Communities, Harvesting, and CGIF: Building the Research Data Graph at NFDI4Culture

Jonatan Jalle Steller[1], Linnaea Charlotte Söhn[1], Julia Tolksdorf[1], Oleksandra Bruns[2,3], Tabea Tietz[2,3], Etienne Posthumus[2,4], Heike Fliegl[2], Sarah Pittroff[1], Harald Sack[2,3], Torsten Schrade[1]
[1]Academy of Sciences and Literature Mainz, Germany
[2]FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany
[3]Karlsruhe Institute of Technology (AIFB), Germany
[1]{firstname.lastname}@adwmainz.de
[2]{firstname.lastname}@fiz-karlsruhe.de
[3]{firstname.lastname}@kit.edu
[4]{firstname.lastname}@partners.fiz-karlsruhe.de

1. Problem: truly linked research data

As a consortium of the *Nationale Forschungsdateninfrastruktur* (NFDI), NFDI4Culture is tasked with developing solutions to systematically make accessible and interconnect the rich decentralised research data available from various providers across its domains. These include architectural studies, art history, musicology, performing arts, and media studies. The overarching goal is to make such data usable for further research in the long term.

Research data in the NFDI4Culture domains largely exists in silos. Even though a large number of data providers subscribe to the use of authority files and controlled vocabularies like the GND, VIAF, Wikidata, or Iconclass to structure their research data, the resources they publish are not automatically 'linked' to the full extent of 5-star Linked Open Data (LOD) (cf. Berners-Lee 2009). While many data providers support their users in getting from an individual resource to authority data, the reverse research path across individual repositories is largely obscured.

NFDI4Culture is building an information system that should enable users to find highly specific resources like, for example, images, objects, and 3D models depicting a specific motif based on authority files and controlled vocabularies. As such, it should also allow participating projects to retrieve related data from other participants in order to connect data based on information such as time, location, resource type, or motif and make them accessible for further research – even beyond the boundaries of individual research domains. The information system is further required to produce FAIR research data, i.e. data that is findable, accessible, interoperable, and reusable (cf. Wilkinson et al. 2016).

The paper is structured in the following manner. First, it reviews existing solutions for interconnecting research data (section 2). Then, an outline of the approach we chose to satisfy the above requirements is given (3). The next section discusses the implementation of the 'Research Data Graph' introduced in this paper (4). The final section outlines ongoing work to enhance and promote the presented solution across and beyond NFDI (5).

2. Review: centralised and federated infrastructures

Multiple approaches are possible to interconnect research data. Centralised infrastructures, for example, contain large amounts of data in a single location, and participating projects need to compile and contribute their data regularly for users to be able to find up-to-date content. Federated infrastructures, on the other hand, may have overarching interfaces but directly pass on requests to the participating data providers and need to collect and output their responses to queries.

Classical examples of centralised information systems in the culture domain are the German Digital Library (DDB) or Europeana(cf. Deutsche Digitale Bibliothek, n.d.; Europeana, n.d.). They ingest large amounts of data about physical objects via a network of aggregators such as museums and other libraries. In the case of Europeana, they ingest community standards such as LIDO and transform the data into the Europeana data model(cf. Isaac 2013, 4–6). While the object-focused data model is too restrictive for all the domains NFDI4Culture covers, Europeana's centralised approach enables them to semantically enhance data by applying a number of vocabularies to each record(cf. Isaac et al. 2015, 2).

Compared to centralised systems, federated infrastructures emphasise a shared API over shared formats. This requires more effort from individual data providers to implement a reliable endpoint, but has the benefit of providing information that is as up-to-date as a data provider is willing and able to deliver. In addition, fully federated systems can be less strict about the licence that data is made available under. The CLARIN Federated Content Search (FCS), for example, requires participants to implement an endpoint for the Search/Retrieve via URL (SRU) protocol and the Contextual Query Language (CQL) with responses serialised as XML(cf. CLARIN, n.d.), but does not require providers to specify a licence that governs how their data may be reused.[1] The technology is being reused by the NFDI consortium Text+ to interconnect linguistic data(cf. Körner et al. 2023), but does not naturally lend itself to NFDI4Culture due to CQL's limitation to text corpora. More recent approaches on this side of the spectrum require REST APIs, as in the case of the FCS developed in the ELEXIS lexicography project (cf. ELEXIS 2022), or SPARQL endpoints, which have federation built into the standard (cf. Prud'hommeaux and Buil-Aranda 2013).

Two existing projects stick out due to their hybrid approaches, which served as inspiration for NFDI4Culture. Firstly, Wikidata combines its centralised storage with participatory data management and the option to query its data via SPARQL (cf. Vrandečić and Krötzsch 2014).[2] Secondly, correspSearch allows participants to hand in correspondence metadata in a limited TEI XML format called CMIF in an effort to allow scholars to find correspondence data across corpora (cf. Dumont 2022).

3. Solution: the Research Data Graph

The solution implemented by NFDI4Culture aims to combine the authority and extensibility of a centralised repository with the diversity of federated APIs. The so-called Research Data Graph (RDG) organises a limited set of metadata on research data from participating providers into a knowledge graph. The goal is to provide data that is as granular as possible, but without demanding a specific level of detail:

[1] Europeana and Wikidata, on the other hand, only allow CC0-licensed content

[2] Some recent infrastructure initiatives, like the DraCor project as part of CLS INFRA, rely on Wikidata as a community data repository with a SPARQL endpoint (cf. Fischer et al. 2019, 4).

while the Corpus Vitrearum Germany, for example, provides metadata on individual images of stained-glass windows, a repository service like RADAR4Culture only has metadata on entire data sets which they store. Both of these data types are clearly marked as such and thus live next to each other in the RDG. The metadata from various contributors is connected to institutional data already available in the Research Information Graph (RIG), which is collated based on the data stored in NFDI4Culture's Culture Information Portal (cf. Tietz, Bruns, Söhn et al. 2023; Tietz, Bruns, Fliegl et al. 2023).[3] The RDG and the RIG together form the Culture Knowledge Graph.

To get metadata into the RDG, providers may implement the lightweight, RDF-based Culture Graph Interchange Format (CGIF) (cf. Bruns, Posthumus, Sack et al. 2023). We designed CGIF by reusing a narrow set of schema.org classes and properties. Resources can be classified as any resource class schema.org provides.[4] In addition to an identifier of the data provider and the data set, it mainly consists of a feed of individual resources with URIs enhanced by date ranges and keywords to express, for example, time, place, and motif. The keywords are IDs from authority files and controlled vocabularies such as VIAF, GND, Wikidata, Getty AAT, Iconclass, and GeoNames (cf. BARTOC, n.d.), which are used in the graph to connect resources across data providers.

As a hybrid of centralised and federated approaches, CGIF may be provided either as embedded metadata, a dedicated API, or a SPARQL endpoint/query that can be harvested periodically, or as a data dump in any RDF serialisation. The goal behind this decision is to make data contributions as easy as possible: regardless of whether a project is fully engaged in LOD and able to SPARQL, uses a content management system with limited access to its inner workings, or uses a workflow based on transforming data from XML, CSV, JSON, or other sources into various formats. As an alternative route, a conversion of LIDO to CGIF has been implemented to utilise existing, fine-grained object metadata available across projects and organisations participating in NFDI4Culture.[5]

The combined triples of the Culture Knowledge Graph (RDG and RIG taken together) are made available via the Culture Information Portal. It hosts a triple store with a SPARQL endpoint, which is also available through a search interface and may be used to query data based on, for example, one of the keywords, a specific data type, a time period, or institutions. The endpoint may also be queried by other websites to retrieve information such as related entries based on a keyword. The Culture Information Portal does not host images or other files harvested via one of the options listed above, but includes, for example, URLs of preview images and IIIF manifests, if available.

4. Implementation: technological choices

To allow for a broad range of resource types, schema.org was chosen over other data interchange options like LIDO(cf. Coburn et al. 2021), which is restricted to information about physical objects, or CIDOC

[3] In the following, ‘research information’ refers to metadata on organisations, funding, publications, and sometimes whole data sets. The Research Information Graph aims for compatibility with services like the OpenAIRE Graph (cf. Manghi et al. 2019) by implementing the CERIF data model. ‘Research data,’ on the other hand, here refers to more granular items in larger data sets. The distinction between the two can be blurry, however, when it comes to metadata ingested from long-term storage repositories like RADAR4Culture.

[4] The schema.org classes and properties have already become a de-facto standard for providing machine-readable data in websites and may, for example, provide structured data about persons, creative works, and intangible entities. They were originally produced by large corporations such as Google, Yahoo!, and Microsoft, but are now shaped and extended by a lively community.

[5] See section 5 for efforts to engage with providers who use further community standards such as Wikibase.

CRM (cf. Bekiari et al. 2022), which requires much more elaborate data than many projects under the umbrella of NFDI4Culture are able to provide. The CGIF is designed to be an intermediary that allows speedy information retrieval, and thus as an abstract addition to more specific formats across various data domains. Using schema.org for the high-level purpose of interconnecting diverse sets of research data has precedent (i.e. Verma et al. 2022, 1065, 1071). Compared to solutions like CLARIN's Component Metadata (cf. Windhouwer and Goosen 2022), schema.org is already widely used by private-sector search providers and goes beyond linguistic data. Projects which implement it as embedded metadata also make their content more machine-readable outside the realm of academic data repositories.

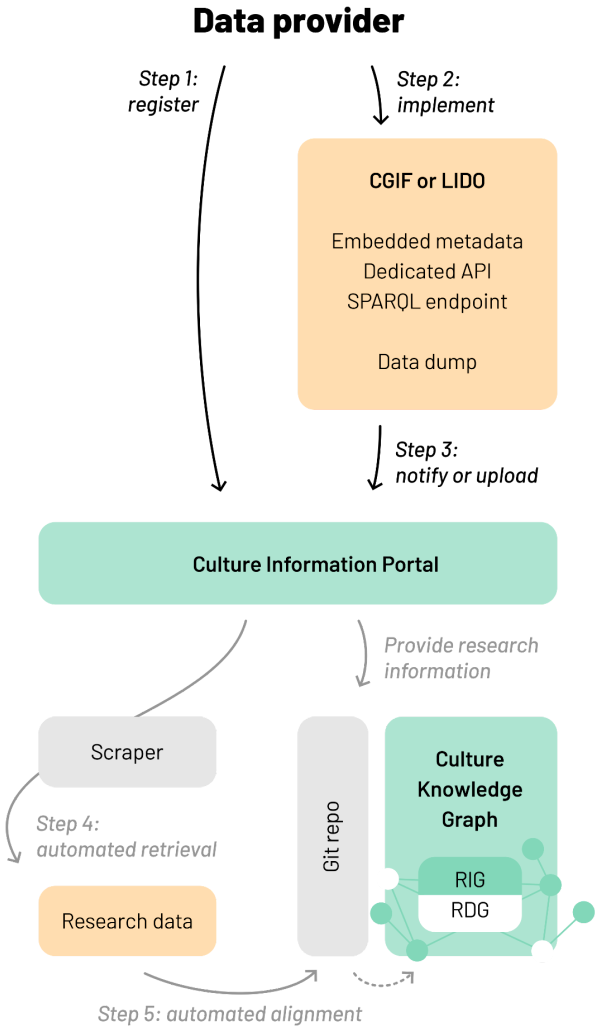

Figure 1. The current process to add research data to the Culture Knowledge Graph.

As fig. 1 illustrates, the process to add data to the RDG currently begins with a data provider (or someone acting on their behalf) registering their data set (and the institutions involved, if not available yet) in the Culture Information Portal (1). They implement CGIF or LIDO (2) and notify the portal when they are ready or upload their transformed research data as a data dump (3). The portal starts a custom scraper to generate triples from embedded metadata, a dedicated API, RDF behind a SPARQL endpoint, or a file

dump (4).[6] In a last step, the data is filtered according to the CGIF specification, aligned with the existing RIG data and the NFDICORE ontology (cf. Bruns, Sack, Posthumus et al. 2023), and saved in a Git repository, which is then ingested into the combined knowledge graph (5). The Git repository is used to version-control triples that are being ingested and to allow for reproducing the entire graph.
While the CGIF was designed as part of a low-threshold ingest pipeline, community members are invited to use, reuse, or enhance the open-source components of the Culture Knowledge Graph. The custom scraping mechanism we currently use, for example, may also be used outside the portal to harvest paginated RDF data or to test and troubleshoot CGIF implementations: the Hydra Scraper (cf. Steller 2023) was originally developed as part of the Corpus Vitrearum, one of NFDI4Culture's participants. The scraper is based on the Python library RDFLib (cf. RDFLib team 2023), due to its compatibility with various RDF dialects. Pyoxigraph (cf. Oxigraph contributors 2023) is used as a triple store due to its speed.
The Culture Knowledge Graph builds on Semantic Web technologies. The alignment routine is necessary to allow for a lightweight interchange format that is compatible with search engine optimisation *and* a graph that is easy and uniform to query via the portal's SPARQL endpoint. As part of the alignment, irrelevant triples are filtered and some of the schema.org literals are converted to allow for reasoning via SPARQL. The schema.org property “temporalCoverage,” for example, is used in the CGIF to mark a resource's temporal origin. For time-based reasoning, however, the property needs to be transformed from a string into at least two standard XML Schema “dateTime” values and may even be automatically mapped to a vocabulary in the future. Additional automated clean-ups and filters may become necessary as we proceed with integrating metadata from further sources.

5. Road ahead: accessibility and network effect

As both the graph itself and the harvesting pipeline are operational, we are now focusing on two areas to iterate upon and improve the Research Data Graph. On the one hand, we are looking to enable scholars to more easily retrieve the data they require by improving the search frontend available in the Culture Information Portal. Since SPARQL is a powerful but also challenging interface for those who are unfamiliar with RDF, we are experimenting with visual interfaces to build the highly specific queries scholars require to retrieve the right information.
On the other hand, our efforts now focus on working with individual projects, communities, and other NFDI consortia to help them contribute data, to come up with sample data transformations, and to make full use of the portal's SPARQL capabilities in web applications. To help connect the vast amount of LIDO data available across NFDI4Culture, for example, we are trialling automated transformations of the relevant metadata into CGIF via a plain ElementTree retrieval in Python, but may yet decide to make this transformation more reusable by reimplementing it in RML (cf. Dimou and Vander Sande 2022) or the web service XTriples (cf. Schrade 2019). In the same vein we are trialling automated transformations for further community standards. Since a number of participants in NFDI4Culture use Wikibase, we are also working towards a best practice for integrating CGIF classes and properties with data managed in Wikibase instances. Last but not least, we are discussing the Research Data Graph with other NFDI consortia and the international Semantic Web community aiming at further adoption, participation, and contribution.

[6] if an endpoint is used to harvest the data, a modification date in the CGIF implementation is periodically checked to see if it needs to reindex a feed and update the graph.

Appendix A

Bibliography

1. BARTOC. n.d. “Vocabularies”. Basic Register of Thesauri, Ontologies & Classifications. Accessed 3 July 2023. https://bartoc.org/vocabularies .
2. Bekiari, Chryssoula, George Bruseker, Erin Canning, Martin Doerr, Philippe Michon, Christian-Emil Ore, Stephen Stead and Athanasios Velios. 2022. *Definition of the CIDOC Conceptual Reference Model.* V. 7.1.2, June. Accessed 11 July 2023. https://www.cidoc-crm.org/sites/default/files/cidoc_crm_v7.1.2.pdf .
3. Berners-Lee, Tim. 2009. “Linked Data”. Design Issues, 18 June 2009. Accessed 30 June 2023. https://www.w3.org/DesignIssues/LinkedData.html .
4. Bruns, Oleksandra, Etienne Posthumus, Harald Sack, Linnaea Söhn, Torsten Schrade, Jonatan Jalle Steller, Tabea Tietz and Julia Tolksdorf. 2023. *Culture Graph Interchange Format Specification.* V. 1.1.0. Mainz, 22 September 2023. Accessed 2 October 2023. https://doi.org/10.5281/zenodo.8369661 .
5. Bruns, Oleksandra, Harald Sack, Etienne Posthumus, Tabea Tietz, and Jörg Waitelonis. 2023. *NFDICORE Ontology.* V. 1.1.0. Karlsruhe, 27 February 2023. Accessed 1 December 2023. https://github.com/ISE-FIZKarlsruhe/nfdicore .
6. CLARIN. n.d. “Federated Content Search (CLARIN-FCS): Technical Details”. CLARIN ERIC. Accessed 1 July 2023. https://www.clarin.eu/content/federated-content-search-clarin-fcs-technical-details .
7. Coburn, Erin, Richard Light, Jutta Lindenthal, Gordon McKenna, Regine Stein, Axel Vitzthum and Michelle Weidling. 2021. *LIDO Schema.* V. 1.1, 30 December 2021. Accessed 11 July 2023. https://lido-schema.org/schema/v1.1/lido-v1.1.html .
8. Deutsche Digitale Bibliothek. n.d. “Kultur und Wissen online”. Deutsche Digitale Bibliothek. Accessed 3 July 2023. https://www.deutsche-digitale-bibliothek.de.
9. Dimou, Anastasia, and Miel Vander Sande. 2022. *RDF Mapping Language (RML).* In collaboration with Ben De Meester, Pieter Heyvaert and Thomas Delva. V. 1.1.1, 16 November 2022. Accessed 3 July 2023. https://rml.io/specs/rml .
10. Dumont, Stefan. 2022. “Correspondence Metadata Interchange Format”. CorrespSearch, 6 March 2022. Accessed 1 July 2023. https://correspsearch.net/en/documentation.html .
11. ELEXIS. 2022. “ELEXIS Protocol for Accessing Dictionaries”. GitHub, 20 April 2022. Accessed 1 July 2023. https://elexis-eu.github.io/elexis-rest .
12. Europeana. n.d. “Discover Europe’s Digital Cultural Heritage”. Europeana. Accessed 3 July 2023. https://www.europeana.eu.
13. Fischer, Frank, Ingo Börner, Mathias Göbel, Angelika Hechtl, Christopher Kittel, Carsten Milling, and Peer Trilcke. 2019. “Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama”. In *Proceedings of DH2019*, 1–6. Utrecht: Zenodo. Accessed 2 October 2023. https://doi.org/10.5281/zenodo.4284002.
14. Isaac, Antoine, ed. 2013. *Europeana Data Model Primer.* The Hague: Europeana, 14 July 2013. Accessed 29 June 2023. https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf .

15. Isaac, Antoine, Hugo Manguinhas, Valentine Charles and Juliane Stiller, eds. 2015. *Selecting Target Datasets for Semantic Enrichment.* The Hague: Europeana, 29 October 2015. Accessed 29 June 2023. https://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_taskforces/Enrichment_Evaluation/EvaluationEnrichment_SelectingDatasets_102015.pdf .
16. Körner, Erik, Thomas Eckart, Axel Herold, Frank Wiegand, Frank Michaelis, Matthias Bremm, Louis Cotgrove, Thorsten Trippel and Felix Rau. 2023. *Federated Content Search for Lexical Resources (LexFCS): Specification.* Genève: Zenodo, 9 May 2023. Accessed 1 July 2023. https://doi.org/10.5281/zenodo.7986303 .
17. Manghi, Paolo, Alessia Bardi, Claudio Atzori, Miriam Baglioni, Natalia Manola, Jochen Schirrwagen, and Pedro Principe. 2019. "The OpenAIRE Research Graph Data Model". Zenodo. https://doi.org/10.5281/zenodo.2643199 .
18. Oxigraph contributors. 2023. *Pyoxigraph.* V. 0.3.18, 13 June 2023. Accessed 11 July 2023. https://github.com/oxigraph/oxigraph/tree/main/python .
19. Prud'hommeaux, Eric, and Carlos Buil-Aranda. 2013. *SPARQL 1.1 Federated Query.* In collaboration with Andy Seaborne, Axel Polleres, Lee Feigenbaum and Gregory Todd Williams. V. 1.1, 21 March 2013. Accessed 1 July 2023. http://www.w3.org/TR/2013/REC-sparql11-federated-query-20130321 .
20. RDFLib team. 2023. *RDFLib.* V. 6.3.2, 26 March 2023. Accessed 11 July 2023. https://github.com/RDFLib/rdflib .
21. Schrade, Torsten. 2019. *XTriples.* V. 1.4.0, 25 March 2019. Accessed 3 July 2023. https://doi.org/10.5281/zenodo.2604986 .
22. Steller, Jonatan Jalle. 2023. *Hydra Scraper.* V. 0.8.4. Mainz, 22 October 2023. Accessed 1 December 2023. https://github.com/digicademy/hydra-scraper .
23. Tietz, Tabea, Oleksandra Bruns, Heike Fliegl, Etienne Posthumus, Torsten Schrade and Harald Sack. 2023. "Knowledge Graph-basierte Forschungsdatenintegration in NFDI4Culture". In *DHd2023: Open Humanities, Open Culture: Konferenzabstracts*, 181–185. Trier: Zenodo, 10 March 2023. Accessed 2 July 2023. https://doi.org/10.5281/zenodo.7715509 .
24. Tietz, Tabea, Oleksandra Bruns, Linnaea Söhn, Julia Tolksdorf, Etienne Posthumus, Jonatan Jalle Steller, Heike Fliegl et al. 2023. "From Floppy Disks to 5-Star LOD: FAIR Research Infrastructure for NFDI4Culture". In *DaMaLOS 2023: 3rd Workshop on Metadata and Research (objects) Management for Linked Open Science*, 1–12. Hersonissos: Publisso, 16 June 2023. Accessed 2 July 2023. https://doi.org/10.4126/FRL01-006444986 .
25. Verma, Shilpa, Rajesh Bhatia, Sandeep Harit and Sanjay Batish. 2022. "Scholarly Knowledge Graphs through Structuring Scholarly Communication: A Review". *Complex & Intelligent Systems* 9, no. 1 (9 August 2022): 1059–1095. ISSN: 2198-6053, accessed 2 July 2023. https://doi.org/10.1007/s40747-022-00806-6 .
26. Vrandečić, Denny, and Markus Krötzsch. 2014. "Wikidata: A Free Collaborative Knowledgebase". *Communications of the ACM* 57, no. 10 (23 September 2014): 78–85. ISSN: 0001-0782, accessed 1 July 2023. https://doi.org/10.1145/2629489 .
27. 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. 160018 (15 March 2016): 1–9. ISSN: 2052-4463, accessed 30 June 2023. https://doi.org/10.1038/sdata.2016.18 .

28. Windhouwer, Menzo, and Twan Goosen. 2022. “Component Metadata Infrastructure”. In *CLARIN: The Infrastructure for Language Resources*, 191–222. Berlin: De Gruyter. https://doi.org/10.1515/9783110767377-008 .