Knowledge Graph-basierte Forschungsdaten-integration in NFDI4Culture

Tietz, Tabea
tabea.tietz@fiz-karlsruhe.de
FIZ Karlsruhe – Leibniz Institut für Informationsinfrastruktur, Deutschland; Karlsruher Institut für Technologie, Institut AIFB, Deutschland

Bruns, Oleksandra
oleksandra.bruns@fiz-karlsruhe.de
FIZ Karlsruhe – Leibniz Institut für Informationsinfrastruktur, Deutschland; Karlsruher Institut für Technologie, Institut AIFB, Deutschland

Fliegl, Heike
heike.fliegl@fiz-karlsruhe.de
FIZ Karlsruhe – Leibniz Institut für Informationsinfrastruktur, Deutschland

Posthumus, Etienne
eposthumus@gmail.com
FIZ Karlsruhe – Leibniz Institut für Informationsinfrastruktur, Deutschland

Schrade, Torsten
Torsten.Schrade@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz, Deutschland

Sack, Harald
harald.sack@fiz-karlsruhe.de
FIZ Karlsruhe – Leibniz Institut für Informationsinfrastruktur, Deutschland; Karlsruher Institut für Technologie, Institut AIFB, Deutschland

Einleitung

Die Nationale Forschungsdateninfrastruktur (NFDI)[1] wurde im Jahr 2020 mit dem Ziel ins Leben gerufen, die Datenbestände aus Wissenschaft und Forschung für das deutsche Wissenschaftssystem systematisch zu erschließen, zu vernetzen und nutzbar zu machen. NFDI4Culture startete im Oktober 2020 als Konsortium der ersten Förderrunde und widmet sich Forschungsdaten zu materiellen und immateriellen Kulturgütern (Altenhöner et al. 2020). Die Interessengemeinschaft von NFDI4Culture reicht von Architektur-, Kunst- und Musik- bis hin zu Theater-, Tanz-, Film- und Medienwissenschaft und besteht aus über 70 beteiligten Organisationen[2]. Dieses multidisziplinäre Konsortium produziert und verarbeitet eine große Menge heterogener Daten und unterhält Repositorien und Services, die für das deutsche Wissenschaftssystem und auch darüber hinaus in Kunst und Kultur von großer Bedeutung sind. Zum Datenspektrum von NFDI4Culture gehören 2D-Digitalisate von Gemälden, Fotografien und Zeichnungen ebenso wie digitale 3D-Modelle kulturhistorisch bedeutender Gebäude, Denkmäler oder audio-visuelle Daten von Musik-, Film und Bühnenaufführungen (Bicher et al., 2022). In den Datendomänen des Konsortiums werden bisher jedoch nur vereinzelt einheitliche offene Standards und Datenmodelle genutzt. Forschungsdaten liegen oftmals in sogenannten Datensilos vor, die weder von außen gefunden noch wiederverwendet werden können, da auch die Zugangsmöglichkeiten uneinheitlich und kompliziert sind. Zudem sind die Lizenzen zur Nutzung der Ressourcen oft nicht sofort ersichtlich oder vollständig geklärt, was deren Nachnutzung zusätzlich erschwert. Weiterhin stellen insbesondere auch die mit digitalen Kulturgütern zusammenhängenden, häufig komplexen datenrechtlichen und datenethischen Aspekte eine Herausforderung dar. Ziel von NFDI4Culture ist es daher, eine bedarfsgerechte Infrastruktur für die Forschungsdaten der Interessengemeinschaft zu schaffen, die den F.A.I.R. Prinzipien folgt und somit das Auffinden, den Zugang, die Nutzbarkeit und Interoperabilität der Ressourcen für alle sicherstellt (Wilkinson et al., 2016).

Die Implementierung der F.A.I.R. Prinzipien erfolgt im Konsortium über die Bereitstellung und Nutzung von domänenspezifischen Ontologien, die Entwicklung von Knowledge Graphs und die Verknüpfung einer Vielzahl von strukturierten Datenbanken und Knowledge Graphs untereinander. Ein einheitlicher und intuitiver Zugriff auf die dezentral vorliegenden Forschungsdatenressourcen des Konsortiums wird über eine zentrale Plattform, das “Culture Information Portal” [3] sichergestellt.

In diesem Beitrag wird der aktuelle Stand und die weiteren Planungen der technisch übergreifenden Task Area 5 von NFDI4Culture zur Knowledge Graph-basierten Integration von Forschungsdaten materieller und immaterieller Kulturgütern vorgestellt. Dies beinhaltet eine Diskussion aktueller technischer und domänenspezifischer Herausforderungen, die Vorstellung der NFDICO Ontologie und des NFDI4Culture Knowledge Graphen sowie eine Darstellung zur Implementation des Culture Information Portals.

Herausforderungen der Datenintegration in NFDI4Culture

Die von NFDI4Culture in den Blick genommene Forschungslandschaft ist durch eine starke Diversität gekennzeichnet. Sie umfasst nicht nur eine Vielzahl von Forschungsdisziplinen, sondern auch unterschiedlichste Organisationen, darunter Universitätsinstitute, Kunst- und Musikhochschulen, Akademien, Galerien, Bibliotheken, Archive, Museen und einzelne Forscher*innen.

Dementsprechend sind die Forschungsprozesse- und ressourcen, die auffindbar, interoperabel und wiederverwendbar gemacht werden müssen, ebenfalls heterogen und liegen nicht nur in divergierenden Standards und Formaten vor, sondern auch in unterschiedlich aufbereiteten Zuständen. Kollektionen sind nicht immer vollständig digitalisiert und erschlossen, weshalb oft nur wenige beschreibende Metadaten zur Verfügung stehen. Andere Datensätze liegen bereits vollumfänglich als Linked Open Data vor und können problemlos mit dem NFDI4Culture Knowledge Graph verknüpft werden. Die Implementierung der F.A.I.R. Prinzipien ist ein Hauptziel des Konsortiums. Daher werden dedizierte Ontologien zur Verfügung gestellt und Maßnahmen getroffen, um alle Akteur*innen dabei zu unterstützen, eigene Daten und Ressourcen langfristig und nachhaltig selbst in Linked Open Data zu transformieren.

Ein wissenschaftsgeleitetes Forschungsdatenmanagement erfordert die aktive Teilnahme aller. Das Konsortium sieht daher umfangreiche Beteiligungsmöglichkeiten für die Nutzenden aller involvierten Fachdisziplinen, aber auch für Kunst- und Kulturschaffende unterschiedlichster Tätigkeitsbereiche und Vertreter*innen der Zivilgesellschaft vor. Es zielt darauf ab, das breite Spektrum der verschiedenen Akteur*innen im Bereich des Kulturerbes zu repräsentieren. Unter anderem ist vorgesehen, dass die Fachgemeinschaft selbst den Knowledge Graph mit eigenen Ressourcen erweitert und dessen Inhalte pflegt. Bedingung dafür ist eine technische Infrastruktur, die es ermöglicht, Ressourcen intuitiv und ohne tiefe technische Kenntnisse zu kuratieren, hinzuzufügen, zu verknüpfen und zu durchsuchen. Diese Infrastruktur wird eine semantische expressive Repräsentation der Daten ermöglichen, um eine vollumfängliche Umsetzung der F.A.I.R Prinzipien zu gewährleisten.

NFDICO und der NFDI4Culture Knowledge Graph

Ein Ziel von Task Area 5 in NFDI4Culture ist die Bereitstellung von Ontologien, um die Forschungsdaten in NFDI4Culture standardisiert und formal zu repräsentieren und miteinander verknüpfen zu können. In einem “bottom-up” Ansatz nach der sogenannten “Waterfalls” Methode (Keet, 2020) und im kontinuierlichen Austausch mit den Domänenexpert*innen des Konsortiums wurde dazu die NFDICO Ontologie entwickelt. Sie verknüpft Datensätze, Forschungsprojekte, Services, Repositorien, Institutionen und Forschungsdisziplinen und dient als Grundlage für den NFDI4Culture Knowledge Graph. Als Modellierungsgrundlage dienten die durch die NFDI4Culture Community übermittelten Beschreibungen ihrer Forschungsressourcen. Die Klasse nfdico:Contribution[4] (Abb. 1) repräsentiert die Beiträge der Fachgemeinschaften und ordnet die Arten der Beiträge (z.B. Datenportal, Datensatz, Kollektion, Software, Infrastruktur, Service) unter.

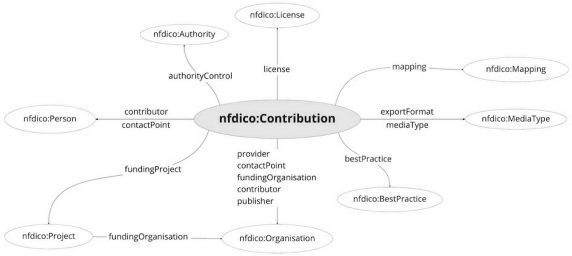

Abbildung 1: Modellierung der Klasse nfdico:Contribution

Instanzen können beispielsweise durch Medientypen, Lizenzangaben, zugehörige Personen und Institutionen und Projekte beschrieben werden. Ein Modellierungsbeispiel für die Klasse nfdico:Service[5] als Unterklasse von nfdico:Contribution ist in Abb. 2 dargestellt. Die Klasse wird beispielsweise durch die akademische Disziplin, Medientypen, verwendete Normen und Ontologien spezifiziert.

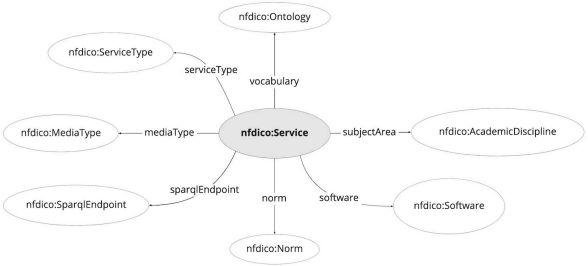

Abbildung 2: Modellierung der Klasse nfdico:Service

NFDICO besteht in der Version 1.1 aus 36 Klassen und 60 Objektattributen. Die spezifisch für den Anwendungsfall in NFDI4Culture definierten Klassen (Prefix nfdico) wurden den “best-practices” in der Ontologieentwicklung folgend zur Sicherstellung hoher semantischer Expressivität und Interoperabilität mit 24 bereits existierenden Ontologien verknüpft, darunter frapo[6], fabio[7], void[8] und schema[9] (Keet, 2020). Die Ontologie ist seit Juni 2022 öffentlich verfügbar und vollständig dokumentiert[10]. NFDICO folgt einem generischen Modellierungsansatz und repräsentiert die Beiträge der Fachgemeinschaft nicht ausschließlich für NFDI4Culture Domänen, sondern ermöglicht die Nachnutzung durch weitere NFDI Konsortien anderer Fachrichtungen. So dient NFDICO beispielsweise als Grundlage der Basisontologie des NFDI Konsortiums NFDI-MatWerk[11]. Gemäß der Waterfalls Methodologie der Ontologieentwicklung wird NFDICO in Absprache mit Domänenexpert*innen iterativ und modular erweitert.

Wie bereits o.g., dient NFDICO als Grundlage für den NFDI4Culture Knowledge Graph. Eine erste Version des Graphen ist publiziert und kann über einen öffentlichen SPARQL Endpunkt[12] abgefragt werden. Der inhaltliche Aufbau des Knowledge Graphen erfolgte auf Basis der Forschungsressourcen und beschreibenden Metadaten, die von der NFDI4Culture Community übermittelt wurden. Alle Beiträge wurden normalisiert und via RDFLib[13]

in den Knowledge Graph integriert. Außerdem wurden (soweit möglich) Verknüpfungen der Entitäten aus dem Knowledge Graph zu Wikidata[14] und der Gemeinsamen Normdatei (GND)[15] hergestellt. Das heißt, die Daten im NFDI4Culture Knowledge Graph sind von Anfang an anschlussfähig und können mittels föderierter SPARQL Abfragen erweitert werden. So können für alle in NFDI4Culture beteiligten Organisationen über die jeweiligen Wikidata-Verknüpfungen zusätzliche Informationen, wie zum Beispiel der Typ der Organisation (z.B. Bibliotheken[16]) in die Ergebnisanzeige mit einbezogen werden, obwohl diese Informationen im NFDI4Culture Knowledge Graph nicht explizit enthalten sind.

Der NFDI4Culture Knowledge Graph enthält aktuell 1796 Entitäten, darunter 173 Organisationen, 156 Forschungsressourcen, die mit ca. 10.000 RDF-Tripeln beschrieben werden. Der Knowledge Graph wird kontinuierlich erweitert und schrittweise für Beiträge durch die Community geöffnet, sodass alle Akteur*innen in Zukunft selbst eigene Ressourcen verknüpfen und öffentlich zugänglich machen können. Die Interaktion der Community mit dem Knowledge Graph erfolgt über eine einheitliche und intuitive Schnittstelle, das im Folgenden beschriebene Culture Information Portal.

Culture Information Portal und Integration des Knowledge Graph

Mit dem Culture Information Portal verfolgt NFDI4Culture das Ziel, einen einheitlichen, intuitiven und zentralen Einstiegspunkt auf die dezentral gespeicherten Forschungsdaten der Community und alle weiteren Dienste des Konsortiums (wie z.B. ein übergreifendes Helpdesk, nachnutzbare Guidelines zu allen Bereichen des Forschungsdatenmanagements oder das konsortiumsweite Identitäts- und Zugriffsmanagement für die gemeinsamen Kommunikations- und Kollaborationswerkzeuge) bereitzustellen. Forschungsdaten sollen durch die Community selbst hinzugefügt, kuratiert und auffindbar gemacht werden. Weiterhin informiert das Culture Information Portal über aktuelle Veranstaltungen und Neuigkeiten des Konsortiums, sowie über beteiligte Akteure und Projektfortschritte. Alle Informationen im Portal sollen außerdem im Sinne der F.A.I.R. Prinzipien als Linked Open Data vorliegen und eine gute Anschlussfähigkeit an europäische Informationsinfrastrukturen wie die European Open Science Cloud gewährleistet sein. Das Portal wurde aus diesem Grund als webbasiertes Forschungsinformationssystem (Current Research Information System / CRIS) auf Basis des Open Source Content Management Systems TYPO3[17] realisiert. Das Datenmodell des Portals orientiert sich am CERIF-Standard der euroCRIS[18], wodurch gleichzeitig eine sehr gute Kompatibilität zu NFDICO gegeben ist. Für die CRIS Implementierung wurde die TYPO3-Extension “Academy” nachgenutzt und weiterentwickelt[19].

Zur Integration des Culture Knowledge Graphen wurden mit dem Ziel einer föderierten Informationsinfrastruktur zunächst existierende Systeme auf die oben genannten Kriterien überprüft, darunter Wikibase[20] und WissKI[21]. Wikibase ist eine freie Software zur kollaborativen Kuratierung von Datenbanken mit der Möglichkeit, Daten im Sinne von LOD zu strukturieren, zu verknüpfen und abzufragen. Wikibase bietet allerdings nicht die Möglichkeit, externe, bereits existierende Ontologien zu integrieren, was die semantische Expressivität der Daten stark begrenzt. Außerdem ist es mit Wikibase nicht möglich, eigene Eingabemasken zu entwickeln, was die intuitive Interaktion mit dem Knowledge Graph einschränkt. WissKI ist eine webbasierte Software zum Sammeln, Strukturieren und Teilen von forschungsbezogenen Daten. Feste Grundlage von WissKI bildet dabei die Optimierung der Software auf CIDOC-CRM, was sich für den hier beschriebenen Anwendungsfall als zu einschränkend gezeigt hat. Die Modellierung in CIDOC-CRM folgt einem Ereignis-basierten Paradigma. Dadurch gerät die Modellierung einfacher Fakten oft sehr komplex und behindert den typischen Anwendungsfall in NFDI4Culture. Durch CIDOC werden einfache SPARQL-Abfragen zu Personen und Organisationen oft hochkomplex und daher ineffizient. Aufgrund der großen Menge an Daten, die in NFDI4Culture erwartet werden, ist eine Abfrage-Effizienz allerdings relevant. Der Zugang zu NFDI4Culture Daten soll für alle Nutzer*innen so einfach wie möglich gestaltet werden, eine CIDOC-CRM-basierte Modellierung schafft aufgrund ihrer Komplexität zusätzliche Barrieren. Dennoch ist CIDOC ein wichtiger und gängiger Bestandteil der GLAM Community. Daher wird ein Mapping des NFDI4Culture Datenmodells nach CIDOC durchgeführt, um einen CIDOC-basierten Export zu gewährleisten.

Als beste Lösung zur Gewährleistung der semantischen Expressivität der erfassten Ressourcen und Metadaten mittels NFDICO und weiterer Ontologien einerseits bei gleichzeitiger Umsetzbarkeit der benötigten Kurationsmechanismen andererseits stellte sich die direkte Implementierung der benötigten Funktionalitäten in TYPO3 heraus. Die TYPO3 Extension “LOD”[22] bietet hierbei einen unmittelbar über der relationalen Datenbank des CMS realisierten “Semantic Layer” mit einem konfigurierbaren IRI-Generator sowie IRI-Resolver für alle Datensätze sowie einem RDF-Serializer für alle Datenbankinhalte. Alle im Portal bzw. Culture CRIS erfassten Ressourcen werden dabei über eine standardisierte LOD API unter Verwendung des Hydra Core Vocabulary[23] in verschiedenen RDF Serialisierungen (z.B. RDFa, Turtle, JSON-LD u.a.) veröffentlicht[24]. Hierdurch können die Daten für den Culture Knowledge Graph bereits jetzt im Kreis der Mitarbeitenden des Konsortiums dezentral kuratiert und kontinuierlich erweitert werden. Die über die LOD API des CMS publizierten Daten werden mittels Ingest-Routinen in den eigentlichen Knowledge Graph integriert. Zum Einsatz kommt oxigraph[25] als nativer RDF-Store mit einem pythonbasierten Wrapper[26] für ein leichtgewichtiges Deployment des öffentlichen SPARQL-Endpoints, der über ein grafisches Interface wiederum direkt in das Culture Information Portal integriert ist.

Weiteres Vorgehen und Zusammenfassung

NFDI4Culture öffnet Forschungsdatensilos und schafft einheitliche und intuitive Zugriffsmöglichkeit auf Forschungsdaten. Dieses Vorhaben wird in der Task Area 5 durch die Bereitstellung dedizierter Ontologien, die Implementation und Verknüpfung von Knowledge Graphen und die Umsetzung des Culture Information Portals erreicht. Akteur*innen der Community können mit ihrer Beteiligung am Vorhaben ihre Daten also auffindbar, interoperabel und wiederverwendbar machen, die Zitierfähigkeit eigener Ressourcen gewährleisten und die multidisziplinäre Verknüpfungen der Daten für ihre Forschung ausnutzen. Alle hier präsentierten Beiträge sind öffentlich verfügbar und nutzbar. Im weiteren Vorgehen wird die Ontologie bedarfsorientiert erweitert und das Culture Information Portal schrittweise für die Community geöffnet werden, um mit den Inhalten im Knowledge Graph zu interagieren und den Knowledge Graph mit Forschungsdaten anzureichern und diese zu kuratieren. Die technische Infrastruktur wird stetig verbessert, um unter anderem teil-automatisierte Qualitätskontrollen der Daten zu umzusetzen (z.B. durch SHACL[27]) und die Integration größerer Datenmengen über den SPARQL Endpunkt zu ermöglichen.

Vergleichbare und verwandte Initiativen werden mit der "Linked Data Platform Finland"[28] und dem "Dutch Digital Heritage Network"[29] umgesetzt. Auch NFDI4Culture ist eine nationale Initiative, gleichsam ist allen Beteiligten die Bedeutung einer globalen Vernetzung bewusst, denn auch deutsche Kulturdaten und deutsches kulturelles Erbe sind international und von internationalem Interesse. In NFDI4Culture wurden bereits Maßnahmen implementiert, um auch Organisationen und Forschende außerhalb des Konsortiums zu involvieren. Beispielsweise dienen dazu das Steering Board[30] und die "Linked Open Data Working Group". Technisch ist eine Vernetzung mit anderen Plattformen und die Nutzung international anerkannter Standards ebenso elementar. Der NFDI4Culture Knowledge Graph enthält bereits Verknüpfungen zu Wikidata, die mit Hilfe der Nutzer*innen stetig erweitert werden. Zudem besteht auf der NFDI4Culture Ontologie-Ebene bereits ein Mapping mit dem "Common European Research Information Format (CERIF)". Diese Verknüpfungen werden in Zukunft weiter ausgebaut werden.

Unter Umsetzung der F.A.I.R. Prinzipien und mit der Beteiligung der gesamten Community wird in NFDI4Culture eine Forschungsinfrastruktur geschaffen, die langfristig und nachhaltig Forschungsdaten auffindbar, zugreifbar und nutzbar für alle macht.

Fußnoten

1. https://www.nfdi.de/
2. https://nfdi4culture.de/
3. https://nfdi4culture.de/
4. https://nfdi4culture.de/ontology#Contribution
5. https://nfdi4culture.de/ontology#Service
6. http://www.sparontologies.net/ontologies/frapo
7. http://www.sparontologies.net/ontologies/fabio
8. https://www.w3.org/TR/void/
9. https://schema.org/
10. https://nfdi4culture.de/ontology
11. https://nfdi-matwerk.de/"
12. https://nfdi4culture.de/de/resources/knowledge-graph.html
13. https://github.com/RDFLib/rdflib
14. https://www.wikidata.org/
15. https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd.html
16. SPARQL Query
17. https://typo3.org/cms
18. https://eurocris.org/. CERIF ist ein von der Europäischen Union empfohlenes, standardisiertes Datenmodell, vgl. https://openaire-guidelines-for-cris-managers.readthedocs.io/en/v1.1.1/introduction.html
19. https://github.com/digicademy/academy
20. https://wikiba.se/
21. https://wiss-ki.eu/
22. https://github.com/digicademy/lod
23. https://www.hydra-cg.com/spec/latest/core/
24. https://nfdi4culture.de/resource/
25. https://github.com/oxigraph/oxigraph
26. https://github.com/epoz/shmarql
27. https://www.w3.org/TR/shacl/
28. https://www.ldf.fi/
29. https://netwerkdigitaalerfgoed.nl/en/
30. https://nfdi4culture.de/about-us/consortium.html#c256

Bibliographie

Altenhöner Reinhard et al. 2020. "NFDI4Culture - Consortium for research data on material and immaterial cultural heritage." Research Ideas and Outcomes 6: e57036. https://doi.org/10.3897/rio.6.e57036

Bicher, Katrin et al. 2022. "Digitalisierung des Kulturellen und digitale Arbeitskultur im Forschungsverbund NFDI4Culture. Community-Arbeit an, durch und mit fachspezifischen Datenkorpora und Elementen der FDM-Infrastruktur". Zeitschrift für Bibliothekswesen und Bibliographie, Jahrgang 69, Heft 1-2, S. 26-36. http://dx.doi.org/10.3196/1864295020691258

Keet, Maria. 2020. "Methodologies for Ontology Development." An Introduction to Ontology Engineering, Engineering Library.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship." Scientific data 3, no. 1: 1-9. https://doi.org/10.1038/sdata.2016.18