Gotta Catch'em All: From Data Silos to a Knowledge Graph

Oleksandra Bruns[1,2], Linnaea Söhn[3], Tabea Tietz[1,2], Jonatan Jalle Steller[3], Etienne Posthumus[1], Torsten Schrade[3], and Harald Sack[1,2]

[1] FIZ Karlsruhe – Leibniz Institute for Information Infrastructure,
Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany
`firstname.lastname@fiz-karlsruhe.de`
[2] Karlsruhe Institute of Technology (AIFB), Kaiserstr. 89, 76133 Karlsruhe, Germany
[3] Academy of Sciences and Literature Mainz, Geschwister-Scholl-Straße 2, 55131
Mainz, Germany `firstname.lastname@adwmainz.de`

Abstract. Diverse research questions, perspectives, standards and formats across culture subject areas have led to the emergence of numerous data silos. NFDI4Culture seeks to overcome this by building a unified KG, facilitating enhanced discoverability and interoperability of distributed and heterogeneous research data. This paper outlines a pipeline for accessing and harvesting cultural heritage meta data from legacy repositories, it discusses the development of a lightweight ontology to facilitate interoperability.

Keywords: research data · knowledge graphs · infrastructure

1 Introduction

Behind every artifact lies a story that can only be uncovered when placed within the context of its cultural and historical surroundings. NFDI4Culture[4] is a consortium within the framework of the German national research data infrastructure programme (NFDI[5]) with the goal to establish an information infrastructure for cultural heritage (CH) research data. Its primary objective is to ensure the findability and interoperability of distributed and heterogeneous research data across five subject areas: architecture, art history, performing arts, musicology, and media sciences [1]. Currently, the consortium manages 68 data portals, which provide access to hundreds of diverse culture datasets[6]. This encompasses a diverse array of tangible and intangible data, ranging from architectural plans and art descriptions to music compositions, historical manuscripts and theatrical events. However the data is contextually related, each discipline approaches data management from its own unique perspective, answering different research questions. This results in the use of a range of standards and formats tailored to

[4] https://nfdi4culture.de/

[5] https://www.nfdi.de

[6] https://nfdi4culture.de/resources/dataportals

specific community needs. Such diversity enhances CH understanding, offering detailed data descriptions and facilitating research. Yet, it poses challenges for data discovery, harvesting, and integration across subjects.

One of the objectives of the NFDI4Culture is to build a Knowledge Graph (KG) to aggregate diverse and isolated meta data from the research landscape and thereby enable discoverability, interoperability and reusability of CH data. One way to achieve this is, following the established initiatives in the cultural domain, e.g. DDB[7], Europeana[8], to build a centralised infrastructure that consolidates data in one location. However, maintaining and updating such a centralized system is very resource-intensive, requiring data providers to regularly compile and contribute their data using a shared standard to access current content. Alternatively, federated infrastructures prioritize a shared API instead of shared formats. Thus, the data providers are required to create their own reliable endpoint that can be accessed and queried from the outside (e.g. CLARIN[9]). However, querying via federation can pose challenges for users, as, for the successful search, they need to be familiar with the various standards used across different data sources. Additionally, much of the data within NFDI4Culture still remains unindexed and inaccessible for querying and federation. It resides in isolated silos within legacy repositories, often challenging to even locate and harvest.

This paper reports on current efforts towards extending the NFDI4Culture-KG, in particular, it presents a KG-based workflow of harvesting research data from legacy repositories, indexing the data, and developing a lightweight ontology for data representation that facilitates easy access and search.

2 From Silos to the NFDI4Culture-KG

The NFDI4Culture-KG interconnects research data within NFDI4Culture. It comprises the Research Information Graph (RIG) and the Research Data Graph (RDG). While the RIG describes metadata about resources, e.g. publishers, contact points, standards, licences, data portals, the RDG aims to harvest and interconnect the content meta data of the resources, e.g. making individual items within data portals accessible for search. Taking into account the challenges and objectives of NFDI4Culture to aggregate and unify CH research data for improved accessibility and interoperability, an ETL (Extract, Transform, Load) environment has been designed. It consists of six modular workflow components, adaptable for independent use or within a comprehensive automated ingest routine. Once harvested and integrated, resources are accessible through a SPARQL endpoint[10] and SHMARQL[11] - a SPARQL endpoint explorer developed in NFDI4Culture, as well as a dashboard[12] for analysis and visualizations.

[7] https://www.deutsche-digitale-bibliothek.de/

[8] https://www.europeana.eu/

[9] https://www.clarin.eu/

[10] https://nfdi4culture.de/sparql

[11] https://nfdi.fiz-karlsruhe.de/shmarql

[12] https://nfdi4culture.de/go/kg-kitchen-dashboard

This section will shortly describe the components of the environment (for more details see[13]).

Step 1: Action. The goal of this step is to define a structured process for harvesting and transforming data feeds[14] into a standardized format compatible with the NFDI4Culture-KG. This involves creating an action RDF/Turtle file with schema.org-based step definitions, connecting the data feed to its metadata (in the RIG), and generating persistent identifiers for imported resources. Ensuring harmonisation and interoperability across harvested data requires mapping to a shared scheme, particularly due to a large amount of formats of sourced data such as CGIF [3], LIDO[15], MEI[16], and other unique annotation schemes. To address this, mapping to the Culture Ontology (see section 3) is provided for each data format outside the pipeline's scope.

Step 2: Cleaning. To ensure harmonisation between the harvested data feed and associated data files while preventing conflicts with information in the consumed resources that may contradict action file definitions, triples are added or deleted based on the information in the action file.

Step 3: Update. After several procedures, e.g. status checks, branch switching, managing bulk operations, and supporting various modes, the changes are pushed to the repository.

Step 4: Stash. If changes in a data feed are pushed, data directories (often called "stashes") are automatically updated or created in response to changes in the data feed. The new version is then made available via a SPARQL endpoint.

Step 5: Endpoint. To prevent downtimes, the construction of a new endpoint is realized through a Docker-based delivery workflow. If a new endpoint must be built, environment variables are adjusted to reference the new container and port for initiating the deployment. Once the new SPARQL endpoint becomes operational, the old container is stopped and removed.

Step 6: Analysis. The last component provides statistics about the integrated data feeds through the Culture Knowledge Graph Dashboard. It supports data analysis and visualizations based on the execution of provided SPARQL queries.

3 Ontologies of NFDI4Culture

The primary objective of the NFDI4Culture-KG is to link research data across NFDI4Culture subject domains through the utilization of ontologies. Every NFDI consortium shares a commitment to creating an interoperable research data infrastructure for a consortium's specific domain. The NFDIcore ontology[17] was developed to be used across consortia [2] to represent metadata about NFDI resources, e.g. persons, projects, data portals, etc., answering shared questions,

[13] https://gitlab.rlp.net/adwmainz/nfdi4culture/knowledge-graph/culture-kg-kitchen

[14] In the context of the ETL process, discussions often revolve around data feeds rather than datasets due to the emphasis on continuous streams

[15] https://www.lido-schema.org/schema/latest/lido.html

[16] https://music-encoding.org/

[17] https://github.com/ISE-FIZKarlsruhe/nfdicore

e.g. "Who is a contact point of a resource?". To answer domain-specific research questions, the NFDIcore ontology is extended following a modular approach, as e.g. with the **cul**ture **o**ntology module (CTO)[18]. This allows for providing additional metadata to describe culture resources, as e.g. property `cto:ddbAPI` enables linking a resource, e.g. a dataset, in the NFDI4Culture Portal to its corresponding entry in the German Digital Library (DDB).

While the RIG facilitates exploration and retrieval of index and metadata for NFDI4Culture resources, the primary objective of the RDG is to represent and interconnect the content of distributed data collections to address domain related research questions, e.g. "Which historical books are written by librettists and include prints showing the iconographic subject "Hercules at the Crossroads?"" [4]. The CTO provides semantics to achieve this level of research granularity in the RDG: the ontology establishes a connection between a data resource (stored and described in the RIG) and its individual component - `cto:DatafeedElement`. The ontology provides light-weight vocabulary for describing element types, subject concepts, and related concepts, including creative works, persons, locations, and temporal information by linking an element to the corresponding concepts in the external vocabularies such as GND[19], Wikidata[20], and ICONCLASS[21]. Additionally, for more detailed research, each data feed element is linked to its source file, where all the information contributed by a data provider is stored and represented using the domain-appropriate standard.

Compared to a generalist model like SCHEMA.org[22], the Culture Ontology (CTO) offers tailored advantages perfectly suited to the intricate landscape of cultural heritage research data. Specifically tailored to meet the unique requirements of the NFDI4Culture community, CTO allows for capturing the nuanced data requirements of each discipline while enabling interoperability across a wide array of cultural datasets. In contrast, utilizing domain-specific ontologies such as CIDOC-CRM[23] requires a comprehensive representation of cultural heritage information. While CIDOC-CRM provides a detailed framework, CTO distinguishes itself by its lightweight design, prioritizing flexibility and adaptability to the dynamic needs of the NFDI4Culture consortium. The primary objective of CTO is not to represent every fine-grained aspect of cultural objects, but rather to integrate the most relevant information essential for ensuring interoperability across cultural domains. This approach allows for efficient data management and harmonization while accommodating the diverse and evolving nature of cultural heritage research within the NFDI4Culture framework.

[18] https://gitlab.rlp.net/adwmainz/nfdi4culture/knowledge-graph/culture-ontology

[19] https://www.dnb.de/EN/Professionell/Standardisierung/GND/gnd_node.html

[20] https://www.wikidata.org/

[21] https://iconclass.org/

[22] https://schema.org/

[23] https://www.cidoc-crm.org/

4 Conclusion

Cultural research data is critical for the understanding of our shared human history and preserving artifacts and sites. However, it is usually dispersed across multiple legacy systems that lack interoperability. In this paper, a KG-based workflow is presented for integrating cultural content into the NFDI4Culture and thus, improving its findability and accessibility.

Both the Culture Kitchen and the Culture Ontology (CTO) are significant steps towards addressing the challenge of integrating and harmonizing diverse data sources within NFDI4Culture. By providing a common vocabulary and framework for representing cultural data, the CTO ontology facilitates standardization and interoperability. Similarly, the Culture Kitchen environment offers a practical approach for data management and integration, providing tools and workflows for aggregating, harmonizing, and exploring cultural research data. Together, these efforts contribute to the ongoing process of improving accessibility and usability within the NFDI4Culture.

In future work, the focus is on the integration of further cultural research data into the NFDI4Culture-KG, further enhancing its findability and interoperability. Additionally, efforts will continue to improve the components of the NFDI4Culture infrastructure, including the SPARQL endpoint, dashboard, SHMARQL, and other tools, to enhance their usability and functionality to accomodate users' requirements.

Acknowledgements. This work is funded by Deutsche Forschungsgemeinschaft (DFG), project number 441958017.

References

1. Bruns, O., Tietz, T., Söhn, L., Steller, J.J., Ondraszek, S.R., Posthumus, E., Schrade, T., Sack, H.: What's Cooking in the NFDI4Culture Kitchen? A KG-based Research Data Integration Workflow. In: 4th Workshop on Metadata and Research (objects) Management for Linked Open Science (DaMaLOS), co-located with ESWC (2024)
2. Sack, H., Schrade, T., Bruns, O., Posthumus, E., Tietz, T., Norouzi, E., Waitelonis, J., Fliegl, H., Söhn, L., Tolksdorf, J., Steller, J.J., Azócar Guzmán, A., Fathalla, S., Zainul Ihsan, A., Hofmann, V., Sandfeld, S., Fritzen, F., Laadhar, A., Schimmler, S., Mutschke, P.: Knowledge Graph Based RDM Solutions: NFDI4Culture - NFDI-MatWerk - NFDI4DataScience. In: 1st Conference on Research Data Infrastructure (2023)
3. Steller, J.J., Söhn, L.C., Tolksdorf, J., Bruns, O., Tietz, T., Posthumus, E., Fliegl, H., Pittroff, S., Sack, H., Schrade, T.: Communities, Harvesting, and CGIF: Building the Research Data Graph at NFDI4Culture. In: DHd2024: Quo Vadis (2024)
4. Tietz, T., Bruns, O., Söhn, L., Tolksdorf, J., Posthumus, E., Steller, J.J., Fliegl, H., Norouzi, E., Waitelonis, J., Schrade, T., Sack, H.: From Floppy Disks to 5-Star LOD: FAIR Research Infrastructure for NFDI4Culture. In: 3rd Workshop on Metadata and Research (objects) Management for Linked Open Science (DaMaLOS), co-located with ESWC 2023. Publisso (2023)