Kontrollierte Vokabulare, Thesauri, Klassifikationen, Normdaten? - Ein Ordnungs- und Bewertungssystem für wissenschaftliche Vokabulare: Das Register für historische und objektbezogene Vokabulare und Normdaten (R:hovono)

Wegener, Marius
marius.wegener@bibliothek.uni-halle.de
Historisches Datenzentrum Sachsen-Anhalt, Martin-Luther-Universität Halle-Wittenberg, Deutschland
ORCID: 0009-0007-6782-1865

Freytag, Julian
julian.freytag@geschichte.uni-halle.de
Historisches Datenzentrum Sachsen-Anhalt, Martin-Luther-Universität Halle-Wittenberg, Deutschland
ORCID: 0009-0002-0622-2184

Liebing, Katja
katja.liebing@geschichte.uni-halle.de
Historisches Datenzentrum Sachsen-Anhalt, Martin-Luther-Universität Halle-Wittenberg, Deutschland
ORCID: 0009-0001-1624-6465

Möller, Katrin
katrin.moeller@geschichte.uni-halle.de
Historisches Datenzentrum Sachsen-Anhalt, Martin-Luther-Universität Halle-Wittenberg, Deutschland
ORCID: 0000-0003-4090-5667

Simons, Olaf
olaf.simons@pierre-marteau.com
Historisches Datenzentrum Sachsen-Anhalt, Martin-Luther-Universität Halle-Wittenberg, Deutschland
ORCID: 0000-0001-9230-4666

Purschwitz, Anne
anne.purschwitz@geschichte.uni-halle.de
Historisches Datenzentrum Sachsen-Anhalt, Martin-Luther-Universität Halle-Wittenberg, Deutschland
ORCID: 0000-0002-2754-8792

Zielsetzung

Die Bedeutung von Vokabularen und Normdaten hat sich in der digitalisierten Welt intensiv gewandelt. Sie besitzen wesentliche Funktionen bei der Entwicklung Künstlicher Intelligenz und Natural Language Processings (Ehrmann 2024), aber auch bei der Datenkuration, -erschließung,-vernetzung und letztlich bei der Datenanalyse. Mit der Digitalisierung steigt daher auch das Interesse an Vokabularen und Normdaten rasant. Gleichzeitig wächst die terminologische Unübersichtlichkeit. Dabei ist heute nicht immer klar, was genau unter kontrollierten Vokabularen, Thesauri oder Ontologien zu verstehen ist, wo Schwerpunkte, Zielsetzungen oder Nutzungsoptionen genau liegen.

Daher werden in vielen Forschungs-, Sammlungs- und Erschließungsprojekten historisch arbeitender Disziplinen direkt oder indirekt kontrollierte Vokabulare, Thesauri, Klassifikationen oder Normdaten erstellt, die äußerst unterschiedliche Zwecken und Zielrichtungen verfolgen. Für Nachnutzende wird es daher immer wichtiger, einen schnellen Eindruck von der Art und Zusammensetzung von Vokabularen zu gewinnen und sich über die Einsatzmöglichkeiten zu informieren. Besonders in den geisteswissenschaftlichen Disziplinen wurden (heute abgeschlossene und in Büchern publizierte) Kategoriensysteme und Taxonomien mit viel Aufwand entwickelt, stehen momentan aber noch nicht als Linked Open Data oder in anderen offenen Formaten zur Verfügung. Andere Projekte verfügen nicht über die Ressourcen oder technischen Systeme, um in der Forschung entstandene Vokabulare einer größeren Community zur Verfügung zu stellen. Dadurch sind grundlegende Arbeiten wenig sichtbar. Erarbeitete Vokabulare sind somit schwer aufzufinden oder können wegen mangelnder Offenheit nicht gemäß der FAIR-Prinzipien nachgenutzt werden. Gleichzeitig fehlen zentrale Informationen über Nachnutzungsmöglichkeiten. Die Arbeitsgruppe Data Connectivity (TA2) des Konsortiums NFDI4Memory hat darum ein Datenmodell entwickelt, mit dem die Erfassung kontrollierter Vokabulare und Normdaten der historisch arbeitenden Disziplinen geleistet wird, mit dem sie diese Lücke schließen möchte.

Das Register historischer und objektbezogener Vokabulare und Normdaten (R:hovono) soll in Zukunft digital vernetzte Forschungsdateninfrastrukturen unterstützen und zwar dadurch, dass es einen Überblick über möglichst viele einschlägige Vokabulare bietet, die relevant für die historisch arbeitende Community sind und die verstreut vorliegenden Informationen zusammenführt. Mit dem Register verbunden ist eine begriffliche Schärfung von Terminolo-

gien über Vokabulare, kombiniert mit einem Bewertungssystem, das Nutzer*innen einen Eindruck von der Qualität und von Verwendungsszenarien verschaffen kann. In diesem Sinne soll hier für die spezifische Community der historisch arbeitenden Disziplinen ein Arbeitsinstrument geschaffen werden, das zugleich Daten an übergreifende Register mit stark reduzierter Informationsbreite liefert und diese ergänzt (Cimiano et al. 2020).

Der Vortrag soll einen Überblick zu den methodischen Überlegungen geben und die Mechanismen der Qualitätssicherung und "Bewertung" diskutieren.

Methode

Zur Erstellung unseres Registers war es essentiell, die spezifischen Bedarfe der Community zu ermitteln. Darum wurde diese bereits im Rahmen der Antragsphase von NFDI4Memory umfassend mit einbezogen. In 95 Problem-Stories wurden Normdaten, Vokabulare und ihre technische Verfügbarkeit sowie die Rolle von Metadaten als wichtiger Anforderungspunkt beschrieben (NFDI4Memory, 2021). Die hohe Bedeutung von Terminologien spiegelt sich auch in der interdisziplinären Querschnittssektion „Metadaten, Terminologien, Provenienz" wieder. Zur Bedarfsermittlung führten wir eine Umfrage bezüglich Arbeitsweisen, Erstellungsprinzipien und Nutzungsmöglichkeiten von Vokabularen mit den Participants sowie Tiefeninterviews durch, deren Ergebnisse wir im Vortrag vorstellen möchten. Auf diese Weise erfolgte eine Analyse von Erfahrungen, Schwierigkeiten, Mängeln und Wünschen im Kontext von Terminology Services (vgl. auch Mayr 2006).

Die daraus ermittelten Bedarfe der Participants verwiesen in der Breite auf die Schaffung einer hohen Datenqualität und gezielten Datenkuration, Beratung und Wissenstransferleistungen. Zum Teil wurden auch Dienstleistungen zur GND gewünscht, sowie einige technische Services und Linked Open Data-Angebote.

Das Register R:hovono greift diese Bedarfe auf. Ziel ist es, nachgewiesene Vokabulare durch das Register für die Community sichtbarer und besser zugänglich zu machen. Dies kann als Basis für die beratende Arbeit von Daten-Stewards, Datenkurator*innen und GND-Agenturen und die weitere Kooperation im Rahmen der Nationalen Forschungsdateninfrastrukturen dienen. Mithilfe der in Entstehung begriffenen GND-Agentur des Konsortiums NFDI4Memory für Geschichtswissenschaft, Normdaten und Datenkuration am Historischen Datenzentrum Sachsen-Anhalt, soll eine weitere Typisierung und Klassifikation der Vokabulare erfolgen, um Interessierten einen Überblick über Möglichkeiten der Nachnutzung und Qualitätsstandards zu erleichtern.

Zudem sollen die Daten an das übergreifende Verzeichnis "Basic Register of Thesauri, Ontologies & Classifications (BARTOC)" mit interdisziplinären Nachweisen von Thesauri, Ontologien und Klassifikationen (https://bartoc.org/) ausgeliefert werden. Dort erfasste Metadaten wurden regulär in R:hovono integriert. BARTOC ist bei bibliothekarischen Einrichtungen schon etabliert und ermöglicht eine nachhaltige interdisziplinäre Übersicht über die fachliche Vielfalt von erschließenden Vokabularen. Es hat aber keine explizite Ausrichtung auf historisch arbeitende Disziplinen und erfasst einen Kerndatensatz. Eine fachgebundene Nutzung benötigt eine vertiefte Erschließung. Dadurch können viele für diese spezifische Community wichtige Informationen nicht inkludiert werden. Diese Lücke schließt nun R:hovono, dessen Strukturierung sich explizit nach den Bedarfen der historisch arbeitenden Community richtet. Zu dieser Community gehören sowohl die klassischen GLAM-Institutionen als auch Wissenschaftler*innen und Bürgerwissenschaftler*innen. Über die entstehende GND-Agentur wurden bereits viele Kontakte auch zu den Citizen Science aufgebaut, wie beispielsweise zum Verein für Computergenealogie und dessen Geschichtlichem Ortsverzeichnis (GOV). Allein dieses Beispiel zeigt, welche relevanten Vokabulare auch in den Citizen Science bestehen. In etwas anderer Hinsicht auch als Verzeichnis wirksam ist die Linked Open Data Cloud, die ebenfalls fachübergreifend Vokabulare erfasst, allerdings für eine begrenzte Anzahl von Domänen. Hier fehlen historisch orientierte Fächer als eigene Domäne. Eine Ausdifferenzierung der Informationen im Sinne der dort etwa etablierten Linguistic Linked Open Data Cloud ist wünschenswert (Chiarcos et al. 2016).

Aufbau und Gliederung des Erfassungsbogens

Aufgrund der Zielsetzung, bessere Bewertungs- und Einschätzungskriterien für Vokabulare zu schaffen, wurden die Metadaten in Bezug auf BARTOC und die Lider-Projekte in ein deutlich erweitertes Spektrum an Informationen erhoben. Sie zielen vor allem auf eine genauere Übersicht zu den verwendeten und hinterlegten Quelldaten (Nachnutzungsmöglichkeiten für historische Forschungen), Zeitstempeln, inhaltlichen Ausrichtungen und vor allem Crosskonkordanzen mit anderen Vokabularen. Selbstverständlich werden die Informationen über Vokabulare im Register als Metadaten frei nachnutzbar veröffentlicht und stehen mit einer CC0 1.0-Lizenz zur Verfügung. Einen zweiten wichtigen Bereich repräsentieren die Formen der Zugänglichkeit und Lizenzen der Weiterverarbeitung der Vokabulare selbst. Im Wesentlichen folgt der Aufbau des Registers diesen Schwerpunktbereichen:

Kopfdaten/Metdadaten

Im Bereich Kopfdaten/Metadaten wird das Vokabular inhaltlich und formal erschlossen, dazu zählen beispielsweise Autorenschaft, Urheberschaft, Publikationsdatum, Absicherung der technischen Umsetzung, Kurzbeschreibung aber auch die Angabe einer funktionalen E-Mailadresse, um eine langfristige Kontaktaufnahme sicherzu-

stellen. Neben diesen eher allgemeinen Metadaten werden auch Quellen- und Datengrundlagen oder im Vokabular repräsentierte Entitäten abgefragt. Unter Entitäten verstehen wir spezifische Klassen von Begriffen oder Kategorien (z.B. Berufsbezeichnungen, Personen). Durch diese sind eine passgenaue Durchsuchbarkeit und Auswahl von Vokabularen gewährleistet.

Zugänglichkeit/Lizenzen

Im Anschluss an die Erfassung der Metadaten folgen Fragen bezüglich der Zugänglichkeit des Vokabulars. Es werden Arten und Bedingungen des Zugangs erhoben. Diese sind für eine mögliche Vernetzung oder Übernahme durch Forschende wichtig. Neben Fragen zu Lizenzierung, Aktualität, Vollständigkeit oder langfristiger Zugänglichkeit, umfasst der Bereich auch Hilfsangebote für ein besseres Verständnis und damit die Nachnutzbarkeit eines Vokabulars, beispielsweise durch eine Dokumentation oder eine Gebrauchsanweisung.

Crosskonkordanzen

Im Bereich Crosskonkordanzen werden die (teil)automatisierbaren Verbindungen eines Vokabulars zu anderen kontrollierten Vokabularen oder Normdaten auf Begriffsebene oder übergeordnete Kategorien erfasst.

Bereitstellung der Daten

Hier werden technische Standards dargestellt, von Transferprozessen über Ein- und Ausgabeformate, die Software, mit der das Vokabular bereitgestellt wird. API-Schnittstelle oder SPARQL-Endpunkte werden, falls vorhanden, erhoben.

Ausblick und optionale Angaben

Im letzten Abschnitt gibt es die Möglichkeit, in einem Freifeld den Bearbeitungsstand des Vokabulars anzugeben, beispielsweise auch erscheinende Versionen, Übersetzungen oder Aktualisierungen.

Erfassungsprozess und Ausgabe in RDF

Zur Adressierung unterschiedlicher Usergruppen erfolgt die Abfrage der Daten sowohl über eine strukturierte Eingabemaske in LimeSurvey als auch über den direkten Eintrag in der Wikibase-Instanz FactGrid. Das in LimeSurvey verwendete „klassische“ Umfrageformat ermöglicht den Beteiligten einen einfachen Zugang und stellt sicher, dass alle wesentlichen Informationen angegeben werden müssen. Von Vorteil ist außerdem, dass eine (semi)automatische Übertragung in FactGrid gewährleistet ist. Daten werden über FactGrid gehostet und ausgeliefert. Dies ermöglicht letztlich die Nutzung vielfältiger und komplexer Abfragen über RDF und SPARQL sowie die Bereitstellung in zahlreichen Datenformaten.

Um eine möglichst barrierearme Eintragungsmöglichkeit zu gewährleisten, haben wir ein Codebuch veröffentlicht (Freytag et al. 2024a). Außerdem bieten wir an, die Eintragung gemeinsam mit den Datengeber*innen online durchzuführen. Wir haben zusätzlich eine Dokumentation veröffentlicht, um unser Register und seine Ziele für die Community nachvollziehbar darzustellen (Freytag et al. 2024b).

Aufgrund der Standardisierung lassen sich die in LimeSurvey generierten Daten in eine Wikibase-Listeneingabe überführen. In der Eingabe wurden bereits die Property-Nummern der FactGrid-Instanz als Identifier in den einzelnen Eingabefeldern genutzt. Die folgende Abfrage listet diese Properties mit den auf ihnen liegenden Nutzungshinweisen und den einzelnen in Selektionen vorgegebenen Antwortoptionen: https://tinyurl.com/28rrusjy.

FactGrid soll einerseits als Repertorium für Metadaten zu Vokabularen, aber auch als Hintergrundressource für das Interface genutzt werden. Die Wikibase-Datenbank wird Feldinformationen entsprechend der spezifizierten Anfragen der Nutzenden auf dem in Entwicklung befindlichen Web-Interface einspielen. Durch die Abfragen werden Datenpakete erzeugt, die sich als csv, tsv oder json (CC0 1.0-Lizenz) herunterladen lassen.

Ein zusätzlicher Nutzen und Service entsteht durch die optionale Einspielung von Vokabularen in die Wikibase-Instanz. Auf diese Weise können sowohl einfache Metadaten von extern vorliegenden Vokabularen verwaltet, als auch komplette Vokabulare inklusive ihrer Kategorien im Objektgefüge der Plattform integriert werden. Dadurch ist auch der einfache Download der gehosteten Vokabulare möglich, der es den Nutzenden beispielsweise ermöglicht, diese für Reconceiling mit OpenRefine zu nutzen oder aber auch zur Erschließung ihrer Forschungsdaten. Unsere ersten Erfahrungen zeigen, dass sich SPARQL-Abfragen sowohl von Vokabularen als auch von Metadaten einfach übertragen lassen.

Es ist in der Wikibase-Instanz möglich, Daten gezielt in Gebrauchskontexten anzufassen. Hierbei werden einzelne Bearbeitungen der jeweiligen Vokabulare mit Angaben zu Bearbeitenden und ihren Überarbeitungen angereichert. Dies soll eine langfristige Erhaltung punktueller Funktionsfähigkeit und eine dynamische, aktuelle Datenlage ermöglichen.

Besonders relevant ist deshalb auch, die Projekte selbst zur Kuratierung der jeweiligen Datenlage zu befähigen und damit Kommunikationsverluste zu minimieren. So kann sich unsere Redaktion auf Beratung, Optimierung von Prozessen und Qualitätssicherung von Arbeitsschritten in der Wikibase-Instanz fokussieren. Das Zusammenwirken von Wikibase-Instanz mit Ein- und Ausgabetools ist innovativ.

Graphdatenbanken werden bei der Etablierung von Standards sehr schnell unhandlich – es gibt keine Eingabeschablonen, sondern sie erzeugen Netze von Verbindungen, die kohärent abgefragt werden müssen. Durch die Eingabe über LimeSurvey und die Datenbankabfrage über ein gesondertes Interface soll Standardisierung bei Datengenerierung und Datenausgabe erzeugt und technikaffine, wie auch inhaltsaffine Bearbeitende optimal unterstützt werden. Im Projekt werden so konsequent vorhandene Ressourcen nachgenutzt und keine neue Software entwickelt.

Für die so gewonnenen Daten möchten wir anschließend eine Bewertungsmatrix entwickeln. Für diese wurden theoretische Vorannahmen gemacht und in ein Stufenmodell überführt, das wir im Vortrag vorstellen möchten. Sie soll aber auch mit den gewonnenen Daten aus der Erhebung validiert werden. Damit möchten wir nicht nur einen Eindruck zu Verwendungszwecken und Datenqualität von Vokabularen vermitteln, sondern diese auch transparenter sichtbar machen. Zugunsten einer breiten Nutzung und Recherchemöglichkeit, werden die Metadaten von R:hovono auf einer separaten Webseite strukturiert und durchsuchbar ausgegeben. Dies kommt den Nutzungsbedürfnissen der breiten Community entgegen.

Fazit

Für die durch die geschichtswissenschaftliche Community in Befragungen geschilderten Problemlagen in Bezug auf Vokabulare und Normdaten, bietet das entwickelte Register einen innovativen Lösungsvorschlag. Durch die zentrale Erfassung von Vokabularen und Normdaten leisten wir einen Beitrag dazu, diese auffindbar, zugänglich, interoperabel und nachnutzbar zu halten. Parallel wurde eine Begriffsschärfung mithilfe der Erarbeitung eines Glossars von Vokabulartypen geleistet. Das Register ist mit dem vorhandenen fächerübergreifenden Verzeichnis BARTOC mithilfe eines Kerndatensatzes verknüpft. Es erbringt aufgrund der fachspezifischen Orientierung und der erweiterten Erfassung von Informationen einen originären Mehrwert für geschichtswissenschaftliche Akteure. Wie dargestellt entspricht dies den kommunizierten Bedürfnissen der Fachcommunity. Um möglichst geringe Beteiligungshürden zu gewährleisten, können Nutzende ihre Vokabulare neben FactGrid auch über das Umfrage-Tool LimeSurvey beitragen. Dort reicht es aus, den Kerndatensatz einzugeben, es sind aber auch umfassende Angaben, zum Beispiel zu Crosskonkordanzen, oder der zugrundeliegenden Methodik möglich. Die bereits erfolgten Einträge sind bei FactGrid abrufbar. Zudem wird aktuell an einem Interface gearbeitet, das ein übersichtliches Durchsuchen von R:hovono ermöglicht.

Bibliographie

Cimiano, Philipp, Christian Chiarcos, John P. McCrae und Jorge Gracia. 2020. *Linguistic Linked Data: Representation, Generation and Applications*. Cham, Switzerland: Springer.

Chiarcos, Christian, John McCrae und Philipp Cimiano. 2013. „Towards Open Data for Linguistics: Lexical Linked Data.“ In *New Trends of Research in Ontologies and Lexical Resources: Ideas, Projects, Systems*, hrsg. von Alessandro Oltramari, 7–25. SpringerLink Bücher. Berlin, Heidelberg: Springer. https://link.springer.com/book/10.1007/978-3-642-31782-8. Zugriff am 24. Juli 2024.

Ehrmann, Maud, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello und Antoine Doucet. 2024. „Named Entity Recognition and Classification in Historical Documents: A Survey.“ *ACM Computing Surveys* 56 (2): 1–47. https://doi.org/10.1145/3604931. Zugriff am 24. Juli 2024.

Freytag, Julian, Katja Liebing, Katrin Moeller, Anne Purschwitz, Olaf Simons und Marius Wegener. 2024. „Codebuch zur LimeSurvey-Umfrage und FactGrid-Eintragung für das Register historischer und objektbezogener Vokabulare und Normdaten (R:hovono)“ 1.0. Zugriff am 26. November 2024. https://doi.org/10.5281/zenodo.11031743.

Freytag, Julian, Katja Liebing, Katrin Moeller, Anne Purschwitz, Olaf Simons und Marius Wegener. 2024. „Dokumentation zum Register historischer und objektorientierter Normdaten und Vokabulare (R:hovono).“ 1.0. Zugriff am 24. Juli 2024. https://doi.org/10.5281/zenodo.11033367.

Mayr, Philipp. 2006. „Thesauri, Klassifikationen & Co - die Renaissance der kontrollierten Vokabulare?“. In *Vom Wandel der Wissensorganisation im Informationszeitalter*, 151–70. Bad Honnef: Bock + Herchen, 2006. DOI: 10.18452/2328. Zugriff am 24. Juli 2024.

NFDI4Memory. 2021. „Problem Stories.“ Zugriff am 24. Juli 2024. https://4memory.de/problem-stories-overview/.

Zeng, Marcia Lei. 2005. „Construction of Controlled Vocabularies: A Primer.“ Kapitel 1-4. Zugriff am 24. Juli 2024. https://marciazeng.metadataetc.org/Z3919/.