In dem auf die Forschungsdaten sprach- und textbasierter Disziplinen ausgerichteten NFDI-Konsortium Text+ spielen Normdaten eine zentrale Rolle für die interoperable Beschreibung und semantische Verknüpfung von verteilten Datenquellen. Insbesondere die Gemeinsame Normdatei (GND) ist ein bedeutender Hub im Zentrum eines im Entstehen begriffenen, domänenübergreifenden Wissensgraphen. Diese Funktion soll im Rahmen von Text+ durch den Aufbau einer GND-Agentur für sprach- und textbasierte Forschungsdaten weiterentwickelt und ausgebaut werden. Ziel ist es, niedrigschwellige, qualitätsgesicherte Beteiligungsmöglichkeiten für Forschende zu schaffen und zugleich den Vernetzungsgrad der GND auch durch Terminologie-Mappings zu erweitern. Spezifische Anforderungen und Nutzungspraktiken werden hierbei anhand der Datendomänen von Text+ exemplifiziert.

Authority data play a central role in the interoperable description and semantic linking of distributed data sources in the NFDI Text+ consortium, which is dedicated to the research data of language and text-based disciplines. The Integrated Authority File (GND) in particular is a key hub at the centre of a nascent cross-domain knowledge graph. This function is to be developed and expanded within Text+ through the establishment of a GND agency for language and text-based research data. The aim is to create low-threshold, quality-assured ways in which researchers can participate. A further aim is to expand the level of networking within the GND, including through terminology mapping. The specific requirements and use practices are exemplified by the data domains of Text+.

JÜRGEN KETT, CHRISTOPH KUDELLA, ANDREA RAPP, REGINE STEIN, THORSTEN TRIPPEL

Text+ und die GND – Community-Hub und Wissensgraph

Einführung

In einer Nationalen Forschungsdateninfrastruktur (NFDI), die zum Ziel hat, das gesamte Wissenschaftssystem zu adressieren, vertreten die komplementär aufgestellten geistes- und kulturwissenschaftlichen Konsortien die Bedarfe einer langen wissenschaftlichen Tradition mit großer Fächervielfalt, hohen Studierenden- und Forschendenzahlen und einem umfänglichen bis in die 1940er-Jahre zurückreichenden digitalen Methodenkoffer. Eine NFDI muss die Entwicklung, die diese Disziplinen getrieben durch den digitalen Wandel vollziehen, durch Infrastrukturkomponenten und Dienste unterstützen. Zugleich ist das kulturelle Erbe, mit dem sich die Geistes- und Kulturwissenschaften insbesondere befassen, auch jenseits des Wissenschaftssystems von allgemeiner Bedeutung. Gegenstand der Geistes- und Kulturwissenschaften sind alle kulturellen Erzeugnisse des Menschen in der gesamten Breite von Sprache, Kunst, Geschichte, Religion und Gesellschaft. Sprache, Sprachen und Texte spielen dabei in vielen Bereichen eine zentrale Rolle, auch über die Geistes- und Kulturwissenschaften hinaus. Ziel des seit Oktober 2021 geförderten NFDI-Konsortiums Text+[1] ist der Aufbau einer Forschungsdateninfrastruktur, deren Konzepte, Tools und Services für alle auf Sprach- und Textdaten basierenden Fragestellungen und Anwendungen interessant sind. Text+ ist daher interdisziplinär und fachübergreifend an zentralen Datendomänen ausgerichtet.

Für die geistes- und kulturwissenschaftlichen Disziplinen, an die sich das Konsortium prioritär richtet, legt Text+ den Fokus zunächst auf Text- und Sprachdaten-Sammlungen, lexikalische Ressourcen und Editionen. Diese drei Datendomänen haben eine lange Forschungstradition und sind mit ausgereiften methodologischen Paradigmen verknüpft, die charakteristische, aber auch bereichsübergreifende Praktiken der Erzeugung, Kuratierung und des Managements von Daten erfordern. Sie sind unabdingbar für eine breite Palette von Fachdisziplinen, u. a. für die Klassische Philologie, die Sprach- und Literaturwissenschaft, Sozial- und Kulturanthropologie, Außereuropäische Kulturen, Judaistik und Religionswissenschaften, Philosophie sowie für die sprach- und textbasierte Forschung in den Sozial- und Politikwissenschaften.

Diese fachliche Breite und Ausdifferenzierung spiegelt sich in den Institutionen, Verbänden und Persönlichkeiten, die sich für Text+ engagieren und dessen Ziele unterstützen. Fundiert wird Text+ über einen Kooperationsvertrag von 34 Institutionen aus der gesamten Palette geisteswissenschaftlicher Einrichtungen: Hochschulen, wissenschaftliche Bibliotheken, Datenzentren der Digital Humanities, Mitglieder der Deut-

schen Akademienunion und der Leibniz-Gemeinschaft sowie führende Rechenzentren, die einen robusten und persistenten Betrieb der Dienste für eine distribuierte Forschungsdateninfrastruktur absichern. Hinzu kommen Fachverbände und weitere Partner wie Fachinformationsdienste, die bereits im Rahmen der Antragstellung ihre Unterstützung zugesagt haben[2] und im Text+ Plenum sowie seinen wissenschaftlichen Arbeits- und Beratungsgremien aktiv mitwirken. Zusätzlich zur Einbindung all dieser Partner von Beginn der NFDI-Entwicklungen und der Antragskonzeption[3] an werden offene Calls for User Stories sowie Calls for Data an die Communitys gerichtet. Der hohe Bedarf und damit auch das hohe Interesse an Text+ wird durch über 120 forschungsgeleitete User Stories dokumentiert, die publiziert und für die Ausgestaltung des Arbeitsprogramms ausgewertet wurden (Rißler-Pipka et al. 2021).[4]

Auch organisatorisch ist das Community-Engagement integrativer Bestandteil der Lenkungsstruktur von Text+: Im Zentrum stehen drei wissenschaftliche Koordinationskomitees für die Datendomänen und eines für Infrastruktur und Betrieb, die im Rahmen eines jährlich stattfindenden Plenums von Text+ Partnern und der Community gewählt werden. Initial sind die Koordinationskomitees mit Repräsentant*innen der Fachverbände besetzt, um eine breite und ausbalancierte Beteiligung der Disziplinen zu gewährleisten. Ihre Aufgabe besteht darin, das Portfolio an Daten, Werkzeugen und Diensten kontinuierlich zu evaluieren und dessen Weiterentwicklung nach den Prioritäten der beteiligten Fachdisziplinen in Abstimmung mit den Infrastrukturanbietern voranzutreiben. Diese Gremien haben somit eine hohe Steuerungsbefugnis und gewährleisten, dass die Bedarfe der Fachcommunitys den Aufbau und Betrieb von Text+ steuern – ganz im Sinne der erklärten Zielsetzung der NFDI. Community-Aktivitäten sind in Text+ zentral als Querschnittsthema in allen Arbeitsbereichen verankert, sie umfassen ein differenziertes und – auch im Sinne der Datendomänen und Fachcommunitys – passgenaues Angebot von Workshops, Trainings, Beratung sowie curricularen Aktivitäten, das von spezifischen Kompetenznetzen getragen und ausgestaltet wird.

Die Forschungsdatenmanagementstrategie stellt das entscheidende Instrument dar, um die übergeordneten Ziele von Text+ im NFDI-Kontext umzusetzen. Sie ebnet den Weg für die Integration von Daten, Werkzeugen und Diensten in eine Infrastruktur, die übergreifenden sowie fachspezifischen Standards genügt und die FAIR- und CARE-Prinzipien umsetzt. Ein wesentlicher Leitgedanke ist dabei ihre Erweiterbarkeit für weitere Datendomänen, die zugleich die Integration von Text+ in die gesamte NFDI optimal gewährleistet und die Bereitstellung von Angeboten, die auch für andere Konsortien relevant sind, unterstützt.

Im Mittelpunkt der Strategie stehen thematische Cluster, die in einer Datendomäne Aktivitäten zu bestimmten Unterarten von Daten und Forschungsmethoden bündeln. Jedes Cluster besteht aus spezialisierten Daten- und Kompetenzzentren. Die Cluster bieten Dienste an, die auf den vom Arbeitsbereich *Infrastructure/Operations* bereitgestellten allgemeinen Diensten aufbauen und auf die clusterspezifischen Anforderungen zugeschnitten werden. Das Arbeitsprogramm ist dabei in die drei Bereiche Datendienste, Community-Aktivitäten und Software-Dienste gegliedert, die übergreifend koordiniert werden und gemeinsam zu den Querschnittsthemen der NFDI beitragen.

Ein Kernelement in Text+ ist der systematische Ausbau von Verlinkungen zwischen den Text+ Datendomänen. Sammlungen und Editionen dienen beispielsweise

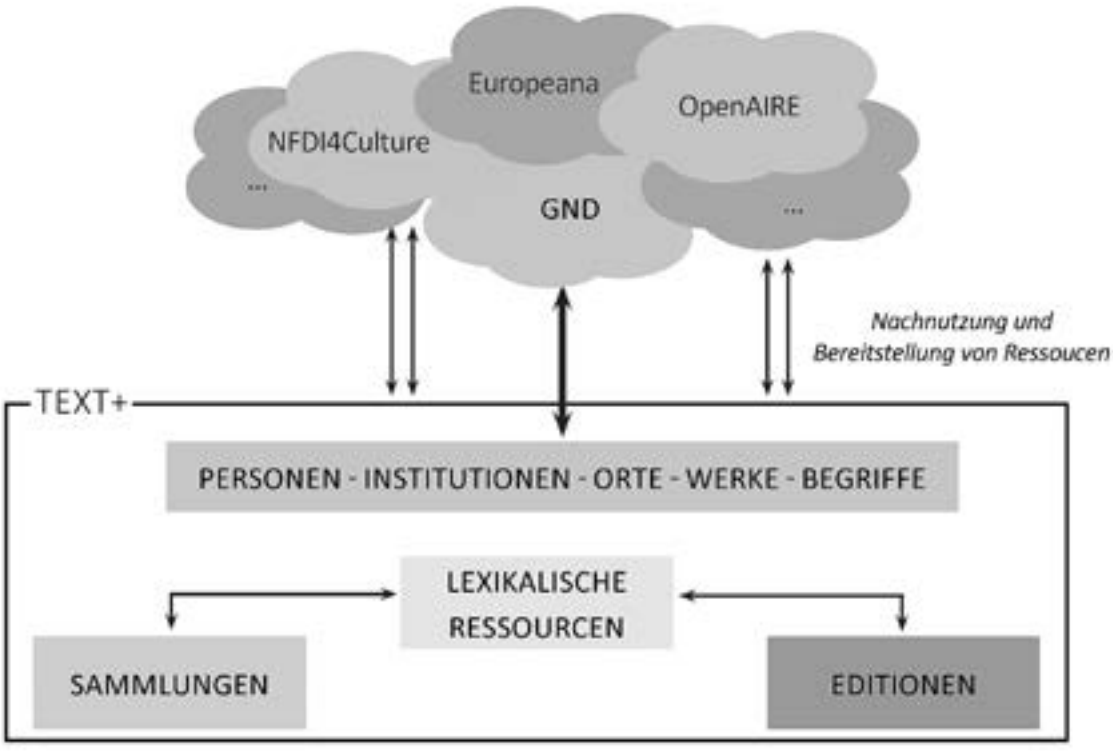
1 Text+ Forschungsdatenmanagementstrategie

als Belegstellen für Einträge in »Lexikalische Ressourcen«. Lexikalische Ressourcen dienen der strukturierten Erschließung von Sammlungen und dem Verweisen auf normierte Entitäten in Editionen. Durch das Einbringen dieser Verlinkungen, d. h. die Verknüpfung mit persistenten Identifikatoren, werden Suchzugänge zu den Forschungsdaten erweitert, ihre Integration in Wissensbasen unterstützt und Möglichkeiten für automatisierte Verfahren wie dem Text und Data Mining ausgebaut. Eine besondere Rolle spielen dabei Normdaten und hier insbesondere die Gemeinsame Normdatei (GND).[5] Die GND als disziplinübergreifend genutzte und ihrerseits mit den Normdateien anderer Nationalbibliotheken verlinkte Ressource ist ein wichtiger Hub in der Linked Open Data Cloud, der im Rahmen von Text+ für sprach- und textbasierte Forschungsdaten weiterentwickelt und ausgebaut werden soll. Die Text+ Akteure verfolgen damit zugleich für die NFDI insgesamt und darüber hinaus relevante Infrastrukturentwicklungen für Kultur und Wissenschaft.

Die GND als Hub

In Zeiten der Digitalisierung und Datenvernetzung benötigen Forschungs- und Kultureinrichtungen Normdateien als gemeinsam genutztes und gepflegtes maschinenlesbares Vokabular, um Bestände, Sammlungen, Forschungsprozesse und -ergebnisse interoperabel zu beschreiben, um zwischen diesen verteilten Datenquellen auf verlässliche Weise semantische Verknüpfungen herzustellen und um sie so in einen größeren Zusammenhang einzubetten. Normdateien können Individualbegriffe (*Named Entities*, z. B. Personen- oder Ortseinträge) und Allgemeinbegriffe (z. B. Schlagworte) enthalten. Sie schaffen zusätzliche Sucheinstiege und thematische Zugänge für die Recherchierenden, etwa durch das Anreichern von Suchindices mit Namensvarianten oder Synonymen und Entsprechungen in anderen Sprachen. Darüber hinaus sind die Angabe von verwandten Begriffen, alternativen Identifikatoren und weiteren identifizierenden Informationen für die Kontextualisierung für die individuelle Forschungsarbeit, aber beispielsweise auch für die Nutzung im Text und Data Mining von großem Vorteil. Zentral ist dabei die Bereitstellung eines dauerhaften, weltweit eindeutigen und maschinenlesbaren Identifikators, also eines persistenten URI, für jeden Eintrag.

Unter dem Leitmotiv »Brücken für die Kultur und Wissenschaft« bietet die GND einer wachsenden Community von Kultur- und Wissenschaftseinrichtungen des D-A-CH-Raums einen zentralen Hub für Normdaten. Diese Brücken – bestehend aus persistent adressierbaren Normdaten zu Personen, Körperschaften, Geografika, Ereignissen, geistigen Schöpfungen und Sachbegriffen – können im Zusammenspiel mit der NFDI ein Rückgrat eines domänenübergreifenden Wissensgraphen bilden und z. B. Thesauri und Fachdatenbanken der verschie-



2 Text+ Forschungsdateninfrastruktur: die GND als Hub

denen wissenschaftlichen Disziplinen über Konkordanzen miteinander verbinden. Hierbei fungiert die GND nicht nur direkt als Datenhub, sondern auch als Vermittlerin zwischen den mit der GND verknüpften Datensets des Wissensgraphen. Auf diese Weise können zu jeder betrachteten GND-Entität indirekt je nach Bedarf z. B. auch Daten aus der Linked Data Cloud wie Wikidata, Wikipedia, VIAF, internationalen Thesauri und Fachdatenbanken hinzugezogen werden.

Integration neuer Communitys ist Teil des Selbstverständnisses

Damit die GND ihre Funktion beim Aufbau einer übergreifenden Dateninfrastruktur im Rahmen der NFDI erfüllen kann, muss sie sich den Bedarfen der diversen neu hinzukommenden wissenschaftlichen Communitys annähern und diese als aktive und gleichberechtigte Anwendende integrieren. Mit der Ausweitung der GND-Nutzung steigen die Anforderungen an die Organisation, Kommunikation, Regelwerksarbeit, Werkzeuge und Infrastruktur. Die GND ist bereits jetzt ein Gemeinschaftswerk von mehr als 1.000 Einrichtungen – Tendenz steigend. Dahinter stehen motivierte Redakteur*innen, die durch ihre Expertise und ihr Qualitätsbewusstsein für ihre hohe Güte, Verlässlichkeit und Verbreitung sorgen. Bei Veränderungen müssen daher viele Einrichtungen und Menschen überzeugt und gewonnen werden.

Es gilt insbesondere, eine Balance zwischen den Interessen der neuen Communitys und den langjährigen Partnern zu finden und dabei die wichtigsten Stärken der GND zu erhalten: Qualität, Verlässlichkeit und Universalität. Die Integration neuer Nutzendengruppen ist in der GND ein fortlaufender Prozess, der seit der Verabschiedung des GND-Entwicklungsprogramms[6] konsequent verfolgt und im Rahmen von Projekten wie GND4C[7] vorangetrieben wird. Der aktuelle Stand der Entwicklungen lässt sich nach Handlungsfeldern gegliedert wie folgt zusammenfassen:

(1) Governance: Die Governance- und Gremienstruktur der GND wird schrittweise um Vertretungen aus den hinzukommenden Communitys erweitert. Ein wichtiges Element war dabei die Einführung von Foren und Arbeitsgruppen, um die Bedarfe der jeweiligen Community zu bündeln.

(2) Kooperative Redaktion: Die Datenpflege verläuft durch ein kooperatives Zusammenspiel von GND-Redaktionen. Diese werden durch einen Verbund von GND-Agenturen koordiniert sowie mit Dienstleistungen und Infrastruktur unterstützt. Ein zentrales Element des Entwicklungsprogramms ist der Aufbau neuer Agenturen und Redaktionen, die für die neuen Communitys infrastrukturelle, redaktionelle und beratende Dienste anbieten.

(3) Regeln und Format: Die Möglichkeit zum Ergänzen community-spezifischer Daten wird 2022 in Zusammenarbeit mit den jeweiligen Arbeitsgruppen und Gremien in die redaktionelle Praxis überführt. Kern eines jeden Normdatensatzes bildet weiterhin ein Set an Elementen, die nach gemeinsamen Regeln erfasst werden (*CORE*-Bereich). Um besonderen Anforderungen einer Anwendungsgemeinschaft der GND gerecht zu werden, kann der *CORE*-Bereich künftig um spezifische Regeln und Elemente mit gesonderter redaktioneller Verantwortlichkeit (sogenannte *PLUS*-Bereiche) ergänzt werden.

(4) Datenbasis: Die Erweiterung der Datenbasis erfolgt häufig im Rahmen von Kooperationen zwischen Forschungsvorhaben und GND-Redaktionen. Anhand von Praxiserfahrungen wurden bereits ein Workflow und ein Kriterienkatalog für Einspielungen und Anreicherungen entwickelt.

(5) Datenvernetzung und Analyse: Von zentraler Bedeutung für die Funktion als Hub sind die Verknüpfungen innerhalb der GND und mit anderen Systemen wie ORCID und VIAF[8], die durch maschinelle Verfahren und einen Claiming-Dienst ausgebaut werden konnten. Auch an einer verbesserten Unterstützung zur kooperativen Pflege von Cross-Konkordanzen der GND-Sachbegriffe zu anderen Thesauri (wie z. B. RAMEAU[9]) wird aktuell gearbeitet.

(6) Anwendungen und Schnittstellen: Mit dem GND-Explorer wird zurzeit eine Anwendung zum Stöbern in den Daten und zur Visualisierung von Hierarchien und semantischen Zusammenhängen entwickelt. Für eine aktive Mit- und Zusammenarbeit über die GND-Redaktionen hinaus werden die Eingabemöglichkeiten über einfache Webformulare erweitert und eine zusätzliche Redaktionsumgebung aufgebaut.

Normdaten als Querschnittsthema der NFDI

Um die Vision eines domänenübergreifenden Wissensgraphen Realität werden zu lassen, müssen auch im Umfeld der GND die Rahmenbedingungen weiterentwickelt werden. Die NFDI bietet für solche Entwicklungen einen idealen Kontext. Die besondere Bedeutung des Themas für die Geistes- und Kulturwissenschaften spiegelt sich nicht zuletzt darin, dass die NFDI-Initiativen NFDI4Culture, NFDI4Memory, NFDI4Objects und Text+ bereits in ihrem *Memorandum of Understanding* (Brünger-Weilandt et al. 2020) den Komplex »Metadaten, Normdaten, Terminologien« (S. 2, Punkt 7) als gemeinsam zu bearbeitendes Querschnittsthema identifiziert haben. Eine mögliche Umsetzung besteht in einem Netzwerk verteilter Datenbanken, die über standardisierte Schnittstellen miteinander kommunizieren. In einem solchen Ökosystem müssen ganz unterschiedliche Anforderungen in Einklang gebracht werden. Nicht alles wird sich in einer übergreifenden Lösung umsetzen lassen. Es gilt, Anforderungen so voneinander abzugrenzen, dass verschiedene Softwarelösungen und -schichten verteilt umgesetzt werden, aber opti-

mal miteinander arbeiten. Hier sind eine enge Abstimmung zwischen den NFDI-Konsortien und gemeinsame Aktivitäten erforderlich. Mit der Einrichtung der Sektion »(Meta-)data, terminologies, provenance« (Koepler et al. 2020) als einer der ersten vier Sektionen im NFDI-Verein ist das Thema auch übergreifend in der NFDI gesetzt.

Die folgenden beiden Abschnitte greifen einige spezifische Anforderungen der im Fokus von Text+ stehenden sprach- und textbasierten Disziplinen auf.

Anreicherung von Metadaten und Forschungsdaten mit Normdaten

Daten, die im Rahmen von Forschungstätigkeiten gesammelt und genutzt werden, sind sehr divers und orientieren sich an den Erfordernissen und Möglichkeiten im jeweiligen Forschungskontext. Aufgrund dieser Diversität erfordert es einen erheblichen fachlichen Aufwand, um sie angemessen – mittels Metadaten – zu beschreiben. Forschungsdaten werden, im Gegensatz zur bibliothekarischen und archivarischen Praxis, in der Regel noch nicht durch ausgebildete Fachkräfte mit Metadaten versehen. Stattdessen erfolgt dies im Zusammenspiel zwischen den Forschenden, die die Daten bereitstellen, und jenen, die später die Bereitstellung der Daten unabhängig von den Forschenden gewährleisten. Ein prototypisches Beispiel ist der Sonderforschungsbereich 833 »Bedeutungskonstitution: Dynamik und Adaptivität sprachlicher Strukturen«, in dem unter anderem Textkorpora, Word-Embeddings für das maschinelle Lernen, Videoaufnahmen, Reaktionszeitexperimente und EEG-Untersuchungen verwendet werden.[10] Text+ muss in allen drei Datendomänen Sammlungen, Lexikalische Ressourcen und Editionen, in denen jeweils eine Vielzahl unterschiedlicher Datentypen auftreten können, damit umgehen, dass Metadaten rein textuell, also ohne Verknüpfungen in andere Wissensbasen, erfasst sind oder sogar gänzlich fehlen. Eine Verwendung der Daten etwa im Umfeld von Linked Data (Ďurčo und Windhouwer 2014) ist aber vielversprechender, wenn die Inventare auf Normdaten und andere Verzeichnisse verweisen.

Named-Entity-Recognition in Metadaten und Forschungsdaten

Innerhalb der Metadaten für Forschungsdaten erscheinen Named Entities, also Eigennamen, an unterschiedlichen Positionen. Auf der einen Seite können sie innerhalb von Beschreibungstexten enthalten sein, auf der anderen Seite erlauben Metadatenbeschreibungen auch die Definition von bestimmten Elementen, in denen – zumindest nach einer Qualitätssicherung – nur Eigennamen vorkommen können. In Metadaten sollte beispielsweise ein Feld für einen Autorennamen nur einen Namen enthalten, selbst wenn dieser nicht eindeutig ist.

In qualitätsgesicherten Metadaten ist der Aufwand zur Extraktion von Named Entities (Named-Entity-Regocnition, NER) relativ gering, da bereits bekannt ist, an welcher Stelle oder in welchen Strukturen sich Eigennamen befinden. Mit der Erkennung der Named Entities können die Metadaten mit deren Identifikatoren sowie weiteren Informationen angereichert und mit anderen Quellen verknüpft werden, wozu die GND sowie andere Normdatensätze verwendet werden können. Mittels Programmierschnittstellen (APIs) geschieht dies automatisch. Dazu extrahiert ein Programm aus den Metadaten die Named Entities und startet über die jeweilige API eine Abfrage in den Normdatenverzeichnissen. Die Referenz zu den Normdaten kann dann als zusätzliche Annotation ergänzt werden. Bei der automatisierten Annotation von Named Entities ist aber nicht sichergestellt, dass die richtige Referenz gefunden wird.[11] In der Praxis werden bei der Anreicherung von Named Entities über Normdaten alle möglichen Ergebnisse aus der Normdatenabfrage als Kandidaten einer Verlinkung angeboten. Im Rahmen von Qualitätssicherungs- und Disambiguierungsverfahren kann durch menschliche Expertise die Verknüpfung verifiziert werden.[12]

Named Entities in Textdaten, also sowohl in Metadaten als auch in textuellen Forschungsdaten, können mittels NER automatisch erkannt werden. Die Daten enthalten in aller Regel Referenzen auf (reale oder auch fiktive) Personen, auf Organisationen und auf Orte. Gängige NER unterscheiden z. B. Personen von Institutionen und Orten. Im einfachsten Fall werden dabei Namen einer Liste in einem Text gesucht. Wird die Zeichenkette gefunden, kann sie ganz analog zu den explizit für Named Entities definierten Metadaten-Elementen in Normdaten nachgeschlagen und mit ihnen verknüpft werden. Die Namenslisten können dabei sehr umfangreich sein, beispielsweise wenn sie aus Normdateien wie der GND direkt erstellt werden. Im Rahmen eines Parsings kann darüber hinaus erkannt werden, dass ein Wort mit hoher Wahrscheinlichkeit ein Name ist, selbst wenn der Name nicht Teil einer Liste ist. Auf diese Weise können textuelle Daten genutzt werden, um eine Namensliste zu vervollständigen und Normdaten zu erweitern. Die Erkennungsleistungen eines NER sind allerdings abhängig von Sprachmodellen, die explizit oder implizit die Trainingsdaten beinhalten, z. B. Informationen zu Sprache, Gattung, zeitlicher Zuordnung. Wenn ein Modell beispielsweise auf der Grundlage von Zeitungstexten einer Sprache erstellt wurde, ist zu erwarten, dass es nicht genauso gute Ergebnisse für Romane erzielen kann und erst recht nicht für andere Sprachen. Bereits andere Sprachstufen oder regionale Varietäten können einen Einfluss auf die Erkennungsleistung haben: so können Named Entities in Editionstexten früherer Jahrhunderte nicht zuverlässig erkannt werden, wenn ein Sprachmodell für moderne Zeitungstexte verwendet wird.

Semantische Anreicherung anhand lexikalischer Ressourcen

Named Entities sind nur ein Beispiel für die Möglichkeit einer semantischen Anreicherung von Texten. Auch Ontologien, Wortnetze, Terminologien und andere lexikalische Ressourcen können genutzt werden, um Wörter in Texten um zusätzliche semantische Informationen anzureichern. Analog erfordert auch die semantische Annotation mittels Verknüpfung zu lexikalischen Ressourcen eine eindeutige Referenzierbarkeit. Für Wortnetze wie dem GermaNet (Hamp und Feldweg 1997; Henrich und Hinrichs 2010) gibt es mit dem Interlingual Index (ILI) einen sprachunabhängigen eindeutigen Index, der für diesen Zweck verwendet werden kann. Für die Wörter, die in den Wortnetzen repräsentiert sind, können damit auch andere lexikalische Ressourcen angereichert und verbunden werden. Gleichzeitig werden auch lexikalische Lücken identifiziert, also Wörter, die nicht verknüpft werden können, wodurch sich ein Reservoir für Ergänzungen in Wörterbüchern ergibt. Auch bei der Verknüpfung zu weiteren lexikalischen Ressourcen gibt es Ambiguitäten, Synonyme etc., die dazu führen, dass nicht jedes Wort mit einer entsprechenden Referenz annotiert werden kann, wodurch eine vollständige, semantisch interoperable Lösung nicht immer möglich ist. Mit Mitteln des Deep Learning und korpuslinguistischen Verfahren können aber zum Teil Rückschlüsse auf unbekannte oder ambige Wörter getroffen werden, etwa durch Wordembeddings.

Anforderungen an eine Normdateninfrastruktur

Für die Erschließung von textbasierten Forschungsdaten mit Metadaten und die Anreicherung der Metadaten wie auch der Forschungsdaten mit Normdatenreferenzen sind die Verfahren der Named-Entity-Recognition und der Verknüpfung mit lexikalischen Ressourcen von großer Bedeutung. Normdateien wie die GND dienen hier einerseits als Quelle und können umgekehrt selbst mithilfe dieser Verfahren potenziell erweitert werden. Um qualitativ hochwertige Ergebnisse zu erzielen, ist die richtige Auswahl und Anpassung der konkreten Verfahren an Fachspezifika der zu verarbeitenden Texte und die passende Kombination aus automatischer Erkennung und intellektueller Prüfung entscheidend. Ein niedrigschwelliger Zugang und eine vereinfachte Anwendung solcher Verfahren für Forschungsprojekte sind daher eine Herausforderung, die in Text+ adressiert wird. Damit zu verbinden sind Kriterienkataloge und Verfahren, nach denen in der GND noch nicht vorhandene Normdateneinträge angelegt oder separat vorgehalten und zugänglich gemacht werden sollen. Ein spezifisches Desiderat ist die Verknüpfung lexikalischer Ressourcen mit den Sachbegriffen der GND, denn damit kann die fachspezifische semantische Anreicherung von Texten bei gleichzeitiger Vernetzung über die GND deutlich verbessert werden.

Nutzung von Normdaten in digitalen Editionen

Ein weiteres Anwendungsfeld, in dem der Einsatz von Normdaten erleichtert und ausgebaut werden soll, sind digitale Editionen. Seit Langem ist es innerhalb der digitalen Editorik gängige Praxis, Entitäten in den Editionsdaten über eindeutige und persistente Identifikatoren mit Datensätzen externer Wissensbasen, insbesondere Normdateien wie der GND, zu verknüpfen. Diese Vorgehensweise hat bislang erst vereinzelt Eingang in die Empfehlungen von Fachverbänden oder Richtlinien der Forschungsförderung gefunden. Innerhalb des sich etablierenden, spezialisierten Rezensionswesens für Digitale Editionen[13] ist jedoch bereits deutlich erkennbar, dass ein Verzicht auf Normdatenverknüpfungen innerhalb der Community of practice durchaus negativ auffällt.

Aktuelle Nutzungspraxis

Normdaten spielen im Bereich digitaler Editionen auf mehreren Ebenen eine Rolle: Erstens auf der Ebene der Erschließung, hier vor allem in den TEI-XML-Daten. Zweitens auf den Ebenen der Publikation und Bereitstellung, konkret also den Webportalen sowie APIs oder speziellen Austauschformaten für die primär maschinelle Verarbeitung.

Erschließung

Im Prozess der Erschließung werden Verknüpfungen mit Normdateien oder ihnen ähnlichen Wissensbasen sowohl auf der Ebene der Metadaten (d. h. im TEI-Header) als auch im annotierten Text der Edition und/oder etwaigen separaten Registerdokumenten (d. h. primär im TEI-Body) vorgenommen. Textsortenabhängig können solche Normdatenverknüpfungen auch in zusätzlichen Elementen im TEI-Header zum Einsatz kommen. Ein sehr prägnantes Beispiel hierfür sind Briefeditionen mit der strukturierten Erfassung der Metadaten von Briefen (Stadler 2016).[14]

Primäre Funktion dieser Verknüpfungen ist die eindeutige Identifikation von Körperschaften, Personen, Orten und anderen Named Entities – einschließlich hierdurch oft geleisteter Disambiguierung – durch die Referenz auf eine externe Ressource in Form eines persistenten URIs. Zugleich ist es eine Vorbedingung für die automatisierte Generierung von Indices, z. B. Personen-, Orts- und Werkregistern sowie für die Durchführung komplexerer Verfahren des Information Retrievals und der Datennachnutzung (Iglesia und Göbel 2014). Zusätzlich führt diese Anbindung zu einer Entlastung der Editor*innen bei der Erstellung von Registern und Stellenkommentaren, da die in externen Wissensbasen vorgehaltenen Informationen hierfür in unterschiedlicher Form nachgenutzt werden können (Dumont 2020). Im D-A-CH-Raum ist der Verweis auf GND-Datensätze die dominante Praxis, für einzelne Entitäten werden

zudem weitere, zum Teil domänenspezifische Wissensbasen verknüpft.[15]

TEI-XML-Daten digitaler Editionen sind, auch wenn sie konform zu den TEI P5 Guidelines sind, aufgrund ihrer spezifischen Anwendungsprofile nicht ohne weiteres projektübergreifend interoperabel.[16] Dies hat zur Folge, dass die projekt- bzw. editionsübergreifende Suche nach identischen Entitäten auf Basis der jeweiligen Projektdaten nur durch auf die jeweilige Auszeichnungsform speziell zugeschnittene Abfragen z. B. via XQuery zu bewerkstelligen ist. Im Hinblick auf Nachnutzungsbedarfe haben sich folglich in bestimmten Kontexten spezielle Austauschformate herausgebildet. Ein Beispiel hierfür ist das aus dem Bereich der Briefedition stammende *Correspondence Metadata Interchange Format* (CMIF) für die Erstellung interoperabler digitaler Briefverzeichnisse. CMIF-Dateien bestehen aus einem im Vergleich zu projektspezifischen Ausprägungen stark reduzierten und zugleich restriktiveren TEI-Header, der nur solche Metadaten enthält, die für den projekt- bzw. editionsübergreifenden Datenaustausch notwendig sind (Dumont et al. 2019). Der Webservice correspSearch[17] demonstriert das Potenzial solcher in einem einheitlichen Austauschformat vorliegenden Briefverzeichnisse beispielsweise in eindrücklicher Weise. Dieser Dienst aggregiert nach vorheriger einmaliger Registrierung unter einer freien Lizenz zur Verfügung gestellte CMIF-Dateien und macht ihre Daten in einem gemeinsamen Suchraum verfügbar.[18]

Publikation und Bereitstellung

Auch auf den Ebenen der Publikation und Bereitstellung werden die in den Daten vorgenommenen Verknüpfungen mit externen Wissensbasen in mehrfacher Hinsicht genutzt:

(1) Nachnutzung der von verknüpften Datensätzen angebotenen Informationen: Diese werden zunehmend dazu verwendet, um die Registeransichten automatisch (und in der Regel selektiv) anzureichern. Anreicherungen können on-the-fly über eine API innerhalb der Webportale geschehen oder bereits zuvor erfolgt sein, indem entsprechende Informationen aus der externen Wissensbasis (automatisch oder manuell) in die eigenen Daten integriert wurden. Häufig findet hierbei eine Kombination beider Varianten statt.[19]

Einen zusätzlichen Mehrwert bieten Dienste wie der *Entity Facts*-Dienst der GND, über den Informationen aus mehreren Quellen zu einem Identifikator zugänglich gemacht sind.[20] Ein anderes Beispiel sind die diversen Abfragemöglichkeiten von correspSearch über die von diesem Dienst bereitgestellte Web-API.[21] Der URI von Friedrich Schleiermacher kann hier zum Beispiel dazu verwendet werden, eine Liste der Briefe und Gegenbriefe Schleiermachers zu generieren, die neben der Darstellung im Portal über die API auch in spezifischen Formaten (z. B. TEI-XML oder CSV) ausgeliefert werden kann. Vereinzelt können URIs externer Ressourcen zudem bereits dazu verwendet werden, um granulare Abfragen der originären Daten einzelner Editionen durchzuführen. So stellt etwa die »Carl Maria von Weber Gesamtausgabe« eine REST-Schnittstelle[22] zur Verfügung, sodass zum Beispiel nach Dokumenten des Typs Brief angefragt werden kann, die eine spezifische Person erwähnen oder in denen diese als Absender/Empfänger im Weber-Briefcorpus auftritt.

(2) Vernetzung digitaler Editionen untereinander und mit weiteren Ressourcen: Ein sehr niedrigschwelliges Format hierfür ist BEACON.[23] Es ermöglicht es digitalen Editionen in ihren jeweiligen Webportalen Links auf externe Ressourcen für identische Entitäten weitestgehend automatisch einzubinden. Gleichzeitig können digitale Editionen über das Anbieten entsprechender BEACON-Dateien ihre Inhalte als Ressourcen für Dritte eindeutig identifizieren und bereitstellen.

(3) Anbieten von Resolverdiensten für spezifische externe URIs innerhalb individueller Editionsportale: Vereinzelt bieten Editionsportale bereits die Möglichkeit (neben der vorgenannten Weber-Edition z. B. die Alfred Escher-Briefedition[24]), GND-Identifikatoren als direkten Einstieg und Link zu spezifischen Registereinträgen zu verwenden, ohne hierbei auf die jeweils projektinternen Identifikatoren zurückgreifen zu müssen.

(4) Bereitstellung der Registerdaten einzelner Editionen als Linked Open Data: Ein weiteres Anwendungsfeld von Normdaten stellt die noch recht junge Entwicklung hin zur Bereitstellung der Registerdaten einzelner Editionen als Linked Open Data dar. So bietet zum Beispiel die Edition »Philipp Hainhofer. Reiseberichte & Sammlungsbeschreibungen 1594–1636« die Registerinhalte als RDF-Daten zum Download an,[25] wobei die Identifikatoren externer Wissensbasen die Datenintegration erleichtern.

Anforderungen digitaler Editionen an externe Wissensbasen

Für die Nutzung externer Wissensbasen durch digitale Editionen ist es unerlässlich, dass die unter einem URI (idealiter über *content negotiation*) bereitgestellten Informationen sich immer auf die gleiche Entität beziehen, der Verweis auf den Datensatz zu z. B. einer Person folglich auch verlässlich Informationen zu exakt dieser bereitstellt.

Weiterhin werden aber auch Dienste benötigt, welche die Identifikatoren diverser Wissensbasen, hier vor allem von Normdateien, aggregieren und als Einstiegsknoten in diesen Graphen genutzt werden können. Anders als bei bestehenden Diensten wie VIAF sollten solche aggregierenden Dienste nicht weitere gleichermaßen redundante und potenziell instabile Identifikatoren erzeugen, sondern die bereits existierenden Normdaten-Identifikatoren als Schlüssel einsetzen.[26]

Eine weitere Anforderung besteht in der Bereitstellung von ausreichend individualisierenden Informationen in einem Datensatz als Entscheidungsgrundlage für eine mögliche Verknüpfung in den eigenen Forschungsdaten. Auch hier bestehen zwischen den Datenbeständen externer Wissensbasen erhebliche Unterschiede. Auch ursprünglich rein bibliothekarische Normdateien wie die GND enthalten oftmals nicht ausreichend individualisierte und damit nicht voneinander unterscheidbare Datensätze zu Personen. Dies kann dazu führen, dass in den Daten digitaler Editionen Verknüpfungen zu einer nicht gemeinten Person hergestellt werden, oder aber auf eine entsprechende Verknüpfung gänzlich verzichtet wird.

Editionsprojekte benötigen zudem Wege, um neue Datensätze in externen Wissensbasen, insbesondere der GND, anzulegen wie auch bestehende Datensätze zu ergänzen und/oder zu korrigieren. Gerade Editionen enthalten oft Entitäten, die in den originären Entstehungszusammenhängen externer Wissensbasen, im Falle der GND also die bibliothekarische Formal- und Sacherschließung, bislang nicht berücksichtigt oder benötigt wurden. Die zum Teil bereits erfolgte und nun nochmals intensivierte Öffnung von Wissensbasen wie der GND bringt für beide Seiten große Vorteile, da Editionen fachwissenschaftlich abgesicherte, hochqualitative Daten und Verweise beisteuern können und im Gegenzug dringend benötigte URIs für die Vernetzung mit weiteren Ressourcen erhalten.

Ausblick

Die vorherigen Abschnitte verdeutlichen, welche Aspekte für Text+ an Bedeutung gewinnen werden. In den ersten Projektmonaten wird es wichtig sein, die vielfältigen Anforderungen der Text+ Community in ein konkretisiertes Arbeitsprogramm zu übersetzen, in dem die benötigten Dienstleistungen und Services spezifiziert sind. Ein wichtiger Baustein besteht im Aufbau einer GND-Agentur in Text+ in enger Kooperation und Abstimmung mit der GND-Zentrale an der DNB.

Aufbau einer GND-Agentur für sprach- und textbasierte Forschungsdaten

Neue Forschungsanforderungen machen Anpassungen der zentralen Infrastruktur und der gemeinsamen Regeln der GND notwendig. Gleichzeitig müssen Partner mit engem Kontakt zu den jeweiligen Fachdisziplinen Forschungsprojekte in Fragen der Normdatenarbeit unterstützen. Diese Daueraufgabe im Kontext des digitalen Wandels auch in den Finanzierungsplänen der Institutionen zu verwirklichen, ist nur ein Baustein neben vielen weiteren in der Gesamtentwicklung hin zu einer nachhaltigen Forschungsdateninfrastruktur. In Text+ wird dieser Aufgabenkomplex insbesondere von den beiden mitantragstellenden Bibliotheken, der Deutschen Nationalbibliothek (DNB) und der Niedersächsischen Staats- und Universitätsbibliothek Göttingen (SUB) getragen. Im Laufe des Projekts implementiert die SUB in enger Zusammenarbeit mit der DNB eine GND-Agentur für sprach- und textbasierte Forschungsdaten als innovativen Dienst, über den die GND in Übereinstimmung mit der GND-Strategie der DNB sowohl in ihrem Umfang als auch hinsichtlich der Anforderungen aus der Forschung community-geleitet erweitert wird.

Niedrigschwellige Beteiligungsmöglichkeiten für Forschende schaffen

Das wohl wichtigste Ziel liegt darin, Forschungsprojekten der Text+ Datendomänen einen möglichst niedrigschwelligen Einstieg in die Anwendung von Normdaten und eine aktive Beteiligung zu ermöglichen, ohne den Anspruch an die Verlässlichkeit von Normdaten zu unterminieren. Hierzu zählen (1) einsteigerfreundliche und intuitive Möglichkeiten, fehlende Entitäten, Eigenschaften und Fehler zu melden oder unterstützt durch geeignete Werkzeuge und Qualitätssicherungsmethoden selbständig Änderungen vorzunehmen, (2) die systematische Erweiterung der Datenbasis um bislang unterrepräsentierte Entitätstypen und Eigenschaften sowie (3) die Verständigung auf notwendige Anpassungen bzw. Erweiterungen der gemeinsamen Regeln und des Datenmodells im Rahmen der dafür zuständigen Gremien und Arbeitsgruppen – z. B. um die Auffindbarkeit, die Datenprovenienz und die Disambiguierbarkeit von Normdaten für Forschungskontexte zu optimieren. Die GND-Agentur wird die Anforderungen der Text+ Community in den entsprechenden GND-Gremien vertreten und kann auf diese Weise als eine Verbindungsstelle für die NFDI in diese Gremien fungieren. Die Agentur wird darüber hinaus in Zusammenarbeit mit der GND-Zentrale Forschende und Forschungsprojekte mit Fortbildungsangeboten und Leitlinien unterstützen, um die Kuratierung ihrer Forschungsdaten und zugehöriger Metadaten im Datenlebenszyklus vom ersten Schritt an entsprechend aufzusetzen.

Terminologie-Mappings unterstützen

Ab einem gewissen Grad von Fachspezifik ist es in der Regel sinnvoller, Daten in verschiedenen Wissensbasen außerhalb der GND zu verwalten und die GND dabei als verbindendes und vermittelndes Element – als Hub und Wegweiser – einzusetzen. Dies gilt sowohl für bestehende Fachdatenbanken, Fachthesauri und internationale Datenhubs, deren vollständige Integration in die GND weder sinnvoll noch machbar ist, als auch für den Aufbau neuer Wissensbasen durch Text+ und die anderen Konsortien. Besonders wichtig für die Interoperabilität zwischen den verschiedenen Disziplinen in der NFDI wird der Aufbau eines Netzwerks im Bereich der Terminologien werden. Dafür kann die GND Knotenpunkte in Form genereller Allgemeinbegriffe bieten.

Auf diese Weise kann ihre Konnektivität mit Ressourcen über Text+ hinaus und in die Gesamt-NFDI hinein ausgebaut werden. Um dieses Ziel zu erreichen, gilt es, den Vernetzungsgrad der GND mit relevanten Quellen durch eine deutliche Steigerung der Anzahl wechselseitiger Links auszubauen und leicht nutzbare Services und Tools zum Melden und Verwalten solcher Mappings anzubieten. Dafür müssen Wege der teilweise nachgelagerten automatischen und semi-automatischen Verknüpfung mit der GND weiterentwickelt und vereinfacht werden.

Ziel von Text+ ist es, hier konkret Konkordanzen zwischen Terminologien aufzubauen, die für die text- und sprachbasierte Forschung von besonderer Relevanz sind. Um solche Mappings effizient erstellen und leicht nutzen zu können, werden integrierende Schnittstellen für Menschen und Maschinen benötigt, die die verteilt gemanagten Teilgraphen (GND, Fachthesauri, Fachdatenbanken) als Einheit präsentieren und in einer Art »One-Stop-Shop« komfortabel zugreifbar machen. Die föderierte Daten- und Metadateninfrastruktur von Text+ mit ihren Werkzeugen zum Schema- und Terminologie-Mapping kommt hier ebenso zum Einsatz wie auch maschinelle Verfahren. Sowohl im Rahmen der Kooperation der Geistes- und Kulturwissenschaften als auch im übergreifenden Kontext der NFDI-Sektion *(Meta-)data*, *terminologies*, *provenance* wird sich Text+ aktiv einbringen. Text+ versteht sich als Motor für die Integration der GND nicht nur als Normdatei, sondern auch als Infrastrukturkomponente in die Entwicklung der einen, alle Wissenschaftsdisziplinen umfassenden NFDI.

Literatur

- BALZER, Detlev, Barbara K. FISCHER, Jürgen KETT, Susanne LAUX, Jens M. LILL, Jutta LINDENTHAL, Mathias MANECKE, Martha ROSENKÖTTER und Axel VITZTHUM 2019. *Das Projekt »GND für Kulturdaten« (GND4C)*. In: OBIB – Das offene Bibliotheksjournal, Ausgabe 2019, 4. Verfügbar unter: https://doi.org/10.5282/o-bib/2019H4S59-97
- BRÜNGER-WEILANDT, Sabine, Kai-Christian BRUHN, Alexandra W. BUSCH, Erhard HINRICHS, Gerald MAIER, Johannes PAULMANN, Andrea RAPP, Philipp VON RUMMEL, Eva SCHLOTHEUBER, Dörte SCHMIDT, Torsten SCHRADE, Holger SIMON, Regine STEIN und Elke TEICH, 2020. *Memorandum of Understanding by NFDI Initiatives from the Humanities and Cultural Studies* [Zugriff am: 08. Dezember 2021]. Zenodo. Verfügbar unter: https://doi.org/10.5281/zenodo.4045000
- DUMONT, Stefan, 2020. Kommentieren in digitalen Brief- und Tagebuch-Editionen. In: Wolfgang LUKAS und Elke RICHTER, Hrsg. *Annotieren, Kommentieren, Erläutern* [online]. De Gruyter, S. 175–194, hier S. 185. Beihefte, 47. [Zugriff am: 14. November 2021]. DOI https://doi.org/10.1515/9783110576788-010
- DUMONT, Stefan, Ingo BÖRNER, Dominik LEIPOLD, Jonas MÜLLER-LAACKMAN und Gerlinde SCHNEIDER, [2019–2020]. Corresponde Metadata Interchange Format. In: *Encoding Correspondence. A Manual for Encoding Letters and Postcards in TEI-XML and DTABf* [online]. Berlin [Zugriff am: 14. November 2021]. Verfügbar unter: https://encoding-correspondence.bbaw.de/v1/CMIF.html
- DUMONT, Stefan, 2016. correspSearch – Connecting Scholarly Editions of Letters. In: *Journal of the Text Encoding Initiative* [online]. Issue 10. [Zugriff am: 14. November 2021]. DOI 10.4000/jtei.1742. Verfügbar unter: http://journals.openedition.org/jtei/1742
- ĎURČO, Matej und Menzo WINDHOUWER, 2014. From CLARIN component metadata to linked open data. In: *Proceedings of the 3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*. Reykjavik, co-located with LREC, 26.–31. Mai 2014: ELRA, S. 24–28.
- HAMP, Birgit und Helmut FELDWEG, 1997. GermaNet – a Lexical-Semantic Net for German. In: *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Madrid, 1997.
- HENRICH, Verena und Erhard HINRICHS, 2010. GernEdiT – The GermaNet Editing Tool. In: *Proceedings of the Seventh Conference on International Language Resources and Evaluation* (LREC 2010). Valletta, Malta, Mai 2010, S. 2228–2235.
- HOTSON, Howard und Thomas WALLNIG, Hrsg., 2019. *Reassembling the Republic of Letters in the Digital Age: Standards, Systems, Scholarship*. Göttingen: Göttingen University Press. DOI https://doi.org/10.17875/gup2019-1146
- IGLESIA, Martin de la und Mathias GÖBEL, 2014. From Entity Description to Semantic Analysis: The Case of Theodor Fontane's Notebooks. In: *Journal of the Text Encoding Initiative* [online]. 28. Dezember 2014. Issue 8. [Zugriff am: 14. November 2021]. DOI 10.4000/jtei.1253. Verfügbar unter: http://journals.openedition.org/jtei/1253
- KOEPLER, Oliver, Torsten SCHRADE, Steffen NEUMANN, Rainer STOTZKA, Cord WILJES, Ina BLÜMEL, Christian BRACHT, Tobias HAMANN, Susanne ARNDT und Johannes HUNOLD, 2021. *Sektionskonzept Meta(daten), Terminologien und Provenienz zur Einrichtung einer Sektion im Verein Nationale Forschungsdateninfrastruktur (NFDI) e.V.* [Zugriff am: 08. Dezember 2021]. Zenodo. Verfügbar unter: https://doi.org/10.5281/zenodo.5619089
- RIßLER-PIPKA, Nanette, Raisa BARTHAUER, Stefan BUDDENBOHM, José CALVO TELLO, Sonja FRIEDRICHS und Lukas WEIMER, 2021. *Community Involvement in Research Infrastructures: The User Story Call for Text+ (1.0.0)* [Zugriff am: 08. Dezember 2021]. Zenodo. Verfügbar unter: https://doi.org/10.5281/zenodo.5384085
- STADLER, Peter, Marcel ILLETSCHKO und Sabine SEIFERT, 2016. Towards a Model for Encoding Correspondence in the TEI: Developing and Implementing. In: *Journal of the Text Encoding Initiative* [online]. 8 September 2016. Issue 9. [Zugriff am: 14. November 2021]. DOI 10.4000/jtei.1433. Verfügbar unter: http://journals.openedition.org/jtei/1433
- TRIPPEL, Thorsten und Claus ZINN, 2016. Enhancing the Quality of Metadata by using Authority Control. In: *LDL 2016 – 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources*, Portorož, Slovenia.

Anmerkungen

1 https://www.text-plus.org/

2 https://www.text-plus.org/ueber-uns/fachverbaende/; https://www.text-plus.org/ueber-uns/weitere-partner/

3 Die Diskussion und gemeinsame Vorbereitung auf die NFDI wurde 2018 mit der Workshopreihe »Wissenschaftsgeleitete Forschungsinfrastrukturen für die Geistes- und Kulturwissenschaften in Deutschland« eingeleitet und setzt sich fort in der Organisierung der NFDI-Initiativen NFDI4Culture, NFDI4Memory, NFDI4Objects und Text+ über ein Memorandum of Understanding (Brünger-Weilandt et al. 2020), auf dessen Grundlage sie aktiv ihre Zusammenarbeit gestalten.

4 https://www.text-plus.org/forschungsdaten/user-stories/

5 https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html

6 https://wiki.dnb.de/display/GND/GND-Entwicklungsprogramm+2017-2021

7 https://wiki.dnb.de/x/dIf9Bw

8 https://orcid.org/; https://viaf.org/

9 https://rameau.bnf.fr/

10 Auch wenn zugehörige Metadaten nach ISO 24622-1 (Component Metadata Infrastructure, CMDI) standardkonform und spezifisch für jeden Datentyp erfasst wurden, gab es keine ausreichenden Inventare für eine Verschlagwortung und Normdatenreferenzen (Trippel und Zinn 2016).

11 Im Fall des SFB 833 zeigte sich, dass gerade diejenigen, die im Rahmen ihrer Promotion oder als studentische Hilfskräfte an Projekten beteiligt waren, nicht in den Normdaten verzeichnet waren. Auch wurden z. B. namensgleiche Autoren aus unterschiedlichen Fächern und Jahrhunderten ausgegeben. Zur Auflösung von Namensgleichheiten war eine Disambiguierung erforderlich, die den weiteren Kontext der Daten einbezog, also etwa das Fachgebiet, die Wirkungsstätte, weitere Publikationen, die mit dem Namen verknüpft sind etc. Für Projektleitende wurde so eine Referenz zu einer gleichnamigen Person aus einem anderen Jahrhundert direkt ausgeschlossen.

12 Dieses Verfahren wurde im SFB 833 bei der Anreicherung der Metadaten verwendet, wodurch die Metadaten, in denen Personen, Institutionen und Orte erscheinen, über die Normdaten mit anderen Datensätzen, Publikationen und Informationen verknüpft wurden. Nach Abschluss des SFB 833 steht ein Named Entity Recognizer (NER) zur Verknüpfung von Named Entities in CMDI-basierten Metadaten in einer verbesserten Version zur Verfügung (siehe https://github.com/SfS-ASCL/BiodataNER).

13 Hier vor allem das RIDE – A review journal for digital editions and resources, aber zum Teil auch Sektionen in den Zeitschriften editio und Variants.

14 Das vergleichsweise junge TEI-Element Correspondence Description mit Unterelementen erlaubt strukturierte Angaben zu Absender, Empfänger, Schreib- und Empfangsort, Datumsangabe sowie bibliografischen Informationen, siehe https://tei-c.org/release/doc/tei-p5-doc/de/html/ref-correspDesc.html. Deren Erfassbarkeit im TEI-Header und nicht nur in Form von Elementen des Brieftextes war zuvor lange ein Desideratum der Special Interest Group Correspondence der TEI-Community und anderer Gruppen (z.B. Hotson 2019) gewesen.

15 Verwendung finden vielfach ORCID- und VIAF-URIs, aber auch ISNI (https://isni.org/) und die Getty Union List of Artist Names (ULAN, https://www.getty.edu/research/tools/vocabularies/ulan/). Geografika werden inzwischen häufig mit Geonames (www.geonames.org/) und/oder dem Getty Thesaurus of Geographic Names (TGN, https://www.getty.edu/research/tools/vocabularies/tgn/) verknüpft. Zu VIAF beachte die Erläuterung in Endnote 26.

16 So lässt etwa auch das Correspondence Description-Element für die Verknüpfung von Entitäten eine ganze Reihe von Auszeichnungsvarianten zu.

17 https://correspsearch.bbaw.de

18 Dies ermöglicht zum einen die projekt- und editionsübergreifende Recherche nach Briefen einer einzelnen Person und zum anderen die temporäre Erzeugung ›virtueller' Briefeditionen auch für solche Personen, für die bislang keine eigenständigen Editionen vorliegen. Dabei ist es nicht zwingend erforderlich, dass CMIF-Dateien mit denselben Wissensbasen arbeiten, solange diese miteinander verknüpft sind. Dies ist bei den verbreiteten Normdateien durch integrierende Dienste wie VIAF relativ gut gewährleistet (Dumont 2016).

19 Im Falle von Einträgen in Personenregistern ist z. B. zu beobachten, dass Teile der in den jeweiligen GND-Datensätzen enthaltenen Informationen, insbesondere zu Geburts- und Sterbeort sowie -datum, in die eigenen Daten integriert werden und um Informationen aus anderen Quellen manuell angereichert bzw. kuratiert werden. Gleichzeitig werden andere Bestandteile, wie z. B. alternative Namensformen, on-the-fly in die Registeransichten eingebunden. Vergleichbares ist für Ortsregister zu beobachten: So werden z. B. die Geokoordinaten für die im Register verzeichneten Orte direkt aus Geonames in die eigenen Daten integriert, alternative Namensformen hingegen über Nutzung der Geo-names-API in die Registeransichten des Portals eingebunden. Innerhalb der Webportale werden die Geokoordinaten dabei z. B. dazu verwendet, um räumliche Visualisierungen der im edierten Text erwähnten Geografika anzubieten (z. B. https://architrave.eu/itinerary.html?lang=de#?tab-id=ParisMap).

20 Entity Facts aggregiert Links, die auf externe Ressourcen für den angegebenen GND-Identifikator verweisen, und stellt die angereicherten Datensätze über eine API zur Verfügung. Aggregationsquellen sind dabei öffentlich verfügbare Wissensbasen und automatisch ausgewertete BEACON-Dateien Dritter. Siehe https://www.dnb.de/entityfacts

21 https://correspsearch.net/de/api.html

22 https://weber-gesamtausgabe.de/api/v1/

23 http://gbv.github.io/beaconspec/beacon.html BEACON ist sehr einfach gestaltet und setzt auf flache, UTF-8 kodierte Dateien. Deren Herstellung ist unter Rückgriff auf die TEI-XML-Daten digitaler Editionen auch mit vergleichsweise geringer technischer Kenntnis z. B. einfach über XSLT realisierbar.

24 https://www.briefedition.alfred-escher.ch/uber-die-edition/technische-grundlagen/, Abschnitt »Schnittstellen«.

25 Siehe https://hainhofer.hab.de/informationen-zur-edition/downloads

26 VIAF setzt derzeit bereits über 50 Wissensbasen durch Clustering-Verfahren automatisch miteinander in Verbindung. Von Nutzenden wird VIAF oftmals als eine weitere Normdatei verstanden: So werden VIAF-Identifikatoren in DH-Projekten häufig zum gleichen Zweck und mit den gleichen Erwartungen (vor allem hinsichtlich Persistenz) eingesetzt wie z.B. GND-URIs, obwohl VIAF eine andere Zielsetzung hat und die Identifikatoren durch das regelmäßige Re-Clustering semantisch instabil sind. Gleichzeitig unterstreicht die weit verbreitete Nutzung von VIAF-URIs innerhalb dieser Community of practice jedoch die Bedarfe nach solchen Diensten. Hier ist ein verstärkter Dialog notwendig, um die Potenziale und Limitierungen etwa von VIAF deutlicher zu artikulieren und Best practices der Dienstnutzung zu vermitteln.

Verfasser*innen

Jürgen Kett, Deutsche Nationalbibliothek, Adickesallee 1, 60322 Frankfurt am Main, j.kett@dnb.de
Foto: DNB, Stephan Jockel

Christoph Kudella, Georg-August-Universität Göttingen, Niedersächsische Staats- und Universitätsbibliothek Göttingen, Papendiek 14, 37070 Göttingen, kudella@sub.uni-goettingen.de
Foto: Johannes Biermann

Andrea Rapp, Computerphilologie und Mediävistik, Institut für Sprach- und Literaturwissenschaft, Technische Universität Darmstadt, Marktplatz 15, 64283 Darmstadt, rapp@linglit.tu-darmstadt.de
Foto: Katrin Binner

Regine Stein, Stellv. Abteilungsleitung Forschung und Entwicklung, Forschungsinfrastrukturen, Text+ Infrastructure / Operations Speaker, Georg-August-Universität Göttingen, Niedersächsische Staats- und Universitätsbibliothek Göttingen, Papendiek 14, 37070 Göttingen, regine.stein@sub.uni-goettingen.de
Foto: Johannes Biermann

Dr. Thorsten Trippel, Seminar für Sprachwissenschaft, Eberhard Karls Universität Tübingen, Wilhelmstraße 19, 72074 Tübingen, thorsten.trippel@uni-tuebingen.de
Foto: privat