# Dokumentation zum Register für historische und objektorientierte Vokabulare und Normdaten (R:hovono)


Julian Freytag, https://orcid.org/0009-0002-0622-2184
Katja Liebing, https://orcid.org/0009-0001-1624-6465
Katrin Moeller, https://orcid.org/0000-0003-4090-5667
Anne Purschwitz, https://orcid.org/0000-0002-2754-8792
Olaf Simons, https://orcid.org/0000-0001-9230-4666
Marius Wegener, https://orcid.org/0009-0007-6782-1865

Kontakt: TA2 NFDI4Memory <ta2-nfdi4memory@geschichte.uni-halle.de>

Eine Arbeit im Rahmen des 4Memory-Konsortiums.[1]

## 1. Einleitung: Ausgangssituation und Zielsetzung

In vielen Forschungs-, Sammlungs- und Erschließungsprojekten historisch arbeitender Disziplinen werden direkt oder indirekt kontrollierte Vokabulare, Thesauri, Klassifikationen oder Normdaten erstellt. Diese können auch für andere Forschende und Institutionen hilfreich sein, um sie für ein eigenes Projekt nachzunutzen, z. B. um Daten zu normieren, zu vernetzen, mit Informationen anzureichern, zu analysieren oder Erkenntnisse aus der Verknüpfung mit eigenen Daten zu gewinnen. Zahlreiche Projekte stellen die innerhalb ihrer Forschungen entstandenen Vokabulare aber nicht regulär einer größeren Community zur Verfügung, was dazu führt, dass grundlegende Arbeiten häufig nicht sichtbar werden. Bereits erarbeitete Vokabulare sind somit schwer auffindbar oder durch ihre fehlende Offenheit nicht den FAIR-Prinzipien gemäß nachnutzbar. Damit geht zentrale Forschungsarbeit verloren! Gelegentlich fehlen auch wichtige Informationen, um sich schnell über Nachnutzungsmöglichkeiten zu informieren.

Aus diesem Bedarf heraus hat die Arbeitsgruppe Data Connectivity (TA2) des Konsortiums NFDI4Memory ein Datenmodell zur Erfassung kontrollierter Vokabulare und Normdaten der historisch arbeitenden Disziplinen entwickelt. Das “Register historischer und objektbezogener Vokabulare und Normdaten (R:hovono)” soll zukünftig digital vernetzte Forschungsdateninfrastrukturen fördern, indem es einen Überblick über möglichst

[1] Diese Arbeit ist im Rahmen des NFDI-Konsortiums 4Memory entstanden (www.4memory.de). Wir danken der Deutschen Forschungsgemeinschaft (DFG) für die finanzielle Unterstützung – Projektnummer 501609550.

viele für die historisch arbeitende Community einschlägige Vokabulare bietet und verstreut liegende Informationen zusammenführt.[2] Ziel ist es, nachgewiesene Vokabulare durch das Register für die Community sichtbar und besser zugänglich zu machen. Dies bietet zudem eine hervorragende Grundlage für die beratende Arbeit von Daten-Stewards, Datenkurator*innen und GND-Agenturen und die weitere Zusammenarbeit im Rahmen der Nationalen Forschungsdateninfrastrukturen. Durch die entstehende GND-Agentur des Konsortiums NFDI4Memory für Geschichtswissenschaft, Normdaten und Datenkuration am Historischen Datenzentrum Sachsen-Anhalt werden eingetragene Vokabulare zudem weiter typisiert und klassifiziert, um Interessierten einen raschen Überblick über Nachnutzungsmöglichkeiten und Qualitätsstandards zu erleichtern.

Ergänzend zur Etablierung eines eigenen Registers besteht mit dem Verzeichnis “Basic Register of Thesauri, Ontologies & Classifications (BARTOC)” bereits ein Register, das Thesauri, Ontologien und Klassifikationen interdisziplinär nachweist (https://bartoc.org/). Dieses, bei bibliothekarischen Einrichtungen bereits gut etablierte Register, gibt einen nachhaltigen interdisziplinären Einblick in die fachliche Breite von erschließenden Vokabularen. Da BARTOC jedoch nicht explizit auf historisch arbeitende Disziplinen ausgerichtet ist und eher einen Kerndatensatz erfasst, können zahlreiche für die Community relevante Informationen dort nicht einfließen. Dies wird durch die Erweiterung des Modellierungsansatzes in R:hovono möglich, dessen Strukturierung aufgrund von Bedarfen und Zielrichtung der historisch arbeitenden Community erarbeitet wurde. Beide Register sollen zukünftig jedoch miteinander korrespondieren.

Ziel ist zudem eine niedrigschwellige, für alle zugängliche und technisch leistungsfähige Arbeitsumgebung. Daher wird das Register eine Dateneingabe einerseits über eine strukturierte Eingabemaske (LimeSurvey) wie auch über den direkten Eintrag in der Wikibase-Instanz FactGrid erleichtern. Alle Daten werden über die Wikibase-Instanz gehostet und ausgeliefert. Dies ermöglicht letztlich die Nutzung vielfältiger und komplexer Abfragen über RDF und SPARQL sowie zahlreiche Datenformate.

## 2. Beteiligte

Das Register wird von der NFDI4Memory Arbeitsgruppe Data Connectivity am Historischen Datenzentrum Sachsen-Anhalt angeboten und gepflegt und kann von Communities genutzt werden. NFDI4Objects wird ebenfalls R:hovono zur Erfassung und Bereitstellung von Vokabularen verwenden.

[2] So zum Beispiel Register zu Vokabularen in Museen und anderen Sammlungsinstitutionen, z.B.: Fachgruppe Dokumentation im Deutschen Museumsbund e. V. (Hg.): Vokabular in der Museumsdokumentation, o. O. 2024, URL: http://museumsvokabular.de/. Bei DigiCult werden momentan über 100 vokabulare und Wertelisten gehostet: DigiCult-Verbund eG (Hg.): DigiCult.xTree, Kiel 2024, URL: https://www.digicult-verbund.de/de/digicultxtree.Weitere Vokabulare führen die Services DANTE und BARTOC auf.

# 3. Zielgruppe

Das Register steht allen Einrichtungen, Institutionen und Forschenden historisch arbeitender und objektbezogener Disziplinen und Infrastrukturen offen. Besonders Participants und Partner*innen von NFDI4Memory, aber auch anderen Interessierten, stehen dabei die Möglichkeiten der beratenden Begleitung und der Eintragung von Informationen in einer Interviewsituation zur Verfügung. Darüber hinaus wird die weitere Community über Calls, Bekanntmachungen und Workshops an der Bereitstellung, Verbreitung und Vernetzung kontrollierter Vokabulare beteiligt, um dadurch eine möglichst umfassende Abdeckung und Bereitstellung zu gewährleisten.

Die Vokabulare selbst können in unterschiedlichsten Granularitäten und Reifegraden vorliegen, es kann sich beispielsweise um Listen, Ontologien, Thesauri, um abgeschlossene und unabgeschlossene, analog oder digital verfügbare Vokabulare handeln.

Eine gewünschte Kontaktaufnahme zu NFDI4Memory erfolgt über die funktionale E-Mail der Arbeitsgruppe: ta2-nfdi4memory@geschichte.uni-halle.de, weitere Informationen können über die Mailingliste der TA2: Data Connectivity [nfdi4memory-ta2@lists.nfdi.de], über den Newsletter von NFDI4Memory [https://4memory.de/] oder den Blog der Task Area Data Connectivity [https://blogs.urz.uni-halle.de/nfdi4memory/] direkt bezogen werden.

Interessierte objektbezogener Vokabulare und Normdaten können sich zur Beratung auch an die Task Area 6 “Qualification, Integration and Harmonisation” des NFDI-Konsortiums NFDI4Objects wenden. Kontaktmöglichkeiten bestehen hier über die E-Mail des Helpdesks [n4o-helpdesk@dainst.de] und die Website: [https://www.nfdi4objects.net/] sowie den Newsletter [https://www.listserv.dfn.de/sympa/info/nfdi4objects].

# 4. Bedarfsermittlung: Welches Register braucht die historische Community?

### Erhebung von Bedarfen durch – Befragung und Interviews mit den Participants

Zur Erstellung eines Registers kontrollierter Vokabulare, das für historisch arbeitende Wissenschaftler*innen sowie Infrastrukturprojekte einen Mehrwert erbringen soll, war es von besonderer Wichtigkeit, die spezifischen Bedarfe innerhalb der Community zu ermitteln. Aus diesem Grund erfolgte bereits in der Antragsphase von NFDI4Memory eine umfassende Einbeziehung der Community. In den 95 verschiedenen Problem-Stories waren Normdaten, Vokabulare, ihre inhaltliche und technische Verfügbarkeit sowie die Rolle von Metadaten ein zentraler Anforderungspunkt an die entstehenden Infrastrukturen.[3] Die

[3] NFDI4Memory (Hg.): Problem Stories, Mainz 2021, URL: https://4memory.de/problem-stories-overview/.

hohe Bedeutung von Terminologien wird nicht zuletzt auch innerhalb der NFDI im interdisziplinären Zusammenwirken der Querschnittssektion “(Meta)daten, Terminologien, Provenienz” sichtbar.[4]

Innerhalb der Task Area Data Connectivity erfolgte nach Projektaufnahme zunächst eine Umfrage zu Arbeitsweisen, Erstellungsprinzipien und Nutzungsmöglichkeiten von Vokabularen bei den Participants. Anschließend konnten auf Basis dieser Erhebung zielführende und auf die jeweiligen Bedarfe zugeschnittene Tiefeninterviews geführt werden, in denen besonders die zum Ausdruck gebrachten Schwierigkeiten, Mängel und Wünsche zum Ausbau von Terminology Services Diskussionen erfuhren. Davon abgeleitet wurden die Zielvorstellungen für das “Register historischer und objektbezogener Vokabulare und Normdaten” entwickelt. Sowohl durch die Umfrage, als auch über die Tiefeninterviews wurden die Bedarfe an Dienstleistungen einer GND-Agentur konturiert sichtbar, die zu kleineren Anteilen aus Services im Bereich von technischen Dienstleistungen, Vernetzungsprojekten und Linked Open Data-Angeboten, in der Mehrzahl aber aus Angeboten für Datenkuration, Beratung und Wissenstransferleistungen bestehen sollen und so eine inhaltliche Passfähigkeit der Daten gewährleisten müssen. Das Register für Vokabulare und Normdaten greift diese Bedarfe auf. Ein Prototyp wurde auf dem NFDI Communityforum im November 2023 in einer ersten Version des Erfassungsbogens vorgestellt und diskutiert.

Basierend auf den Befragungen, Interviews und Diskussionen wurden die Anforderungen konkretisiert und systematisiert und nunmehr mit dem LimeSurvey-Erfassungsbogen und der Modellierung in FactGrid die Grundlage für das Register R:hovono und ein nachhaltiges Angebot geschaffen.[5]

# 5. Der Erfassungsbogen - Wie werden die Vokabulare für das Register erhoben?

## 5.1 Erfassungsmethoden

Für die Erfassung der Daten und ihre Anreicherung mit Normdaten gibt es zwei Zugangsmöglichkeiten: Einerseits kann die Erfassung über die strukturierte Beantwortung von Fragen im Umfrage-Programm LimeSurvey durchgeführt werden, andererseits ist eine direkte Eingabe in der Wikibase-Instanz FactGrid möglich, in der dann jedoch strikt dem Verfahren des Codebuches gefolgt werden sollte, um alle notwendigen Daten selbstverantwortet einzutragen. Letztlich werden alle Daten in FactGrid zusammenge-

[4] NFDI (Hg.): Sektion (Meta)daten, Terminologien, Provenienz, Karlsruhe 2024, URL: https://www.nfdi.de/section-metadata/.

[5] Eine Veröffentlichung zu weiteren Ergebnissen aus unseren Interviews folgt.

führt. Eine noch im Entstehen begriffene Website soll zudem ein strukturiertes Ausgabeformat ermöglichen, um den niedrigschwelligen Zugang auch zu den Arbeitsergebnissen zu erleichtern.

Bei LimeSurvey handelt es sich um ein Programm zur Erstellung und Veröffentlichung von Online-Umfragen, das die Ergebnisse in einer Datenbank erfasst und damit selektive Abfragen und strukturierte Ausgaben ermöglicht. Die Umfrage kann über den Link: [https://umfrage.uni-halle.de/194457?lang=de] aufgerufen werden. Die Verwendung einer 'klassischen' Umfrage bietet den Beteiligten eine einfache Zugänglichkeit und stellt zugleich sicher, dass keine erforderlichen Informationen und Angaben vergessen werden, was letztendlich den Arbeitsaufwand minimiert. Zugleich soll durch diesen niedrigschwelligen Einstieg eine möglichst hohe Beteiligungsbereitschaft der Nutzenden und eine Minimierung des zeitlichen Aufwandes erreicht werden. Sofern alle Informationen vorhanden sind, dauert die Beantwortung des Fragebogens ca. 20 Minuten. Es ist zu empfehlen, sich vorab über das Codebuch einen Überblick über die verankerten Fragen zu verschaffen und die entsprechenden Informationen vorzuhalten.

Die Umfrage besteht aus offenen Fragen (mit Antworten in freien Textfeldern), geschlossenen Fragen (mit Einfach- oder Mehrfachantworten) und kombinierten Fragen (beispielsweise geschlossene Frage mit freiem Textfeld). Zudem sind Filterfragen installiert, die es den Teilnehmenden ermöglichen, unzutreffende Erhebungsbereiche zu umgehen. Für jedes Vokabular muss dabei ein separater Fragebogen ausgefüllt werden, eine gleichzeitige Aufnahme mehrerer Vokabulare ist nicht vorgesehen.

Ein zusätzlicher Vorteil von LimeSurvey besteht darin, dass eine (semi)automatische Übertragung der Eingaben in FactGrid[6] möglich ist. Hierfür sind die dafür erforderlichen P-Nummern an die Antworten der Erfassungsfragen gekoppelt, diese spiegeln die entsprechenden Erfassungsfelder in FactGrid wieder und können dort auch direkt aufgerufen oder abgefragt werden.

Zur Erläuterung und als Ausfüllhilfe für den Erfassungsbogen wurde ein Codebuch[7] verfasst. Dieses soll mögliche Missverständnisse minimieren und zur allgemeinen Verständlichkeit der Fragen beitragen. Am Ende der Dokumentation findet sich ein Glossar mit für die Erstellung des Registers elementaren Begriffen und deren Definitionen, da viele Fachbegriffe in unterschiedlichen Disziplinen heterogene Bedeutungen tragen.[8] Damit sollen innerhalb des späteren Registers eine einheitliche Semantik und innerhalb der Erhebung aussagekräftige Antworten auf gleicher definitorischer Grundlage sichergestellt werden.

[6] Olaf Simons (Hg.): FactGrid. A Database for historians, Gotha 2024, URL: https://database.factgrid.de/wiki/Hauptseite.

[7] Julian Freytag, Katja Liebing, Katrin Moeller, Anne Purschwitz, Olaf Simon und Marius Wegener: Codebuch zur Dokumentation des Registers historischer und objektbezogener Vokabulare und Normdaten (R:hovono), Halle 2024, DOI: 10.5281/zenodo.11031743.

[8] Siehe Kapitel 7.

Für erfahrene Nutzende besteht damit auch die Möglichkeit, die erforderlichen Daten direkt über FactGrid einzutragen. FactGrid ist eine Wikibase Graphdatenbank, die speziell die wissenschaftliche Community der historisch arbeitenden Disziplinen anspricht. Während die Software und die zentralen Aspekte der Datenmodellierung mit dem Wikidata-Projekt parallel laufen, ist die Community hier auf die Nutzung von Klarnamen-Konten verpflichtet. Dies ermöglicht, vergleichsweise transparent organisiert, eine unmittelbare Beteiligung an der kollektiven Datengenerierung und -kuratierung. Voraussetzung ist die Einrichtung eines Nutzerkontos durch Administrator*innen der Datenbank [ta2-nfdi4memory@geschichte.uni-halle.de, Betreff: Nutzerkonto FactGrid]. Kontoinhaber*innen können danach Daten jederzeit eigenhändig auffrischen. Die Datenbank erlaubt es im selben Moment, ganze Vokabulare zu integrieren und eigene Forschungsprojekte mit ihnen auf der Plattform zu unterstützen. Wir weisen bei diesbezüglichem Interesse gerne in die Handhabung der Datenbank ein und bieten Beratungen dazu an.

### 5.2 Erfassung über LimeSurvey

Über den Link [https://umfrage.uni-halle.de/194457?lang=de] können die Nutzer*innen die Umfrage aufrufen, wo sie sich im Folgenden mit maximal 48 Fragen in den fünf Bereichen Metadaten, Zugänglichkeit, Cross-Konkordanzen, Bereitstellung der Daten (Lizenzen) und Ausblick konfrontiert sehen. Eine Zwischenspeicherung des Fragebogens ist möglich, so dass bei Unklarheiten eine spätere Bearbeitung eingefügt werden kann. Erst der Button ‘Absenden’ am Ende des Fragebogens übermittelt die strukturierten Daten. Die Antworten auf die Fragen sind nicht anonymisiert.

Der Erfassungsbogen soll die Übernahme eines Nachweises des entsprechenden Vokabulars in das Register ermöglichen, um Interessierten einen Überblick über das Vokabular zu geben und sie zu einem Urteil zu befähigen, inwiefern es einen Mehrwert für ihre Forschung bietet.

#### 5.2.1 Aufbau und Gliederung des Erfassungsbogens

Um die Crosskonkordanz zu BARTOC zu gewährleisten, gelten die Sacheinträge des Basisregisters gleichzeitig als Minimal- bzw. Kerndatensatz des entstehenden Registers R:hovono. Einträge müssen also zwangsläufig die in BARTOC erfassten Informationen abdecken, um ins Register aufgenommen zu werden. Dieses Kerndatenset (Titel, Typ, Beschreibung, Sprachen, Schlagwörter, Link sowie Zugangsmöglichkeiten, Informationen zu Urheber*innen und Kontaktdaten) wird in LimeSurvey und im Codebuch jeweils mit dem Symbol eines roten Ausrufezeichens (**!**) als Bestandteil des Kerndatensatzes gekennzeichnet. Darüber hinaus werden weitere Informationen abgefragt, um eine möglichst präzise Beschreibung der Vokabulare zu erlangen und so eine bessere Entscheidung über Nutzungsmöglichkeiten treffen zu können.

Bei in Zusammenhang mit den abgegebenen Erfassungsbögen auftretenden Fragen ist die Kontaktaufnahme möglich, um etwaige Unklarheiten zu beseitigen oder gemeinsame Interviewtermine zu vereinbaren. Alle über diese Minimalanforderungen hinausgehenden

Informationen erhöhen die Qualität von Datenangeboten, verbessern die Nachnutzbarkeit und Attraktivität der verzeichneten Vokabulare und werden später in der Web-Repräsentation und auf FactGrid entsprechende Berücksichtigung erfahren.

Die Fragen werden je nach Zweck als offene Fragen mit Freifeldern, als Mulitple Choice oder auch als Kombination offener und geschlossener Fragen gestellt. Ebenfalls enthalten sind Filterfragen, um Fragen, die nur bei bestimmten Vokabularen von Interesse sind, auch nur in diesen Fällen vorzulegen. So wird beispielsweise in Frage 40 erhoben, ob die Nutzendengruppen des Vokabulars bezeichnet werden können; nur wenn diese Frage mit “Ja” beantwortet wird, erscheint im Erfassungsbogen die Frage “Falls es Ihnen bekannt ist, geben Sie hier bitte an, von wem Ihr Vokabular extern bereits genutzt wurde.” Zu beachten ist dieses Prinzip weniger in der LimeSurvey-Umfage als vielmehr bei der Eintragung in FactGrid, wo diese Anpassungsleistungen nicht automatisch erfolgen (weitere Fragen wären: Frage 6 als Bedingung für Frage 7; Beantwortung Frage 30 als Bedingung für Frage 31; Frage 32 als Bedingung für Frage 33).

### 5.2.2 Kopfdaten/Metadaten

Mit den Abfragen im Bereich Metadaten erfolgt eine inhaltliche und formale Erschließung des jeweiligen Vokabulars. Dies beinhaltet die Erhebung von Kontaktinformationen, die zeitliche Einordnung, den Umfang und die Sprachen des Vokabulars. Zudem werden eine inhaltliche Beschreibung in unterschiedlichen Formen, die Vergabe von Schlagwörtern, eine Bestimmung des Vokabular-Typs, die Benennung von Entitäten innerhalb des jeweiligen Vokabulars und die vorhandene Quellengrundlage abgefragt.

Gleichzeitig erfasst das Register Angaben zu Verantwortlichkeiten wie der Autorenschaft, Urheberschaft (Frage 8), die Absicherung der technischen Umsetzung (Frage 9) und eine (funktionale) Kontakt-E-Mail (Frage 12). Die Angabe einer funktionalen E-Mailadresse stellt sicher, dass die Mailadresse gemeinsam innerhalb eines größeren Arbeitsbereiches genutzt wird und eine überindividuelle Erreichbarkeit gegeben ist. Gleichzeitig wird so eine erfolgreiche Kontaktaufnahme langfristiger erwartbar.

Auf der zeitlichen Ebene wird das Jahr der erstmaligen Publikation, bzw. das geplante Publikationsdatum abgefragt (Frage 13), um eine Einschätzung der Aktualität bzw. auf den Zeitpunkt eines möglichen Zugriffs machen zu können.

Die Größe des Vokabulars ist ebenfalls von Bedeutung, um einschätzen zu können, inwiefern das Vokabular für das eigene Anliegen hilfreich sein kann; darum wird die Objektanzahl des Vokabulars abgefragt (Frage 16). Auch die “Tiefe” des Vokabulars wird erfasst, indem die vertikalen Hierarchieebenen abgefragt werden. Diese geben Auskunft über die Granularität (Frage 18).

Für die Anschlussfähigkeit des Vokabulars ist die Frage nach Sprachen, in denen das Vokabular vorhanden ist, zentral (Frage 17). In der Befragung als Antwortmöglichkeiten angelegt sind die am häufigsten erwarteten Sprachen in Form von Multiple-Choice, um eine möglichst große Normierung zur besseren Auswertbarkeit zu erreichen. Weitere

Sprachen können entsprechend des ISO-639-Sprachcodes (Set 2, dreistellig) angegeben werden.[9] Zentral für die Sprachangabe ist nicht die Übersetzung von Einzelbegriffen, sondern die Hauptsprachen des überwiegenden Teils des Vokabulars und seiner möglichen Übersetzungen.

Lediglich zur Erleichterung von Ausgaben in inhaltlichen Zusammenhängen der verschiedenen geisteswissenschaftlichen Konsortien wird die entsprechende Zugehörigkeit in Frage 1 thematisiert. Ebenso wird der (vorrangige) Zweck der Nutzung des Vokabulars (u. a. Quantitative/Qualitative Analysen, Datenkuration, Identifizierung und Matching) über eine Multiple-Choice-Option erhoben (Frage 5), um den späteren Nutzenden einen schnellen Überblick zu geben, für welche Zwecke das Vokabular sinnvoll anwendbar sein könnte.

Darin beinhaltet sind inhaltliche Beschreibungen, die die Betrachter*innen dazu befähigen sollen, die inhaltliche Relevanz eines Vokabulars für ihre Forschungsvorhaben einschätzen zu können. Allgemein zählt dazu der Name inkl. Akronym (Frage 3), eine Kurzbeschreibung des Vokabulars (max. 1500 Zeichen inklusive Leerzeichen) auf Deutsch (Frage 21) und auf Englisch (Frage 22); um auch eine internationale Verständlichkeit zu gewährleisten. Das Vokabular soll zudem in einem Satz zusammengefasst werden (Frage 24), dies dient der späteren Kurzanzeige, die bei Eingabe des Vokabulars als Suchbegriff in FactGrid und auf der Homepage erscheint.

Zur besseren inhaltlichen Suche und Auffindbarkeit werden Schlagwörter für die Vokabulare erfasst, die eine Einordnung in (historische) Wissensfelder ermöglichen (Frage 14). Die Eingabe erfolgt möglichst präzise und wird anschließend durch die Arbeitsgruppe klassifiziert und normiert. Anhand der Tiefenschärfe der übermittelten Erfassungsbögen wird entschieden, welche konkrete Einteilung historischer Wissensfelder letztlich am sinnvollsten genutzt werden kann. Eine Möglichkeit wäre die DFG-Fachsystematik der Wissenschaftsbereiche,[10] gegebenenfalls aber auch eine fachlich erweiterte Systematik. Da die Erfassenden in diesem Zusammenhang frei agieren können und nicht genötigt werden, sich einer bereits bestehenden Klassifikation anzupassen, ergibt sich für uns die Option einer relativ freien Systematisierung bei Vorliegen einer ausreichend großen Menge an Vokabularen, die aber auch mit bestehenden Vokabularen abgeglichen werden kann. Ebenfalls zu den Metadaten gehören Fragen nach der Art des Vokabulars (Frage 4).

Der anschließende Fragenblock richtet sich auf die im Vokabular repräsentierten und beschriebenen Entitäten (Frage 15). Unter einer Entität werden spezifische Klassen von Begriffen oder Kategorien (z. B. Orte, Personen, Berufsbezeichnungen, Musikinstrumente, Verwandtschaftsbeziehungen) verstanden. Die Angabe ermöglicht eine

[9] Library of Congress / Bayerische Staatsbibliothek (Hg.): Sprachencode nach ISO 639, München 2019, URL: https://www.bib-bvb.de/web/kkb-online/rda-sprachencode-nach-iso-639.

[10] DFG (Hg.): Systematik der Fächer und Fachkollegien der DFG für die Amtsperiode 2020-2024, Bonn 2021, URL: https://www.dfg.de/resource/blob/175334/89ba4a3464c99aaea40fdef47367e7b2/fachsystematik-2020-2024-de-grafik-data.pdf.

passgenaue Durchsuchbarkeit und Auswahl von Vokabularen. Ergänzt bzw. erweitert wird die Erhebung durch die Frage nach den Eigenschaften oder Merkmalsausprägungen dieser Entitäten bzw. Begriffe, die der Klassifizierung sowie Kategorisierung zugrunde liegen oder diese definieren (Frage 20).

Ebenfalls zu den Metadaten gehört die Benennung der für das Vokabular elementaren Quellen und Datengrundlagen. Gerade für Forschungs- und Provenienzprojekte ist es sinnvoll, die Quellenbasis von Terminologien und ihre Quellennähe auch hinsichtlich von Zeit- und Ortsnachweisen genau zu dokumentieren. Hier begnügt sich das Register mit ungefähren Mengenangaben. Gleichzeitig sollen die genaueren Grundlagen von Nachweisen und theoretischen Konzepten erfolgen. Dabei unterscheidet das Register Originalbezeichnungen von Quellen und Fachbegriffen. Letztere repräsentieren “aus der Wissenschaft generierte, verallgemeinerte Fachbegriffe” (Codebuch). Daher können historische Fachbegriffe aus heutiger Perspektive durchaus auch Quellenbegriffe sein, weil sie mittlerweile nicht mehr als Fachbegriffe genutzt werden.

### 5.2.3 Zugänglichkeit/Lizenzen

Der Erfassung der Metadaten folgen Fragen zur Zugänglichkeit des Vokabulars. Hier werden die Arten und Bedingungen des Zugangs erhoben. Dieser Bereich ist von hoher Relevanz für die Vernetzung und spätere Übernahme durch andere Forschende.

In diesem Komplex werden generelle Fragen zur Zugänglichkeit, Lizenzierung, Versionierung, Aktualisierung und Dokumentation, Tutorials und Lernmaterialien, Zitation sowie Fragen zu Vollständigkeit, Bearbeitungs- und Ergänzungsoptionen und Erschließungsgrundsätzen thematisiert.

In Hinsicht auf die Verfügbarkeit macht es bspw. für Matchings einen erheblichen Unterschied, ob ein Vokabular nur intern digital verfügbar, in Gänze herunterladbar, nur zum Teil oder ausschließlich als Printpublikation zur Verfügung steht (Frage 26); so kann die Printpublikation erst mit einem erheblichen Umfang an Vorarbeiten für digitale Verknüpfungen nutzbar gemacht werden, aber dennoch für die Community von hoher Bedeutung sein. Für die Beurteilung der Zugänglichkeit ist es wichtig zu wissen, unter welchen lizenzrechtlichen Bedingungen eine Anwendung möglich ist (Frage 27). Für diese Erfassung sind mehrere Lizenzen als Antwortmöglichkeiten angegeben. Der Fokus liegt hierbei auf den verschiedenen offenen Lizenzen, da nur diese den Fair-Prinzipien entsprechen (dazu zählen: CC0 1.0., CC BY 4.0, InC-EDU 1.0 u. a.). Sollte die entsprechende Lizenz sich nicht in der Auflistung finden, ist auch eine Eingabe im Freifeld möglich. Intendiert ist an dieser Stelle, dass sich die Erfassenden, wenn noch nicht geschehen, über eine Lizenzierung ihres Vokabulars verständigen.

Um die Aktualität des Vokabulars einschätzen zu können, wird die Datierung der aktuellen Version abgefragt (Frage 29). Gerade langfristige Zugänglichkeit hängt vom Ausmaß der Datenkuration, Regelwerk und institutioneller Pflege eines Vokabulars ab (Frage 35). Entstammt es einem abgeschlossenen Projekt, ist eine Weiterentwicklung nicht zu erwarten. Wird die Pflege langfristig oder dauerhaft durch eine Organisation

getragen, bilden Erweiterungen, Ergänzungen und Korrekturen vielversprechende Optionen auf qualitativ höherwertige Daten. Ebenfalls für Verknüpfungen und Nachnutzungen von Bedeutung ist die Vollständigkeit des Vokabulars (Frage 34). Hier steht die Überlegung im Raum, wie umfassend oder vollständig alle relevanten Begriffe einer entsprechenden Entität oder eines Themas repräsentiert werden. Gleichzeitig verschaffen diese Informationen einen Eindruck, ob Terminologien für Erweiterungen offen stehen und welche organisatorischen Schritte hierfür in Betracht kommen. In diese Überlegung integriert ist die Frage nach den Arbeitsabläufen und Berechtigungen für Erweiterungen und Korrekturen (Frage 36). Es werden auch die Prinzipien der begrifflichen Erschließung abgefragt. Dies ermöglicht einen Überblick über den Typ des Vokabulars, es gibt an, ob es sich beispielsweise um eine einfache Begriffsliste oder um ein standardisiertes Vokabular mit Regelwerk handelt.

Ein weiterer wichtiger Bestandteil für ein besseres Verständnis und die Nachnutzbarkeit eines Vokabulars sind Hilfsangebote. Hierzu zählen Codebücher bzw. Dokumentation (Frage 30 und 31) oder auch Gebrauchsanweisung (Frage 32 und 33) im Sinne von Lehrmaterialien oder Tutorials.

### 5.2.4 Crosskonkordanz

Die Fragen zu den Crosskonkordanzen erfassen die (teil)automatisierbaren Verbindungen eines Vokabulars zu anderen kontrollierten Vokabularen oder Normdaten auf der Ebene der Begriffe oder übergeordneten Kategorien. Auf diese Weise können auch standardisierte, persistente Normdaten mit flexibleren und spezialisierten Vokabularen durch ein Matching verknüpft und verlinkt werden und so ganz unterschiedlichen Zwecksetzungen dienen. Solche Matchings erleichtern Prozesse der Datenkuration, -anreicherung und -plausibilisierung erheblich und stellen die Grundlage für einen vielfältigen Nutzungsprozess sicher. Die Abfrage soll einen ungefähren Eindruck davon geben, mit welchen anderen Vokabularen eine bestimmte Terminologie also abgleichbar ist und welche Transformationsmöglichkeiten von Daten sich damit ergeben. Dies sind wesentliche Parameter der Qualitätsbestimmung und ihre Erfassung kann späteren Anwendenden die Entscheidung für oder gegen ein bestimmtes Vokabular erleichtern.

Zur Angabe wird der Name des verknüpften Vokabulars erfasst und einerseits mit einem Worturteil (“vollständig, teilweise, geringfügig”) sowie andererseit mit einer quantitativen Beurteilung des Matchingergebnisses (“geschätzte Vollständigkeit in %”) beschrieben. Dabei unterscheidet die Erfassung das Matching des eigenen Vokabulars mit dem jeweiligen Zielvokabular und umgekehrt. Gleichzeitig kann dieses Ergebnis für verschiedene Sprachen unterschiedlich ausfallen und wird daher für jede Sprache separat eingeschätzt (Frage 38).

Für spätere Matchings von besonderem Interesse ist die Frage, ob Vokabulare über stabile Identifikatoren verfügen.

Ebenfalls im Bereich Crosskonkordanz findet sich die Erhebung von aktuellen und potentiellen Nutzendengruppen, die später in Form von Vorschlagsoptionen in der Web-Präsentation genutzt werden können.

### 5.2.5 Bereitstellung der Daten

Einen wesentlichen Teil des Zugangs zu Vokabularen und Forschungsdaten stellen die technischen Standards dar, die im vierten Fragenblock detailliert erfasst werden. Für Transferprozesse sind Ein- und Ausgabeformate von Daten von entscheidender Bedeutung. Daher sollten nach Möglichkeit alle Ausgabeformat eines Vokabulars angegeben werden. Diese erfolgen jeweils durch die Angabe des meist dreistelligen Dateityps bzw. der Dateiendung. Zur Auswahl stehen die gängigen Datenformate (bspw. csv. xlsx, pdf, skos, rdf, json, xml), die beliebig ergänzt werden können. In diesem Zusammenhang interessiert ebenfalls die Software, mit der das Vokabular bereitgestellt wird (Frage 42). Direkt erhoben werden, wenn vorhanden, API-Schnittstelle (Frage 44) oder SPARQL-Endpunkte (Frage 45) von Datenangeboten.

Abschließend besteht in diesem übersichtlichen Fragenkomplex die Option, bereits bestehende Services zum Vokabular (Beratung, Matching etc.) anzugeben (Frage 46).

### 5.2.6 Ausblick und optionale Angaben

Im letzten Abschnitt wird in einem Freifeld der Bearbeitungsstand des Vokabulars erfasst, hier können beispielsweise erscheinende Versionen, Übersetzungen, Aktualisierungen etc. angekündigt werden (Frage 47). Das Freifeld wurde hier gewählt, um den vielen verschiedenen Bearbeitungsständen und eventuellen Anmerkungen der Erfassenden Raum zu geben sowie weitere Informationen zu hinterlassen. Intendiert ist hier die Absicht, bei Häufungen bestimmter Ergänzungen/Informationen eine Modifikation des Fragebogens vornehmen zu können; gleichzeitig werden so individuelle Informationen möglich (Frage 48).

## 5.3 Einpflegen in FactGrid

Die über das Umfragetool LimeSurvey generierten Daten lassen sich dank der Standardisierung, die die Eingabeschablone mit sich bringt, leicht in eine eindeutige Wikibase-Listeneingabe überführen. Die vorliegende Dokumentation nutzte bereits die Wikibase Property Nummern der FactGrid Instanz als Identifier der einzelnen Eingabefelder. Die folgende Abfrage listet diese Properties mit den auf ihnen liegenden Nutzungshinweisen und den einzelnen in Selektionen vorgegeben Antwortoptionen: [https://tinyurl.com/24fcfwce].

Ziel ist es, die FactGrid-Instanz sowohl als Repertorium für Metadaten zu Vokabularen wie als Hintergrundressource für das Interface zu nutzen, mit dem Nutzendesich im Bereich der historischen Vokabulare orientieren können. Die Wikibase-Datenbank wird dabei zusätzlich Feldinformationen auf spezifizierte Anfragen hin in das in Entwicklung befindliche Web-Interface einspielen, das auf einer URL des Historischen Datenzentrums

Halle laufen wird. Datenbankabfragen, die Nutzende auf der Suche nach Vokabularen durchführen, werden dabei Datenpakete erzeugen, die sich als csv, tsv oder json-Dateien unter der freien CC0 Lizenz herunterladen lassen.

Der Vorteil der Wikibase-Nutzung liegt darin, dass die Instanz sowohl schlichte Metadaten – extern vorliegender – Vokabulare verwalten kann, als auch ganze Vokabulare mitsamt ihren Kategorien im Objektgefüge der Plattform integriert. Nutzende können sich im FactGrid gehostete Vokabulare leicht herunterladen, diese zum Reconciling mit OpenRefine nutzen und dieselben Vokabulare auf der Plattform auch einfach zur Erschließung ihrer Forschungsdaten und Linked Open Data nutzen.

SPARQL-Abfragen sowohl von Vokabularen wie von Metadaten zu Vokabularen lassen sich im selben Moment problemlos in die BARTOC-Datenbank einspielen, so unsere Erfahrung mit den ersten Vokabularen, die wir in FactGrid vorhalten.

Die Abfrage der einzelnen Properties unseres Fragebogens ist vergleichsweise einfach möglich. Interessant werden Anfragen, die bis auf die Wortebene der Vokabulare gehen und die Worte Vokabularen zuordnen. Die Graph-Datenbank macht es hier möglich, dass beliebige Worte jederzeit in mehreren Vokabularen notiert sein können. Verschiedene Kategorisierungen und Hierarchisierungen können auf denselben Worten liegen und diese ganz kategorisieren.

Die Wikibase-Instanz speichert im Gegensatz zu herkömmlichen Repositorien keine Datenpakete mehr. Sie generiert vielmehr beliebige Pakete im Moment einer jeweiligen Abfrage aus der Datenlage. Das hat den Vorteil, dass nicht länger Gesamtüberarbeitungen von Datenpaketen im Rahmen laufender Versionierungskampagnen anstehen. Die Daten können punktuell in Gebrauchskontexten angefasst werden. Versioniert werden dabei die einzelnen Bearbeitungen auf der Ebene der jeweiligen Vokabel mit Angaben zu den Bearbeitenden und ihren Überarbeitungen. Das Ziel ist hier die Arbeit mit einer laufend punktuell funktionsfähig gehaltenen, dynamischen und tagesaktuellen Datenlage. Von besonderem Interesse ist es darum, Projekte, die Metadaten von Vokabularen oder ganze Vokabulare einspeisen, selbst zur Kuratierung “ihrer” Datenlage zu befähigen und in diesem Verfahren Kommunikationsverluste zu minimieren. Unsere Redaktion kann sich so auf Beratung, die Optimierung von Prozessen und die Qualitätssicherung von Arbeitsschritten in der Wikbase-Instanz konzentrieren. Die Erfahrungen mit der namentlich authentifizierten Nutzerschaft sind hier sehr positiv. Die Bearbeitungen fremder Daten werden, wo sie punktuell vorgenommen werden, eigentlich immer mit einer Rückbindung an die Datenlage vorgenommen, die Großprojekte flächendeckend kaum leisten können.

Das Zusammenspiel von Wikibase-Instanz mit Eingabe- und Ausgabe-Tools wird innovativ sein: Graphdatenbanken werden bei der Etablierung von Standards sehr schnell unhandlich – es gibt in ihnen keine Eingabeschablonen, sie schaffen Netze von Verbindungen, die nachher kohärent abgefragt werden müssen. Hier soll die Eingabe über LimeSurvey und die Datenbankabfrage über ein eigenes Interface Standardisierung von beiden Seiten, der Datengenerierung und Datenausgabe, schaffen und sowohl die

Bedürfnisse von eher technikaffinen wie inhaltsaffinen Bearbeitenden optimal unterstützen.

## 6. Ausblick / Zeitplan

Am 23.04.2024 wurde die Umfrage freigeschaltet und allen Participants zur Verfügung gestellt. Um Participants und Interessierte in die Erfassung einzuführen und möglichen Unsicherheiten beim Ausfüllen vorzubeugen, finden bei Bedarf einer größeren Gruppe Register-Workshops statt, wie dies bei der Einführung des Tools erfolgt. In Form eines Briefings zeigen wir den Participants und allen Interessierten, wie die Erfassung mit Limesurvey funktioniert. Wir werden unsere Dokumentation vorstellen und anhand von Anwendungsbeispielen Musterlösungen bereitstellen. Die Vorstellung soll aufgezeichnet und über einen Youtube-Kanal von HistData [https://www.youtube.com/channel/UCit6Az_C32PIRfUy1R812XA] zur Verfügung gestellt werden.

Einlaufende Daten sollen quartalsweise in FactGrid übernommen und im selben Rahmen an BARTOC weitergemeldet werden. Vereinbarungen zu Schulungen für FactGrid können über das Team NFDI4Memory HistData getroffen werden.

Die Erfassung über LimeSurvey bleibt nach Möglichkeit für den Förderzeitraum bis 2028 aktiv und sichert ein fortlaufendes Einpflegen neuer Daten. Um die Einträge aktuell zu halten, werden die innerhalb der TA2 registrierten Bereitstellenden von Daten regelmäßig daran erinnert, Veränderungen im Register einzupflegen. Alle Änderungen von Einträgen des Registers erfolgen anschließend in FactGrid. Durch die Erfahrungen beim Einpflegen von Nutzer*innen, aber auch aus der Sicht der Projektverantwortlichen, werden etwaige Besonderheiten evaluiert. So wird der Fragebogen weiter optimiert.

Beschreibungen und Dokumentation werden über eine projektspezifische Website präsentiert, über die später auch die Ausgabemaske zu vorgefertigten Abfragen aus FactGrid bereitgestellt werden soll.

## 7. Glossar zentraler Begrifflichkeiten

Genannte Begrifflichkeiten werden je nach Kontext/Fachdisziplin sehr unterschiedlich genutzt, darum sollen kurze Definitionen ein grundlegendes Verständnis ermöglichen, was genau hier mit bestimmten Terminologien angesprochen wird. Es wurden sehr knappe und nur wesentliche Begriffsdefinitionen entwickelt.

**Kontrollierte Vokabulare:** Können eine Sammlung von Wörtern, Phrasen, Kategorien, Notationen oder anderen Bezeichnungen (Sprachvarianten, Termen, terms) sein, die aufgrund von Begriffen (concepts, records) zur Definition, zur Erschließung und zum Auffinden von Informationen dienen.[11] Begriffe werden durch Konzepte und

[11] Leicht modifiziert nach: Jessica Sandrock: Kontrollierte Vokabulare, in: DigiCult.xTree (Hg.): Thesaurus-Handbuch, 2023, URL: https://digicult.atlassian.net/wiki/spaces/XTREE/pages/3287056514/1+Kontrollierte+Vokabulare+In+Bearbeitung.

Beschreibungen eindeutig definiert. Ihnen werden die Bezeichnungen (Sprachvarianten) zugeordnet.

Der Begriff des „kontrollierten Vokabulars“ umfasst daher ganz unterschiedliche Formen, angefangen von einer einfachen Wortliste bis hin zu hierarchisch organisierten, komplexen Klassifikationen und Thesauri. Je nach Verwendungszweck sind die Ausformungen von Regeln, die Abbildung der Definitionen, die Organisationsformen der Erstellung und Erweiterung von Begriffen sowie die Dokumentation und Zugänglichkeit von Vokabularen sehr unterschiedlichen Anforderungen unterworfen. Diese werden hier als Qualitätskriterien verstanden. In der Regel wird ein Vokabular in einer einheitlichen Sprache angeboten, kann aber über Übersetzungsleistungen auch in zahlreiche weitere Sprachen überführt werden. Dabei ist die genaue Passfähigkeit der Definition bzw. Konzepten von Begriffen und Bezeichnungen zu gewährleisten. Gleiches gilt für die Disambiguierung in Raum und Zeit. Dies fördert, wie auch das Matching mit ähnlichen oder verwandten Vokabularen, die Internationalisierung eines Angebots. So entstehen Crosskonkordanzen, die eine Übertragung von einem System in ein anderes ermöglichen.

Oft werden kontrollierte Vokabulare für die Erfassung einer bestimmten **Entität** oder zur Erschließung eines bestimmten Sammelgebietes eingegrenzt. So gibt es Vokabulare, die sich nur der Erfassung von Orten, Berufen oder Personen (Entitäten) widmen oder eben einezlne oder mehrere Wissenschaftsgebiete thematisch oder zeitlich ordnen möchten (Sammlungsgebiete, Fachgebiete).

Bezeichnungen, Begriffe und Definitionen werden in der Regel mit persistenten Identifikatoren versehen und können zusätzlich taxonomische Aufgaben erfüllen. Für den Aufbau eines vernetzten, webbasierten Datenraumes werden zunehmend webbasierte, persistente Identifikatoren (persistente Links) notwendig. Über Linked Open Data entstehen so verknüpfte Wissensräume. Insgesamt lassen sich unterschiedliche Schwerpunktsetzungen mit verschiedenen Unterarten kontrollierter Vokabulare finden:

**Wortlisten und Kategoriensysteme**: In vielen Projekten entstehen Listen von Bezeichnungen, die bestimmten Definitionen, Konzepten oder Beschreibungen zugeordnet werden, um sprachliche Ausdrücke zu systematisieren. In der Regel werden solche Definitionen und ihre Zuordnungen in Dokumentationen festgehalten. Häufig fungiert die Kategorie gleichzeitig als Identifikator. Diese Kategoriesysteme gelten als kontrollierte Vokabulare, weil sie eine überindividuelle Zuordnung und Nachvollziehbarkeit von Analyseprozessen ermöglichen.

**Thesaurus / Thesauri**: dienen der regelbasierten Normierung natürlicher Sprache und semantischer Formen/Inhalte, indem sie die Relationen/Verknüpfungen von Wörtern mit ähnlichen oder gegensätzlichen Bedeutungen klären und zuordnen. Ein Thesaurus muss in der Regel nicht im Sinne einer Klassifikation strukturiert sein, sondern besitzt weitgehend assoziative, nichthierarchische Relationen. Thesauri dienen dazu:

- Bezeichnungen (Numerus, Genus, Wortreihenfolge) nach vorgegebenen Regeln zu vereinheitlichen

- Mehrdeutigkeiten von Bezeichnungen aufzulösen (Disambiguierung) und über feste Mechanismen der terminologischen Kontrolle bestehende Homonyme zuzuordnen
- Synonyme und bedeutungsähnliche Bezeichnungen zu identifizieren, Begriffen zuzuordnen und über Vorzugsbezeichnungen (Deskriptoren/Schlagworten) auffindbar zu machen.[12]

**Klassifikationen/Taxonomien**: werden mithilfe einer Methode zur Organisation von Informationen in hierarchischen oder taxonomischen Strukturen gebracht, wobei Elemente in Gruppen oder Klassen eingeteilt werden, die auf gemeinsamen Merkmalen und Eigenschaften basieren. Sie sind daher auf eine Entität bezogen und messen in der Regel bestimmte Merkmale oder Eigenschaften nach vorgegebenen, regelhaften Kriterien. Dabei werden Anforderungen der Thesauri zugrunde gelegt. Hierzu muss eine Klassifikation keine semantische Zuordnung im Sinne von Vorzugsbezeichnungen oder Relationsbeziehungen auf gleicher Hierarchieebene enthalten. Besonders Taxonomien ordnen Objekte nach bestimmten Rangstufen oder auch metrischen Systemen. Klassifikationen dienen dazu:

- Begriffe einer Entität eindeutig hierarchisch nach einem bestimmten Merkmal zu messen
- Ordnungskriterien und Hierarchien zu entwickeln und
- Analysesysteme für Bezeichnungen und Begriffe für wissenschaftliche Anwendungen und Ordnungssysteme auszugeben·

**Normdaten (Authority Data)**: Normdaten sind öffentlich verfügbare, standardisierte kontrollierte Vokabulare, die für die Repräsentation, Beschreibung und Identifikation von bestimmten Entitäten durch eine größere Community gemeinsam genutzt werden und über gemeinsam gepflegte Regeln, persistente Identifikatoren und strukturelle stabile Organisationen betrieben werden. Aufgrund der Bedeutung solcher Normdaten für die Community sind solche Vokabulare gut dokumentiert und beschrieben, besitzen Möglichkeiten für den Download in verschiedenen Dateiformaten und werden durch weitere Serviceleistungen ergänzt.

**Ontologie/Schemata**: Eine Ontologie ist ein informatisches Konzept, das dazu dient, die Ordnungen von Vokabularen und ihrer Konzepte in ein übergeordnetes Sinnsystem, technisches Schema oder formale Modelle zu überführen, um sie über webbasierte Protokolle zu verknüpfen. Dabei beschreiben sie vor allem die Beziehungen zwischen den unterschiedlichen Begriffen und ihren Konzepten bzw. hierarchische Ordnungsprinzipien.

[12] Jessica Sandrock: Was ist ein Thesaurus?, in: DigiCult.xTree (Hg.): Thesaurus-Handbuch, 2023, URL: 2.1 Was ist ein Thesaurus? (In Bearbeitung) - digiCULT.xTree - Confluence (atlassian.net).