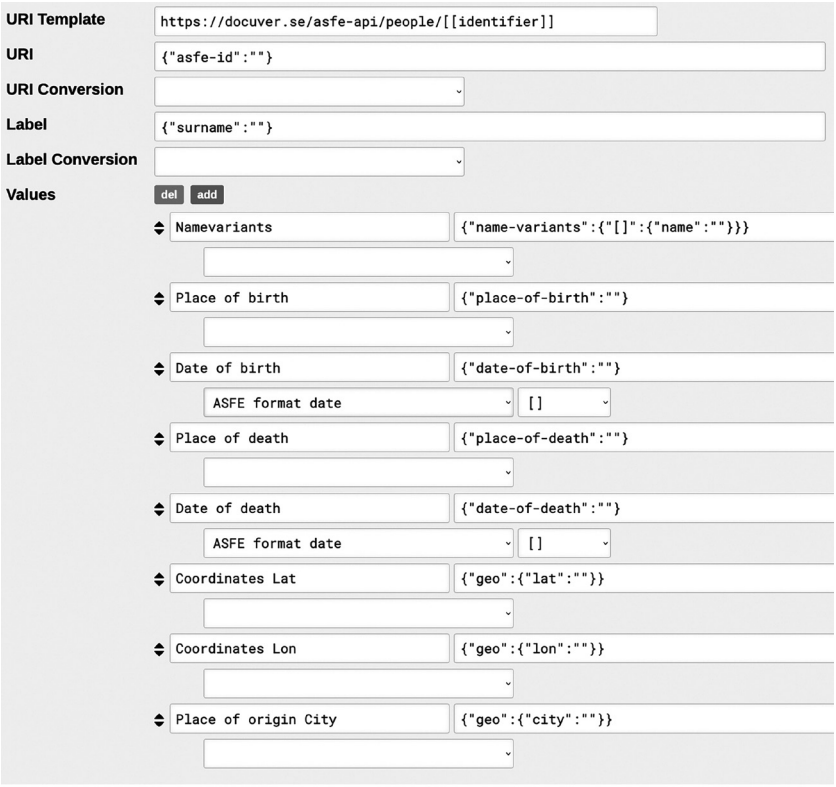

Abb. 1: Datenmapping mit dem DDI-Modul. Datenfelder der Quelldatenbank (rechts) und der Zieldatenbank (links)

8 Vgl. die Dokumentation von nodegoat zu einigen praktischer Anwendungsfällen des DDI-Moduls: https://nodegoat.net/guide.s/132/ingestion-processes (Zugriff: 18.02.2022), vgl. für das in diesem Beitrag vorgestellte Fallbeispiel: Gubler 2021.

doi.org/10.35468/5952-09

Für dieses Datenmapping müssen die Strukturen sowie die Inhalte der Quelldatenbanken möglichst im Detail bekannt sein, um bei diesem Projektschritt bereits Daten harmonisieren zu können. Dies ist somit ein entscheidender Moment, da nun die vier Projekte ihr projektspezifisches Wissen einbringen und eine gemeinsame Ontologie entwickeln. Bei diesem Vorgang sind zwei Aspekte zu beachten, die zum Gelingen einer Datenharmonisierung beitragen. Erstens müssen die Beteiligten einen Überblick wie auch vertiefte Einblicke zu Daten aus den Quelldatenbanken erlangen und zweitens sich darüber einfach verständigen können. Dies kann bei großen und komplexen Datenbanken zur Herausforderung werden. Aus diesen Gründen wurde für das DDI-Modul eine Benutzeroberfläche entwickelt, mit der Datensätze zu Testzwecken abgefragt und das Ergebnis, also die Struktur eines Datensatzes und die Inhalte der Datenfelder, übersichtlich dargestellt werden können.[9] Damit wird für alle an einem Datenmapping beteiligten Projekte vor dem Datenimport deutlich, welche Datenfelder in welcher Zusammensetzung vorhanden sind. Falls dabei die Strukturen der Datenfelder von Quell- und Zieldatenbank nicht zu stark voneinander abweichen, können mit dem Datenmapping entsprechend viele Daten bereits durch den Import harmonisiert werden. Dies auch deshalb, da der Import nicht auf bestimmte Datentypen begrenzt ist. Wir können zum Beispiel Daten zu Personen, Institutionen, Texten oder geografischen Koordinaten importieren. Da sich die Daten nach dem Import in einer Forschungsumgebung befinden, können wir, wie erwähnt, bei guter Datenkonsistenz mit den Auswertungen fortfahren oder die Daten weiter bearbeiten. Sind aber die Unterschiede zwischen Quell- und Zieldatenbank zu groß, können wir weitere Module für die Harmonisierung der Daten einsetzen. Dazu gehört das Modul für eine nachträgliche Kategorisierung (Markierung) der Daten, die von Hand oder automatisiert durchgeführt werden kann. Dieses Modul, ‚Reversals' genannt, funktioniert so, dass auf Grundlage einer Abfrage (Datenfilter), die wir in der Benutzeroberfläche erstellt haben, die entsprechenden Datensätze mit einem frei wählbaren Begriff (der Bezeichnung des Reversals) markiert werden. Die Reversals können wir folglich als Markierungen, Auszeichnungen, Kategorien, Attribute oder auch als ‚Tags' ansehen, abhängig vom Forschungsinteresse. Wir können sie für jedes Datenbankobjekt verwenden und so etwa Personen, Orte, Beobachtungen kategorisieren oder Texte auszeichnen. Mit den Reversals ist es schließlich auch möglich, die Daten nachträglich zur Erhebung zu klassifizieren. Die Reversals werden damit zu einem vielseitigen Werkzeug für einen datengesteuerten Zugriff, der uns die Daten ihrer Zusammensetzung überblicken oder auch Einzelheiten genauer erkennen lässt. Speziell bei großen Datenmengen ist dies hilfreich. Weiter können wir die Reversals einsetzen, um Ordnungen in Datenbestände zu bringen oder die Datenkonsistenz automatisiert zu überprüfen,

9 Bei den Abfragen handelt es sich um SPARQL/API-Abfragen, die via grafische Oberfläche und Schnittstelle in nodegoat auf die Quelldatenbanken zugreifen.

doi.org/10.35468/5952-09

indem fehlerhafte Datensätze markiert werden. Auch für qualitative Kommentare zu den Daten können wir die Reversals nutzen. Zum Beispiel, um auf Unsicherheiten oder Mehrdeutigkeiten in den Daten hinzuweisen oder Daten von Auswertungen auszuschließen.[10]

Zusätzlich zu den Reversals haben wir mit dem Reconciliation-Modul ein weiteres Instrument für die Datenharmonisierung zur Verfügung. Dieses Modul können wir für den automatisierten Abgleich von Daten (*Data Reconciliation*) oder auch für deren Anreicherung verwenden. Außerdem können wir es für Datenchecks einsetzen, um etwa nach einem Import die Daten in Quell- und Zieldatenbanken zu vergleichen und so zu prüfen, ob die Daten korrekt übertragen worden sind. Das Modul funktioniert im Grunde so, dass wir Referenzwerte, etwa ein gemeinsames Vokabular, mit den Daten unserer Datenbank abgleichen lassen und bei Übereinstimmung die Treffer, also die Begriffe eines Vokabulars, im entsprechenden Datensatz speichern können. Der Algorithmus sucht folglich nach vordefinierten Mustern (Begriffen) gemäß eines *pattern matching*. Er ist damit nicht so eingestellt, von sich aus neue Muster zu entdecken.[11] Nach diesen mehr allgemeinen Betrachtungen zu den Methoden und Werkzeugen der Datenharmonisierung kommen wir zu den wichtigen Punkten der praktischen Umsetzung. Vorab einige technische und inhaltliche Bemerkungen.

Die Testreihen des SPARK Projekts, bei denen Daten aus unterschiedlichen Quellen mit dem DDI-Modul in nodegoat importiert wurden, haben erwartungsgemäß zwei wesentliche Herausforderungen von Linked Open Data (LOD) gezeigt. Bereits auf der technischen Ebene wird der Import erschwert durch fehlende Standards: Es mangelt an verbindlichen Vorgaben für die Ausgestaltung und Dokumentation einer Schnittstelle sowie für Daten- und Strukturformate offener Daten, die dann via diese Schnittstelle etwa für eine eine Vernetzung oder für eine Nachnutzung bereitgestellt werden. Bisweilen fehlt auch eine aussagekräftige Dokumentation zu beidem, zu den technischen Angaben zur Schnittstelle wie zur Datenausgabe. Man kann dies alles in technischer Hinsicht als eigentlichen Wirrwarr bezeichnen, durch den man sich zuweilen kämpfen muss, um an die Daten für eine Vernetzung zu gelangen. Neben diesen technischen Hürden kommen die erwähnten, inhaltlichen Hemmnisse aufgrund der Besonderheiten historischer Quellen und des ‚Eisberg-Effekts' hinzu. Es liegt an den erwähnten Gründen, dass trotz zahlreicher Initiativen und kräftiger Forschungsförderung noch kein Projekt der Geisteswissenschaften bahnbrechende Erkenntnisse durch die Vernetzung von Linked Open Data gewinnen konnte. Ein

10 Dies ist nur eine Anwendungen von nodegoat für die Erfassung und Visualisierung unsicherer oder mehrdeutiger Daten. Vgl. dazu die Serie von Blogposts: https://nodegoat.net/blog.s/42/how-to-store-uncertain-data-in-nodegoat (Zugriff: 10.02.2022). Mit Beispielen zum Umgang mit unsicheren, widersprüchlichen oder mehrdeutigen Daten sowie mit unvollständigem Quellenmaterial.

11 Eine Funktion, die aber bei Bedarf im Rahmen der Weiterentwicklung des Moduls hinzugefügt werden könnte.

doi.org/10.35468/5952-09

weiterer Grund hierfür liegt vor allem in der Art der Kommunikation zwischen den Geisteswissenschaften und der Informatik zu verorten, was zugleich ein Licht auf das Rollenverständnis beider Disziplinen wirft. Je nachdem, welche der Disziplinen in einem Projekt dominiert, werden bei den Geisteswissenschaften mehr inhaltliche Schwierigkeiten, insbesondere die Uneinheitlichkeit oder Mehrdeutigkeit von Daten, diskutiert. In der Regel bleibt es in solchen Konstellationen bei Absichtserklärungen, ohne Datensätze exemplarisch und konkret zu vernetzen. Dominiert die Informatik, werden zwar zuweilen große Mengen an LOD-Daten in einer Datenbank gesammelt, die Qualität der Daten und der Forschungskontext aber zu wenig berücksichtigt. Die Betrachtungen und Interpretationen zu solchen Datensammlungen bleiben, aus Sicht der Geisteswissenschaften, letztlich an der Oberfläche und ergehen sich bisweilen in Diskussionen, inwiefern solche Daten bereits als Informationen angesehen werden können oder nicht. Dies ist durchaus eine wichtige Frage, da LOD allgemein als Teil eines „Semantic Web" als wegweisend für die Zukunft eines „Internet des Wissens" gepriesen wird, eines Internets, in dem die Nutzer:innen LOD-Daten auf strukturierte und standardisierte Weise abrufen und in Informationen und Wissen umwandeln können. Wir sind, namentlich in den digitalen Geisteswissenschaften, aber noch weit davon entfernt, um mit solchen Daten wirklich zu neuen Erkenntnissen zu gelangen. Im besten Fall können wir Datenbestände eigener Forschungsprojekte mit LOD-Daten „semantisch anreichern", wie es jeweils verheißungsvoll genannt wird, doch ist das bloße Hinzufügen (Anreichern) von Daten noch keine Analyse, die dann in der Regel ausbleibt oder wenig bringt, wenn wir die Inhalte und die Qualität der Daten nicht angemessen bewerten und im Forschungskontext einbetten können. Auch in diesem Punkt verweisen Datenvernetzungen darauf, die Geisteswissenschaften gegenüber der Informatik etwas hervorzuheben, indem sie mehr ins Zentrum einer Untersuchung zu stellen sind, die, ausgehend von Quellen und Forschungskontext mit der Hilfe digitaler Ressourcen und Werkzeuge durchgeführt wird. Nicht zuletzt aufgrund der erwähnten Faktoren, der Verbesserung der Kommunikation zwischen beiden Disziplinen und der Rückenstärkung der Geisteswissenschaften, wurde das DDI-Modul entwickelt. Als Hilfestellung für die Geisteswissenschaften wurde für das Modul eine Benutzeroberfläche entwickelt, mit der ohne Programmierkenntnisse Schnittstellen für den Datenimport aus Quelldatenbanken definiert und eingerichtet werden können. Wie bereits gezeigt, kann danach, ebenfalls in der übersichtlichen Benutzeroberfläche, eine Quelldatenbank via eine solche Schnittstelle abgefragt und das Ergebnis der erwähnten Testabfrage für die Zuweisungen der Datenfelder beim Mapping von Quell- zu Zieldatenbank verwendet werden. Mit dem Ergebnis der Testabfrage erhalten wir, wie angetönt, eine Übersicht zur Ordnung und zum Inhalt der Daten, die damit bereits auch einen Eindruck geben können, ob eine Vernetzung der Daten überhaupt lohnt im Hinblick auf mögliche Einsichten, sei es zu bestimmten Forschungsfragen oder aus den vernetzten Daten selber. Grundsätzlich unterstützt damit das DDI-Modul auch die Kommunikation zwischen den Geistes-

doi.org/10.35468/5952-09

wissenschaften und der Informatik. Damit funktioniert das Modul als eine Art Übersetzungstool einer visuellen Kommunikation, das Missverständnissen vorbeugen soll, die nicht selten vorkommen, falls allzu abstrakt über Daten gesprochen wird.


```
"info": "Welcome to the nodegoat Data API. For more information visit https://documentation.nodegoat.net/API.",
"timestamp": "2022-02-21T08:50:15+00:00",
"authenticated": true,
"data": {
        "objects": {
                "8194479": {
                        "object": {
                                "nodegoat_id": "ngOk7N97xNUR9NaCgOIBzZWylN17yekSiNB",
                                "object_id": 8194479,
                                "object_name": " Johannes Vliet di Nikolaus da Leida",
                                "object_name_plain": null,
                                "object_name_style": [],
                                "object_style": [],
                                "object_sources": [],
                                "object_version": "",
                                "object_dating": "2021-09-11T13:15:54Z",
                                "object_locked": null
                        },
                        "object_definitions": {
                                "16275": {
                                        "object_description_id": 16275,
                                        "object_definition_ref_object_id": null,
                                        "object_definition_value": "Johannes Vliet di Nikolaus da Leida",
                                        "object_definition_sources": [],
                                        "object_definition_style": []
                                },
                                "16280": {
                                        "object_description_id": 16280,
                                        "object_definition_ref_object_id": 7742264,
                                        "object_definition_value": "M",
                                        "object_definition_sources": [],
                                        "object_definition_style": []
```


Abb 2: Beispiel einer Testabfrage mit dem DDI-Modul zu einem Personendatensatz aus dem Padua-Projekt.

Die Übersetzungsfunktion führt uns wieder zur Frage nach dem Selbstverständnis und damit grundsätzlich auch zur Stellung der digitalen Geisteswissenschaften im Forschungsprozess. Die entscheidende Bedingung für eine erfolgreiche Umsetzung digitaler Projekte in den Geisteswissenschaften ist im gegenseitigen Verstehen von Geisteswissenschaften und Informatik zu sehen und nicht etwa in der Fähigkeit der Geisteswissenschaften, selber programmieren zu können, da dies heutzutage für die Informatik selbst eine enorm hohe Herausforderung darstellt, besonders aufgrund der sich rasch wandelnden Neuerungen bei den Programmiersprachen. Allerdings müssen die Geisteswissenschaften zwingend mit den grundlegenden Prinzipien von Programmierung sowie den relevanten Instrumenten für die Datenanalysen vertraut sein, um mit den Daten zu neuen Erkenntnissen gelangen zu können. Die digitalen Geisteswissenschaften sollten demzufolge eine übersetzende Schnittstellenfunktion in einem Forschungsprozess einnehmen. Ein Prozess, in dem besonders (aber nicht ausschließlich) durch die Verwendung digitaler Ressourcen und der Anwendung von Auswertungstools neue Erkenntnisse gewonnen werden können. Es wäre demnach

doi.org/10.35468/5952-09

treffender von datenbasierten Geisteswissenschaften zu sprechen, um nicht analoge Daten aus einem Forschungsprozesse auszuschließen. In dieser Sicht nehmen die digitalen Geisteswissenschaften eine Position vergleichbar der Wirtschaftsinformatik ein, die interdisziplinär und anwendungsorientiert die Kompetenzen der Disziplinen vereinigt. Die Stärke der Wirtschaftsinformatik ist eine integrative Betrachtung von Informations- und Kommunikationssystemen, namentlich im Rahmen der Konzeption und Abwicklung digitaler Projekte. Sie nimmt die zentrale Übersetzungsfunktion zwischen der Informatik und den Anforderungen von Unternehmen ein, vor allem bei der Entwicklung von Informationssystemen, bei denen aus Daten Informationen und Wissen generiert werden (vgl. Leimeister 2021, 137–205). Nichts anderes als vielschichtige Informationssysteme haben wir in historischen Quellen vor uns. Diese beinhalten aber grundsätzlich ‚weiche' Daten, die zudem unter variablen politischen, wirtschaftlichen, sozialen wie kulturellen Bedingungen entstanden sind und dadurch schließlich unterschiedliche mentale Einstellungen, Wahrnehmungen und Weltanschauungen der Zeitgenossen beinhalten können, welche es bei der Datenerhebung und der Kontextualisierung zu berücksichtigen gilt.

Kommen wir nun aber wieder zur Fallstudie und zu ihrer praktischen Umsetzung. Die Studie wurde erstellt für eine Präsentation im Rahmen der Jahrestagung des Atelier Heloïse, die im März 2021 virtuell in Bologna abgehalten wurde. Das Atelier ist ein internationaler Verbund datenbasierter Forschungsprojekte mit einem gemeinsamen Interessen an einer Geschichte der Universitäten vom europäischen Mittelalter bis in die Neuzeit. Ein methodischer Schwerpunkt der Projekte bildet eine kontextualisierte Prosopographie im Rahmen der Wissenschafts- und Wissensgeschichte.[12] Für das SPARK Projekt wurden aus dem Atelier vier Projekte mit vergleichbarer Ausrichtung ausgewählt, aus denen die Daten mit dem DDI-Modul vernetzt werden sollten.[13] Die Zieldatenbank, die mit den Daten der Projekte erstellt wurde, enthält gegen 200'000 Datensätze zu Studenten und Absolventen aus Europa für den Zeitraum von 1200 bis 1800, wobei, wie erwähnt, die Projekte unterschiedliche Zeiträume abdecken: das ASFE zu Bologna 1500–1800, das Studium Parisiense zu Paris 13.-16. Jahrhundert, das Projekt zu Padua 13.-20. Jahrhundert sowie das RAG 1250–1550. Die Datendichte ist somit für das 14.-16. Jahrhundert am höchsten. Zwingende Voraussetzung für den dynamischen Import der Daten ist, dass die vier Projekte diese im Format JSON über eine Schnittstelle bereitstellen. Bislang können mit dem DDI-Modul somit nur Daten im JSON-Format importiert werden. Allerdings hat dieses Format, das sich besonders für strukturierte Daten eignet, mittlerweile eine hohe Verbreitung erreicht. Anders formatierte Daten kön-

12 Weitere Informationen zum Atelier wie auch zur Konferenz in Bologna auf der Projektwebsite: https://heloise.hypotheses.org (Zugriff: 18.02.2022).

13 Die vier Projekte wurden zufällig ausgewählt, um einen Prototyp zu erstellen; es gibt zahlreiche weitere wichtige Datenbankprojekte im Atelier, die nun mit dem DDI-Modul ebenfalls ihre Daten zusammenführen oder in eine gemeinsame Datenbank einspeisen können.

doi.org/10.35468/5952-09

nen zudem ebenfalls importiert werden, dies aber nicht im DDI-Modul, sondern via Schnittstelle, was Vertrautheit mit der Programmierung voraussetzt. Optional können Daten aber auch von Hand im CSV-Format via Benutzeroberfläche hochgeladen und so importiert werden. Für den dynamischen Datenimport wird nun wie folgt vorgegangen. Zuerst muss im DDI-Modul für jedes der vier Projekte eine eigene Schnittstelle, Linked-Data-Ressource genannt, definiert werden.

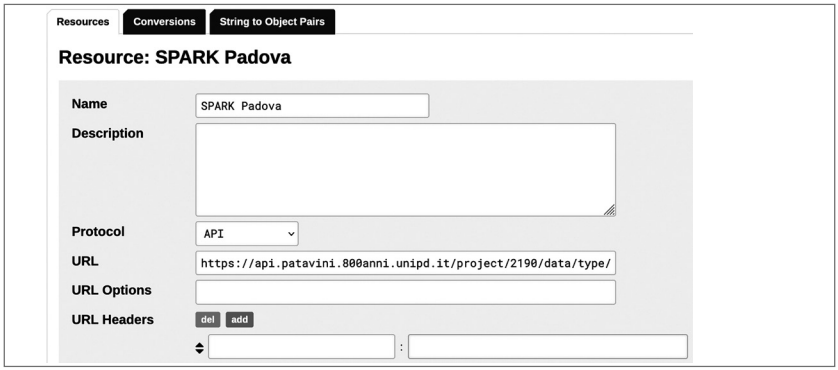

Abb 3: Definition einer Linked-Data-Resource im DDI-Modul

Dazu gehört die Angabe, ob die Daten von einer API-Schnittstelle oder von einem SPARQL-Endpunkt importiert werden sollen.[14] Anschließend wird die Testabfrage für einen bestimmten Datensatz definiert, beispielsweise für eine Person. Dazu wird der Identifikator einer Person, zum Beispiel die GND-Nummer, die in der Quelldatenbank hinterlegt ist, als Variable in die Testabfrage eingefügt. Das Ergebnis der Testabfrage wird im DDI-Modul, wie erwähnt, übersichtlich angezeigt und gibt ein deutliches Bild davon, wie die Daten der Quelldatenbank organisiert sind. Die Testabfrage bildet damit die Vorlage für das Datenmapping, für die Zuweisung der Datenfelder also von der Quell- zur Zieldatenbank. Falls die Datenfelder beider Datenbanken eine hohe Übereinstimmungen aufweisen, können, wie erwähnt, viele Daten bereits mit dem Mapping harmonisiert werden. So wie etwa bei den Namensangaben zu den Studenten der vier Projekte, die zusammen mit den Namensvarianten (andere Schreibweise oder latinisierte Namen) in ein neues Namensfeld der Zieldatenbank importiert wurden. Damit wurde der erste Schritt für die datenbankübergreifende Suche gemacht. Als weitere biographische Angaben wurden die Herkunftsorte der Studenten importiert, welche in den Matrikeln oder in anderen Akten der Universitäten überliefert sind, in der Regel mit Datumsangaben. Da diese in den Projekten nicht durchgehend einheitlich formatiert sind, mussten sie

14 Die Daten müssen nicht zwingend ohne Zugangsbeschränkung verfügbar sein, da im DDI-Modul Login-Angaben für eine Schnittstelle angegeben werden können.

doi.org/10.35468/5952-09

entsprechend konvertiert werden. Dazu wurde das DDI-Modul mit einer speziellen Funktion ausgestattet, die es ermöglicht, Datumsangaben der Quelldatenbanken vor dem Importvorgang automatisiert in ein passendes Format zu konvertieren. Dies geschieht mit einem Skript (Javascript), das im DDI-Modul erfasst und anschließend beim Mapping für das zu konvertierende Datenfeld ausgewählt wird. Mit solchen Skripts können aber nicht nur Datumsangaben, sondern beliebige Daten konvertiert werden. Da Javascript eine einfache, aber zugleich mächtige Skriptsprache ist, gibt es entsprechend zahlreiche Anwendungsmöglichkeiten, um auch an diesem Punkt des Imports die Daten weiter zu harmonisieren. Für die Projekte war die einheitliche Formatierung der Datumsangaben zwingend, um die Herkunftsorte auf einer Karte dynamisch visualisieren zu können im zeitlichen Ablauf. Dadurch ist erkennbar, wie sich die Herkunftsräume aufbauen, verändern oder sich auch gegenseitig bedingen. Die Geodatensätze zu den Herkunftsorten wurden für die Harmonisierung zuerst in separate Tabellen der Zieldatenbank importiert und anschließend diese Tabellen automatisiert mit einer Referenztabelle mit dem Reconciliation-Modul abgeglichen. Da die Orte in den Quelldatenbanken keine Identifikatoren enthalten, mussten mit dem Algorithmus des Moduls die Namen der Orte nach Gleichheit oder Ähnlichkeit einander zugeordnet werden. Dieses Vorgehen wurde auch deswegen gewählt, da das Projekt Studium Parisiense nur die Namen der Herkunftsorte in seiner Datenbank hinterlegt hat, aber keine geografischen Koordinaten. Für diesen Fall, der in Forschungsprojekten immer wieder auftritt, ist ein solcher automatisierter Abgleich mit einer Referenztabelle, die Koordinaten und Identifikatoren enthält, eine effektive Methode, um die Geodaten des eigenen Projekts mit den entsprechenden Angaben zu ergänzen und sie dadurch interoperabel zu machen und visualisieren zu können. Bei Geodatensätzen können ungenaue Angaben eine Herausforderung für die Darstellung auf Karten sein, Angaben wie etwa ‚Diözese' oder ‚Region'. Hierzu wurde im SPARK Projekt ein pragmatisches Vorgehen gewählt und die ungenauen Angaben ebenso als exkate Punkte visualisiert. Auf den Karten werden diese Punkte zur Interpretationshilfe farblich oder mit einem Zeichen hervorgehoben. Nachdem der Abgleich zu den geografischen Angaben aller Projekte fertiggestellt war, mussten die Ergebnisse auf fehlerhafte Zuweisungen durch den Algorithmus geprüft werden, die bei identischen oder ähnlichen geografischen Bezeichnungen vorkommen können. Vereinfacht wurde diese Kontrolle dadurch, dass nun sämtliche Orte auf einer Karte visualisiert werden und damit Fehler rasch entdeckt werden konnten. Sind dann die fehlerhaften Zuweisungen bereinigt, ist eine solche Karte von großem Nutzen, da sie interaktiv ist: Per Mausklick auf einen Punkt auf der Karte, der zum Beispiel für einen Ort, für eine Institution, für eine Person oder auch für ein Werk steht, können die mit diesem Punkt verknüpften Informationen (aus anderen Datenbankobjekten) in Listenform angezeigt oder bearbeitet werden. Diese interaktiven Karten ermöglichen damit visuelle Datenexplorationen- oder analysen auf unterschiedlichen Ebenen und zu beliebigen Themen, abhängig von unserem Datenmodell (vgl. Keim 2004). Der

doi.org/10.35468/5952-09

Erkenntnisgewinn ist am größten, wenn, wie erwähnt, die beteiligten Projekte auch solche visuellen Analysen gemeinsam durchführen, da sie in ihren Daten Muster, Entwicklungen und Zusammenhänge am besten erkennen. Wie bei den Karten, so kann mit dem Modul der Netzwerkanalyse gleichermaßen explorativ vorgegangen werden und die Daten damit mit einer weiteren Visualisierungstechnik betrachtet werden, worauf wir hier aber nicht weiter eingehen können.[15]

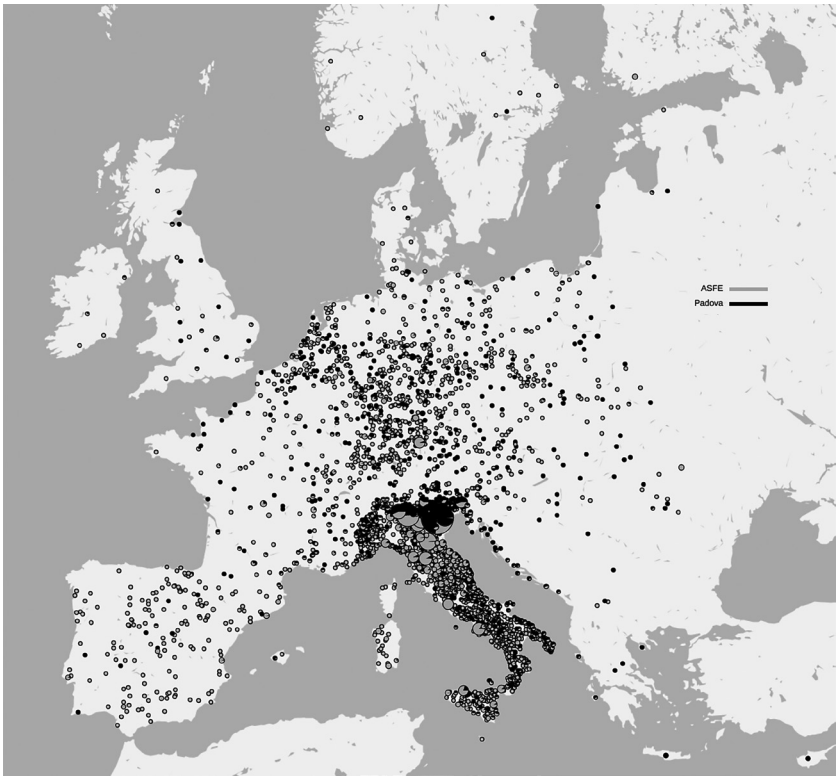

Abb. 4: Herkunftsräume der Studenten der Projekte zu den Universitäten in Bologna (grau) und Padua (schwarz) (Datenbreich 14.-17. Jh.). Mit einer Darstellung in Farbe ließen sich die Herkunftsräume mehrerer Universitäten auf einer Karte abbilden.

Die bisher importierten Daten zu den Namen der Studenten und ihrer geografischen Herkunft waren einheitlich genug, um sie mit dem Mapping oder dem

15 Auf die Netzwerkanalyse können wir an dieser Stelle aus Platzgründen nicht weiter eingehen. Siehe einige Arbeiten zur Anwendung in der Bibliographie zu nodegoat: https://www.zotero.org/lab1100/tags/nodegoat/items/9EDWQCYM/library (Zugriff: 18.02.2022).

doi.org/10.35468/5952-09

Datenabgleich (Reconciliation) harmonisieren zu können. Weichen dagegen Strukturen und Inhalte von Quell- und Zieldatenbank zu sehr voneinander ab, können die Daten nach dem Import mit den beschriebenen Reversals harmonisiert werden. Die Reversals eignen sich somit insbesondere bei Daten, für deren Harmonisierung die Quelldatenbanken ihre Strukturen und/oder ihre Datenausgabe (Schnittstellen) in größerem Umfang anpassen müssten. Im SPARK Projekt wurden Reversals etwa dazu verwendet, die Abschlüsse (Graduierungen) der Studenten zu harmonisieren. Die Graduierungen wurden von den Projekten teilweise verschieden erhoben und bezeichnet. Zudem kann die Bedeutung eines akademischen Grades nach Universitäten variieren, dies besonders im europäischen Vergleich. Das Graduierungssystem der Universitäten im Alten Reich etwa ist zwar bemerkenswert einheitlich, unterscheidet sich aber punktuell von den Systemen der Universitäten in Frankreich oder Italien.[16] Wenn wir nun bei den Graduierungen eine Vergleichbarkeit mit den Reversals erreichen wollen, empfiehlt es sich, wie auch sonst bei den Datenanalysen, vom Allgemeinen zum Besonderen zu gehen. Dazu stehen uns Selektions- und Visualisierungstechniken zur Verfügung, mit denen wir Teilmengen der Daten zusammenstellen und aus unterschiedlichen Perspektiven betrachten können. Diese Techniken können wir nun in Kombination mit den Reversals nutzen. Da wir bei der Definition der Reversals frei sind, zeigt das folgende Vorgehen nur eine von vielen möglichen Anwendungen. Um uns einen ersten Überblick über die Daten zu verschaffen, markieren wir zuerst alle Juristen der Projekte mit einem Reversal ‚Jurist' und einem Akronym des jeweiligen Projekts, zum Beispiel ‚Jurist – RAG'. Anschließend verfeinern wir die Reversals und dringen damit tiefer ins Datenmaterial vor. Wir können dazu etwa den Studienort im Reversal angeben und so einen Juristen im RAG, der in Bologna studiert hatte, entsprechend mit ‚Jurist – RAG – Bologna' kategorisieren. Gleich verfahren wurde für die anderen Projekte. Diese Reversals können wir nun bereits für Datenfilter nutzen und damit alle Juristen in Listenform betrachten oder die Daten auch exportieren. Auf dieselbe Weise wie die Juristen können wir die Graduierungen nach Projekten mit Reversals kategorisieren, etwa einen Juristen im RAG mit entsprechender Promotion an der Universität Basel mit ‚RAG – Promotion Dr. iur. – Basel' kategorisieren, oder eben auch Unterschiede oder Übereinstimmungen zwischen den Graduierungen nach der Methode der Datenaufnahme oder nach den Quellen der Universitäten entsprechend markieren. Da wir nun solche Reversals auch als Grundlage für die Kategorisierung weiterer Reversals verwenden können, ist eine Vielzahl an Varianten und Zugängen für die

16 Diese Gegenüberstellung verweist im Übrigen auf eine Quellensituation, die ein anschauliches Beispiel einer heterogene Überlieferung ist: Im europäischen Mittelalter sind Universitätsmatrikeln praktisch nur für das Gebiet des Alten Reiches überliefert, jedoch nicht für wichtige Bildungsregionen wie Frankreich oder Italien. Erst in der Frühen Neuzeit kommen Universitätsmatrikel auch außerhalb des Alten Reiches auf. Zu den Gründen dieser Überlieferung vgl. Schwinges 2020.

doi.org/10.35468/5952-09

Gruppierungen der Daten bis zur Feingliedrigkeit möglich. Beispielsweise, indem wir die Personen nicht nur nach den belegten Fachrichtungen (römisches Recht, Kirchenrecht) kategorisieren, sondern tiefer vordringen und auch ihre Werke (Texte) oder Korrespondenzen nach bestimmten Begriffen mit Reversals markieren, sei es nach einem Vokabular oder auch nach Zeiträumen.
Die Reversals sind, wie erwähnt, nicht begrenzt auf gewisse Objekt- oder Feldtypen. Zudem erleichtern sie den Zugriff auf die Daten im Rahmen der Datenanalyse grundsätzlich, da wir zum Beispiel komplexe Zusammenstellungen von Daten nicht immer wieder neu erfassen müssen für eine Datenabfrage, sondern einfach auf den bereits erstellten Reversal zurückgreifen können. Ein einfaches Beispiel sind die Juristen, deren Reversal wir mit ihrer geografischen Herkunft kombiniert auswerten können. Umgehend können wir etwa eine Karte erstellen und dadurch erkennen, aus welchen Regionen Europas die Juristen stammten und wo sie studierten. Natürlich müssen wir die Definitionen der Projekte zu den Juristen berücksichtigen und dies unter Umständen mit einem Reversal differenziert markieren. Ein Jurist im RAG wird etwa auch als ein solcher bezeichnet, wenn er seine Studien ohne Abschluss beendete. Weiter können wir etwa nur Juristen, die römisches Recht studiert haben, nach Herkunfts- und Studienorten auf der Karte visualisieren und damit zugleich die Bedeutung der Universitäten für dieses Fach sichtbar machen, dies vor allem in quantitativer und räumlicher Hinsicht. Grundsätzlich können wir jedoch mit der vorgestellten Methode Bildungs-, Wissens- oder Kommunikationsräume nicht nur allgemein visualisieren, sondern gezielt rekonstruieren, indem wir eine Untersuchung ausgehend von bestimmten Datenbankobjekten (Personen, Orte, Institutionen, Werke etc.) kombiniert durchführen. Falls wir weitere Daten zur Verfügung haben, zum Beispiel Informationen über die Tätigkeiten und Ämter der Juristen, so wie im RAG, können wir auf Karten und in Netzwerken nachverfolgen, wohin ihr juristisches Wissen mit ihnen gelangte. Dies, indem wir die Studenten und Gelehrten als Wissensträger betrachten. So können wir etwa die Verbreitung eines Fachgebiets im vormodernen Europa, zum Beispiel des römischen Rechts, anhand der personellen Mobilität der Personen in großen Linien nachzeichnen und diese mit qualitativen Untersuchungen zu juristischen Werken verfeinern. Hierfür können wir Bibliotheksbestände der Gelehrten einbeziehen oder auch die Einflüsse gelehrter oder fachkundiger Wissenszirkel oder von Personen (zum Beispiel Privatgelehrte, Verwandte) untersuchen, wofür wir etwa die Netzwerkanalyse einsetzen können. Diese allerdings nicht nur für Personen, sondern etwa auch für Werke, Lehr- oder Lerninhalte. Damit können wir auch erzieherische Traditionen und ihre Wirkungen erforschen oder allgemein Veränderungen im Bildungswesen. Mit den vorgestellten Techniken können wir somit beliebige Objekttypen zur Grundlage einer Analyse nehmen, die wir in Makro- oder Mikroperspektive sowie auch in komparatistischer Perspektive durchführen können. Kombiniert mit geographischen Raumanalysen ist es sodann möglich, ganze Wissens- und Bildungslandschaften entstehen zu lassen.

doi.org/10.35468/5952-09

Für eine räumliche Analyse zeichnen wir dazu mit GeoJSON eine Fläche auf Kartenbasis und kopieren den generierten JSON-Code in das Datenfeld der Benutzeroberfläche, in dem wir die Geodatensätze hinterlegt haben. Anschließend können wir innerhalb dieser Fläche die Daten nach den uns interessierenden Kriterien abfragen und auf den Karten differenziert visualisieren.[17]
Zusammenfassend ist festzuhalten, dass die Module (DDI, Reversal, Reconciliation) auf verschiedene Weise, je nach Strukturen und Inhalten der Quelldatenbanken, für eine Vernetzung von Forschungsdaten verwendet werden können.[18] Zu ergänzen ist, dass mit dem DDI-Modul Daten auch nur anreichern können. Zum Beispiel können wir mit dem Modul zu einer Person, basierend auf dem GND-Identifikator, weitere Daten von Wikidata oder von LOBID hinzufügen.[19]

Abb. 5: Angereicherte Daten für einen Personendatensatz (Josias Simmler) mit Links zu Portalen und mit Publikationen in Listenform, die hier ebenfalls hinzugefügt wurden.

17 Dies mit der Funktion der ‚Conditions', vgl. dazu die Dokumentation von nodegoat, https://nodegoat.net/documentation.s/88/conditions (Zugriff: 18.02.2022).

18 Reklassifizieren etwa mit dem CIDOC-Referenzmodell (https://www.cidoc-crm.org, Zugriff: 18.02.2022) mit weiteren Hinweisen zu möglichen Anwendungen.

19 Einen sehr wertvollen Beitrag für Datenvernetzungen leistet das Projekt LOBID, das vom Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen betrieben wird. LOBID gehört zu den wenigen Projekten, die eine Datenschnittstelle mit reichhaltigen und gut strukturierten Daten auf aktuellem technischen Stand zur Verfügung stellt, siehe https://lobid.org (Zugriff: 18.02.2022).

doi.org/10.35468/5952-09

Dies können etwa auch Korrespondenzen sein, die ein Archiv bereitstellt und die wir, basierend auf dem GND-Identifikator für die Personen importieren können, wie im dargestellten Beispiel. Entsprechend können wir auch Texte importieren und dann die Analyseinstrumenten darauf anwenden können.[20] Letztlich können wir sämtliche Daten, die wir zusammengestellt, vernetzt oder angereichert haben, via Schnittstelle im JSON-Format für die Forschung wieder bereitstellen.[21]
Anhand des Fallbeispiels universitätsgeschichtlicher Datenbankprojekte wurden Zugänge für eine Vernetzung von Forschungsdaten in den Geisteswissenschaften aufgezeigt. Betont wurde die Notwendigkeit, bei der Erhebung und Vernetzung von Forschungsdaten mehr mit qualitativen Methoden sowie Kenntnissen zur Quellen- und Forschungslage vorzugehen, um den Besonderheiten der Quellenüberlieferung gerecht zu werden, welche oft heterogen und lückenhaft ist und entsprechend eine ‚weiche' Datengrundlage bildet. Erst wenn dies berücksichtigt wird, können Daten mit Erkenntnisgewinn vernetzt werden. Hierzu wurde im Fallbeispiel eine virtuelle Forschungsumgebung eingesetzt, die mit ausgeklügelten Techniken nicht nur einen quantitativen, sondern besonders einen gezielten qualitativen Datenzugriff erlaubt und Datenanalysen aus verschiedenen Perspektiven ermöglicht.

Literatur

Beretta, F. & Riechert, T. (2016): Collaborative Research on Academic History using Linked Open Data: A Proposal for the Heloise Common Research Model. CIAN – Revista de Historia de las Universidades, Instituto Figuerola de Historia y Ciencias Sociales – Universidad Carlos III de Madrid 19 (1), 133–151. DOI: https://doi.org/10.20318/cian.2016.3147

Blasch, E.; Ravela, S. & Aved, A. (2018) (eds.): Handbook of Dynamic Data Driven Applications Systems Cham: Springer.

Burrows, T. (2017): The History and Provenance of Manuscripts in the Collection of Sir Thomas Phillipps: New Approaches to Digital Representation. In: Speculum. A Journal of Medieval Studies 92 (1), 39–64. DOI: https://doi.org/10.1086/693438

Carius, H. (2021): Europäische Gelehrtennetzwerke digital rekonstruieren: Vernetzung von Briefmetadaten mit Early Modern Letters Online (EMLO), in: Bibliotheksdienst 55 (1), 29–41. DOI: https://doi.org/10.1515/bd-2021–0008

Dumont, S. (2015): correspSearch – Connecting Scholarly Editions of Letters, in: Journal of the Text Encoding Initiative (Selected Papers from the 2015 TEI Conference) 10. DOI: https://doi.org/10.4000/jtei.1742

20 Siehe die Anwendung des Moduls mit einer Software zur Handschrifterkennung (Transkribus): https://nodegoat.net/blog.s/58/connect-your-nodegoat-environment-to-wikidata-bnf-transkribus-zotero-and-others (Zugriff: 18.02.2022).

21 Dies kann beispielsweise auf nützlich sein, wenn Daten in der Zieldatenbank bereinigt und vereinheitlicht wurden und in verbessertem Zustand wieder in die Quelldatenbank re-importiert werden. Solche Bereinigungen können auch bei einer vollumfänglichen Datenbankmigration stattfinden, für die das Datenmapping des DDI-Modul freilich auch eingesetzt werden kann. Vgl. auch einen weiteren Anwendungsfall für eine Datenbereinigung via nodegaot bei Carius 2021.

doi.org/10.35468/5952-09

Gubler, K. (2021): Data Ingestion Episode III – May the linked open data be with you. In: HistData, 30.03.2021. URL: https://histdata.hypotheses.org/2130 (Zugriff: 18.02.2022)

Gubler, K.; van Bree, P. & Kessels, G. (2022a): Server-side data harmonization through dynamic data ingestion. A centralized approach to link data in historical research. In: Brizzi, G. P.; Frova, C. & Treggiari, F. (Hrsg.): Fonti per la storia delle popolazioni accademiche in Europa/Sources for the History of European Academic Communities. Dixième Atelier Héloïse/Tenth Workshop Heloise, Bologna: Il Mulino.

Gubler, K.; Hesse, C. & Schwinges, R. C. (2022b) (Hrsg.): Person und Wissen. Bilanz und Perspektiven [RAG Forschungen; 4], Zürich: vdf Hochschulverlag.

Gubler, K. (2022c): Von Daten zu Informationen und Wissen. Zum Stand der Datenbank des Repertorium Academicum Germanicum, in: Gubler, K.; Hesse, C. & Schwinges, R. C. (Hrsg.): Person und Wissen. Bilanz und Perspektiven [RAG Forschungen; 4], Zürich: vdf Hochschulverlag, 19–47.

Hafner, U. (2019): Die Zeitmaschine ist kaputt. In: NZZ am Sonntag, 15.12.2019. URL: https://magazin.nzz.ch/wissen/venice-time-machine-knatsch-um-millionen-projekt-ld.1528382 (Zugriff: 18.2.2022)

John, T. & Pankaj, M. (2017) (eds.): Data Lake for Enterprises, Birmingham: Packt.

Keim, D. A. (2004): Datenvisualisierung und Data Mining. In: Kuhlen, R.; Seeger, T. & Strauch, D. (Hrsg.): Handbuch zur Einführung in die Informationswissenschaft und -praxis, Bd. 1: Grundlagen der praktischen Information und Dokumentation. München: De Gruyter, 362–370.

Koho, M.; Burrows, T. … & Wijsman, H. (2021): Harmonizing and publishing heterogeneous premodern manuscript metadata as Linked Open Data, in: Journal of the Association for Information Science and Technology 73 (2). DOI: https://doi.org/10.1002/asi.24499

Kuhlen, R.: Information. In: Kuhlen, R.; Seeger, T. & Strauch, D. (Hrsg.): Handbuch zur Einführung in die Informationswissenschaft und -praxis, Bd. 1: Grundlagen der praktischen Information und Dokumentation. München: De Gruyter, 3–20.

Leimeister, J. M. (2021): Einführung in die Wirtschaftsinformatik. 13. Aufl. Berlin/Heidelberg: Springer Gabler. DOI: https://doi.org/10.1007/978–3–662–63560–5

Randeraad, N. (2018): Dutch Social Reformers in Transnational Space, 1840–1914: Reflections on the CLARIAH Research Pilot 2TBI. URL: https://cris.maastrichtuniversity.nl/files/31584584/Dutch_Social_Reformers_2TBI_report_article.pdf (Zugriff: 18.02.2022)

Schwinges, R. C. (2020): Warum gab es fast nur im deutschen Reich allgemeine Universitätsmatrikeln? Eine Frage der Reichweite, in: Henkel, N.; Noll, T. & Rexroth, F. (Hrsg.): Reichweiten. Dynamiken und Grenzen kultureller Transferprozesse in Europa, 1400–1520, Bd. 1: Internationale Stile – Voraussetzungen, soziale Verankerungen, Fallstudien [Abhandlungen der Akademie der Wissenschaften zu Göttingen, N. F.; 49.1). Berlin/Boston: De Gruyter, 37–58.

van Bree, P. & Kessels, G. (2013): nodegoat: a web-based data management, network analysis & visualisation environment. URL: http://nodegoat.net (Zugriff: 18.02.2022)

van Bree P. & Kessels, G. (2015): Mapping Memory Landscapes in nodegoat. In: Aiello, L. M. & McFarland, D. (eds): Social Informatics [Lecture Notes in Computer Science; 8852]. Cham: Springer. DOI: https://doi.org/10.1007/978–3–319–15168–7_34

van den Heuvel, C. M. J. M & Alvarez Frances, L. (2014): Mapping Notes and Nodes in Networks: Exploring potential relationships in biographical data and cultural networks in the creative industry in Amsterdam and Rome in the Early Modern Period. Eindrapport KNAW PPS project.

doi.org/10.35468/5952-09