Knowledge Representation and Discovery for Cultural Heritage Research Data with CTO and SHMARQL

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

[1]FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Eggenstein-Leopoldshafen, Germany
[2]Institute of Applied Informatics and Formal Description Methods (AIFB) of KIT, Karlsruhe, Germany
[3]Academy of Sciences and Literature Mainz, Geschwister-Scholl-Straße 2, 55131 Mainz, Germany

Abstract

This paper presents an approach to representing, discovering, and exploring research (meta)data in the cultural heritage (CH) domain. One key component is the NFDI4Culture Ontology (CTO), a modular ontology for CH research data. Aligned with the Basic Formal Ontology (BFO) and extending the mid-level ontology NFDIcore, CTO enables structured, interoperable metadata integration across diverse CH domains. Another key component is the lightweight Linked Data platform SHMARQL, which supports querying and storytelling with RDF data, offering new possibilities for data discovery, reuse, and cross-domain integration. Both CTO and SHMARQL are integrated within NFDI4Culture, a consortium of the German NFDI programme for the national research data infrastructure that focuses on material and immaterial CH data. NFDI4Culture aims to make heterogeneous and decentralized research data findable and interoperable through the NFDI4Culture-KG and Portal. Designed with modularity and reuse in mind, both tools demonstrate great generalizability and have been successfully applied beyond their original CH context.

Keywords

Ontologies, Cultural Heritage, Research Data Management, Interoperability, Linked Data Platform, Exploration

1. Introduction

In response to the accelerating growth of scientific publications and research artifacts, research knowledge graphs (RKGs) have gained prominence[1] [1, 2], as they have the potential to represent scholarly information in a machine-understandable, semantically rich, and interlinked way. Capturing research resources like datasets, collections, and software and their relationships in a knowledge graph (KG) enables advanced searches, analysis, and reuse of knowledge across disciplines. A crucial foundation of this approach is the use of ontologies and standardized vocabularies to provide a common, context-sensitive data model, along with persistent identifiers to uniquely reference resources. This ensures that the KG remains interoperable and actionable at scale, addressing the heterogeneity and fragmentation that often affect traditional scholarly data management. The domain of cultural heritage (CH) exemplifies both the need for and the potential of RKGs, as it brings together diverse, richly contextual research data that are typically distributed across heterogeneous and decentralized sources. The CH research landscape is diverse and spans various fields, each with its own perspectives, metadata standards, and data formats. In Germany the consortium NFDI4Culture (part of the National Research Data Infrastructure, NFDI[2]) brings together numerous subject-specific data portals and collections from the domains architecture, art history, musicology, performing arts, and media studies, each maintained

5th International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment, Nov 2024, Nara, Japan

tabea.tietz@fiz-karlsruhe.de (T. Tietz); etienne.posthumus@partners.fiz-karlsruhe.de (E. Posthumus); oleksandra.bruns@fiz-karlsruhe.de (O. Bruns); linnaea.soehn@adwmainz.de (L. Söhn); jonatan.steller@adwmainz.de (J. J. Steller); joerg.waitelonis@fiz-karlsruhe.de (J. Waitelonis); torsten.schrade@adwmainz.de (T. Schrade); harald.sack@fiz-karlsruhe.de (H. Sack)

0000-0002-1648-1684 (T. Tietz); 0000-0002-0006-7542 (E. Posthumus); 0000-0002-8501-6700 (O. Bruns); 0000-0001-8341-1187 (L. Söhn); 0000-0002-5101-5275 (J. J. Steller); 0000-0001-7192-7143 (J. Waitelonis); 0000-0002-0953-2818 (T. Schrade); 0000-0001-7069-9804 (H. Sack)

[1] NFDI-MatWerk Knowledge Graph https://nfdi.fiz-karlsruhe.de/matwerk/

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

by different institutions and using distinct metadata standards. Although this diversity reflects the richness of CH data ranging from art object descriptions to musical scores and historical performance events, it also resulted in data silos that prevent unified discovery and analysis. The lack of a common representation and access point for these distributed resources not only constrains data findability and interoperability, but also limits the potential for broader scholarly knowledge integration, as e.g., linking CH data with other scientific domains.

NFDI4Culture has recognized the need for an approach that represents and links CH research data in a unified way, while respecting the specific requirements of the domain. In line with the broader vision of scientific KGs, NFDI4Culture has been creating a unified KG[3] for CH research data with the goal of providing a centralized, semantic index over decentralized and heterogeneous data sources. In doing so, CH datasets from different institutions and disciplines are discoverable and queryable, enabling researchers to ask complex questions across silos. One key component of this effort is the NFDI4Culture Ontology (CTO) [3], a domain-specific ontology designed to represent CH research resources and their metadata in a semantically rich and consistent way. The CTO extends the mid-level ontology NFDIcore [4] and is aligned with the Basic Formal Ontology (BFO) [5, 6, 7] and other cross-domain standards. This alignment ensures that, while the ontology is tailored to the nuanced needs of CH data, it remains interoperable with ontologies and data from other domains. CTO employs a lightweight and modular design to prioritize flexibility and integration of only the most relevant information needed for cross-domain interoperability.

The SHMARQL tool is another key component as part of the NFDI4Culture effort. The Linked Data publishing platform has been created to easily browse the structure of the data in a KG. The tool, originally created for NFDI4Culture, has now been adopted by a number of different disciplines and projects. Both, CTO on the side of data representation and SHMARQL on the side of discovery and the publication of LOD are fully integrated into the NFDI4Culture Information Portal[4]. The portal makes available the data by means of a public SPARQL endpoint[5] with currenly more than 70m triples has been made publicly available[6].

This paper is guided by the following research questions:

RQ1: How can cultural heritage research resources be represented to establish a unified index and centralized access point for decentralized and heterogeneous research data within the (German) cultural heritage landscape?

RQ2: How can the requirements of the cultural heritage domain for the representation and discovery of research resources be met, while also providing interoperability to other domains, i.e. other NFDI consortia and beyond?

RQ3: How can cultural heritage research resources be published and queried effectively to enable the discovery of decentralized research resources through a unified and centralized access point?

The contributions in this paper include:

C1: **Knowledge representation for CH research**: The NFDI4Culture Ontology (CTO) has been created for CH research (meta)data. CTO extends NFDIcore and maps to BFO and standard vocabularies (e.g. Schema.org). This paper also contributes lessons learned from applying earlier versions of CTO in a productive system. These experiences informed the development of version 3.0 and may serve as guidance for similar efforts in other domains. CTO was released on github[7]. A documentation is provided[8] along with a list of resources[9]. The NFDI4Culture-KG, built on

[3] NFDI4Culture-KGhttps://nfdi4culture.de/services/details/culture-knowledge-graph.html
[4] NFDI4Culture Portalhttps://nfdi4culture.de/
[5] Public SPARQL Endpointhttps://nfdi4culture.de/resources/knowledge-graph.html
[6] Public Dashboardhttps://superset.nfdi4culture.de/superset/dashboard/culture-kg-kitchen/
[7] https://github.com/ISE-FIZKarlsruhe/nfdi4culture
[8] https://nfdi.fiz-karlsruhe.de/4culture/
[9] https://nfdi.fiz-karlsruhe.de/4culture/ontology/

CTO and NFDIcore, is accessible via a SPARQL endpoint[10] and currently contains more than 70m triples. Its current statistics are available through a dashboard[11]

C2: **Discoverability via SHMARQL**: The versatile SHMARQL tool provides a lightweight web user interface for scholars to intuitively navigate the graph, create queries, and visualize results without deep technical expertise. It further serves as a semantic web application framework on which experienced practitioners can build custom projects disseminating datastories explaining and analyzing their KGs. It is available as open-source software [12] with online documentation [13] .

In the following sections, the contributions of this work are placed within their scientific context (Section 2), followed by a description of the newly published CTOv3.0, which marks a substantial revision from earlier versions and is presented for the first time (Section 3), as well as the implementation of SHMARQL (Section 4). Section 5 highlights how these contributions are embedded within the NFDI4Culture framework.

2. Related Work

This section relates the contributions of this paper to two areas central to this approach: ontologies developed for representing research data, and tools that support the publication and exploration of Linked Data.

2.1. Ontologies

The Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) [8] is a foundational ontology developed to capture ontological categories rooted in natural language and human common sense. Although DOLCE has seen widespread use in domains such as CH and digital humanities (DH), the ontologies presented in this contribution are designed to support interoperability across a broader range of domains, including the natural sciences, engineering, and research data management. To better align with the ontological frameworks commonly used in these fields, the decision was made to base the presented ontologies on BFO, which is widely adopted in scientific and data-intensive research contexts.

The VIVO ontology [9] models the scholarly context of researchers–including their outputs, interests, and accomplishments. It was developed for the VIVO software and integrates established standards such as BFO, Dublin Core[14], and the Event ontology[15]. While VIVO offers detailed representations of academic activities, it is relatively complex and tailored to research information systems. In contrast, this paper presents the modular BFO-compliant mid-level ontology NFDIcore and the domain extension CTO, designed to integrate heterogeneous research domains.

Although CIDOC-CRM[16] is widely used in the CH and digital humanities domains for research data management, it was originally designed to represent cultural objects, events, and their relationships. While CIDOC is well-suited for many heritage contexts, its event-based modeling paradigm is insufficient for the interdisciplinary requirements of the research addressed in this contribution, which demands a process-based approach.

The Scientific Knowledge Graphs Interoperability Framework (SKG-IF)[17] addresses interoperability by providing a high-level reference model for aligning heterogeneous scholarly graphs. The Core Data Set for Research (KDSF)[18] is a standardized model to represent research information within the German

[10] https://nfdi4culture.de/resources/knowledge-graph.html
[11] https://superset.nfdi4culture.de/superset/dashboard/culture-kg-kitchen/
[12] https://github.com/epoz/shmarql
[13] https://shmarql.com/
[14] https://www.dublincore.org/resources/glossary/ontology/
[15] http://motools.sourceforge.net/event/event.html
[16] https://www.cidoc-crm.org/
[17] https://skg-if.github.io/
[18] https://kerndatensatz-forschung.de/

academic system, designed to harmonize research reporting across universities and research institutions. However, KDSF is tied to national structures and reporting requirements, which makes it overly specific for the broader, cross-disciplinary scope of this work.

The DCAT vocabulary[19] enables standardized dataset and data services descriptions to enhance discoverability across web-based catalogs. However, DCAT operates exclusively at the metadata level and lacks the semantic depth required to model domain-specific content, relationships, and provenance information. Although DCAT may complement certain aspects of research data infrastructure, it is not suitable for the sophisticated ontological modeling presented in this contribution.

Schema.org[20] is a community-driven standard for structuring web data, including research data, to enhance discoverability and integration with web technologies. However, its limited semantic expressivity, inherent ambiguity, and incompatibility with certain domain standards (e.g., materials science), render it insufficient for the requirements of this work. To address this limitation, the contributed NFDIcore ontology provides Schema.org mappings, while the NFDI4Culture ontology selectively reuses relevant terms.

2.2. Discovery Tools

Several tools have been developed to support RDF data exploration and querying, addressing user needs ranging from visual interfaces to advanced querying functionalities. The following section reviews prominent examples focussing their relationship to SHMARQL functionality. The RDF Playground [10] simplify RDF data interaction through comprehensive SPARQL query support. It provides an educational platform enabling users can create SPARQL queries within a web-based environment. While RDF Playground shares SHMARQL's SPARQL-based approach to RDF manipulation, it priorizes learning and experimentation over streamlined KG exploration. Similarly, Linked Data Fragments (LDFs) [11] enable users to interact with RDF data through minimal setup requirements, facilitating querying and exploration of fragmented KGs. However, LDFs' primary focus on the enhancement of large KG exploration by decomposing them into smaller, more manageable fragments. This fragmentation approach enables more efficient querying and reduces the operational complexity associated with large RDF datasets.

Linked Data browsers enable users to navigate Linked Data through interfaces that emphasize entities and their interconnected relationships. For instance, LodLive [12] facilitates interactive browsing with dynamic Linked Data exploration, supporting a graph-centric perspective. LODmilla [13] emphasizes entity-centric browsing, enabling users to explore Linked Data through specific entities and their associated properties. LodView[21] is a widely-used Linked Data browser that supports both an entity-centric and graph-centric exploration of RDF datasets. It presents Linked Data through an intuitive interactive web interface, allowing users to navigate RDF triples and visualize their relationships seamlessly. In contrast to these Linked Data browsers, SHMARQL empowers users to both browse RDF data and execute custom SPARQL queries, enabling targeted filtering and retrieval of specific information from RDF dataset.

Another category of RDF discovery tools comprises graph visualization platforms, such as GraphDB[22] and Stardog Studio[23]. These tools enable users to interact with RDF data through intuitive visualizations that illuminate relationships between entities, thereby facilitating navigation of complex datasets for non-expert users. While these platforms enhance discovery through graphical representations of RDF data, SHMARQL adopts a fundamentally different approach centered on querying and exploring RDF data via SPARQL endpoints. This enables users to directly interact with and retrieve targeted data from any RDF knowledge graph without depending on pre-configured visual interfaces.

In summary, all aforementioned tools aim at providing streamlined and user-friendly approaches for

[19] https://www.w3.org/TR/vocab-dcat-3/

[20] https://schema.org/

[21] https://github.com/linkeddatacenter/app-lodview

[22] https://graphdb.ontotext.com/documentation/11.0/index.html

[23] https://www.stardog.com/studio/

RDF data discovery and exploration. SHMARQL distinguishes itself by offering an on-the-fly approach that enables users to immediately explore RDF data by simply providing the URL of any SPARQL endpoint. Unlike other tools, SHMARQL's integration with full-text search via Fizzysearch[24] simplifies the incorporation of full text queries into SPARQL. This integration makes SHMARQL capable of both effortlessly querying existing KGs and publishing new ones with minimal configuration requirements.

3. Data Representation with the NFDI4Culture Ontology (CTO)

NFDI4Culture aims to establish an information infrastructure for material and immaterial CH research data. Throughout the NFDI4Culture project, a KG has been developed and integrated with the Culture Information Portal, utilizing the NFDI4Culture Ontology (CTO) as its foundational model. CTO supports the integration of research metadata through a dedicated ETL (Extract, Transform, Load) pipeline. The NFDI4Culture-KG provides a single point of access to decentralized cultural heritage research resources serving as an index of the communities' research data. The comprehensive development of the KG and Portal has been presented in previous work [3, 14, 15, 16]. The following section introduces CTO v3.0, which differs significantly from earlier versions and has not been described in a peer-reviewed publication.

3.1. Requirements for the Ontology Modeling and Development Process

The requirements for the development of CTO were initially derived from user stories collected together with the scientific communities engaged in NFDI4Culture at the beginning of the project involving GLAM institutions (galleries, libraries, archives, museums), universities, academies, as well as individual researchers. The user stories are publicly available[25] and the extracted requirements are discussed in [17]. Based on these initial inputs, workshops with domain experts were conducted to further refine the ontology's scope and structure. The subsequent integration of CTO into the NFDI4Culture Portal and KG infrastructure has revealed additional requirements:

REQ1: **Complexity:** While the initial CTO design followed a deliberately ultra-lightweight approach aimed at broad applicability with minimal complexity, the integration of real-world data exposed several limitations and ambiguities. The need emerged to increase the expressivity of the ontology while still maintaining the overall goal of a lightweight and pragmatic design.

REQ2: **Domain Requirements and Granularity:** Individual domains like musicology and the performing arts required more expressive representations than initially anticipated.

REQ3: **Authority Data:** In CH data references to authority data and external vocabularies are essential for identification, classification, and for revealing connections between datasets by linking entities across different sources. Supporting these references in a way that is more structured, understandable, and effectively queryable way became a key consideration for the revised design.

REQ4: **Licensing and Rights Statements:** The integration of research metadata into the KG and, consequently, into the portal revealed the need for a more nuanced representation of licenses and rights statements, i.e. to support both structured license URIs and natural language usage restrictions, depending on the level of detail and format given by the data providers.

REQ5: **BFO Alignment:** CTO extends NFDIcore, which in earlier versions was aligned with BFO 2.0. Following a joint decision by the NFDIcore stakeholders to update the ontology and align with BFO 2020, a new requirement for the revised version of CTO was to adopt this alignment, ensure consistency with the NFDIcore ontology, and support interoperability across NFDI consortia.

REQ6: **Persistent IRIs and Label Independence:** Experience gained from integrating CTO in a productive system revealed that human-readable labels needed to be decoupled from technical identifiers to ensure long-term maintenance, reliable referencing, and stable querying as the ontology evolves.

[24] https://github.com/ISE-FIZKarlsruhe/fizzysearch

[25] https://nfdi4culture.de/resources/user-stories.html

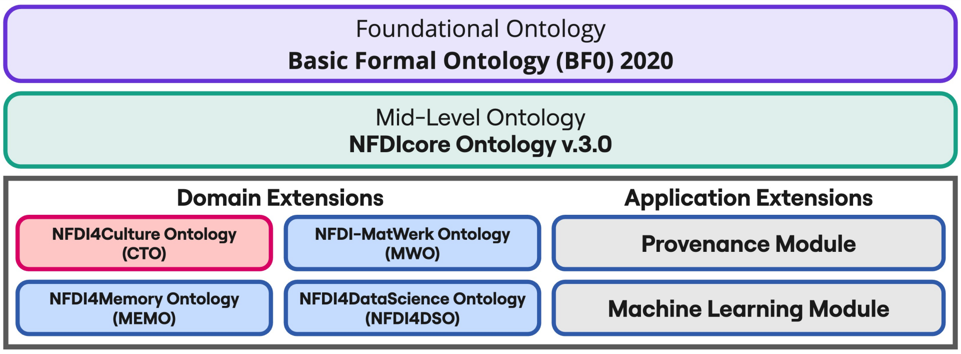

Foundational Ontology
Basic Formal Ontology (BFO) 2020
Mid-Level Ontology
NFDIcore Ontology v.3.0
Domain Extensions
Application Extensions
NFDI4Culture Ontology (CTO)
NFDI-MatWerk Ontology (MWO)
NFDI4Memory Ontology (MEMO)
NFDI4DataScience Ontology (NFDI4DSO)
Provenance Module
Machine Learning Module

Figure 1: Overview of the modular structure and relationship of NFDIcore, BFO and CTO

REQ7: **Transparency, Collaboration, and Quality Control:** A more transparent, reproducible, and quality-controlled ontology development process was needed to better support collaboration across contributors and facilitate easier reuse of CTO.

3.2. A Modular Design Approach with BFO and NFDIcore

To support interoperability across the diverse domains represented within the NFDI, which spans all scientific disciplines and humanities, and to enable federated queries across multiple consortia, a modular ontology design has been adopted. This approach balances the shared goals and structures of NFDI consortia with their individual disciplinary requirements, enabling both cross-domain knowledge discovery and domain-specific representation (REQ5).

At the center of this modular approach is NFDIcore [4]. The mid-level ontology is aligned with BFO and integrates established models such as the Information Artefact Ontology (IAO) [18], the Software Ontology (SWO) [19], and the Ontology of Bioscientific Data Analysis and Management [20]. NFDIcore provides a structured framework for representing both the organizational landscape of NFDI and the research data contributed by participating projects. The ontology captures a range of concepts including organizations, projects, datasets, research outputs, technical standards, software, services, and geographical information. It supports consistent data modeling, enhances interoperability, and facilitates the discovery and reuse of research data across different disciplines and infrastructures.

NFDIcore was initially developed based on user stories and competency questions provided by the consortia NFDI4Culture, NFDI4Memory[26], NFDI-MatWerk[27], and NFDI4DataScience[28] [21]. The ontology has since evolved through regular stakeholder meetings, active participation in NFDI working groups, and continuous issue tracking and discussion on GitHub. It is maintained by an engaged community, supported by a discussion forum, defined release schedules, and clearly documented milestones.

To meet the specific needs of individual domains, NFDIcore is designed to be extended by application ontologies such as CTO, the NFDI-MatWerk Ontology (MWO) [22] on materials science, the NFDI4Memory Ontology (MEMO) [23] on history, and the NFDI4DataScience Ontology (DSO) [24]. These ontologies build upon NFDIcore's shared structure while providing the expressivity required for their respective communities. Functional extensions applicable across consortia, such as provenance models or machine learning components are currently under development. In regular community exchanges, it is evaluated which concepts from domain-specific extensions are suitable to be represented as part of the shared mid-level ontology NFDIcore and which elements remain domain and function specific. Any consortium and domain is welcome to provide their own extension to NFDIcore. A curation process that improves the possibilities of participation is currently in development.

[26] https://4memory.de/
[27] https://nfdi-matwerk.de/
[28] https://www.nfdi4datascience.de/

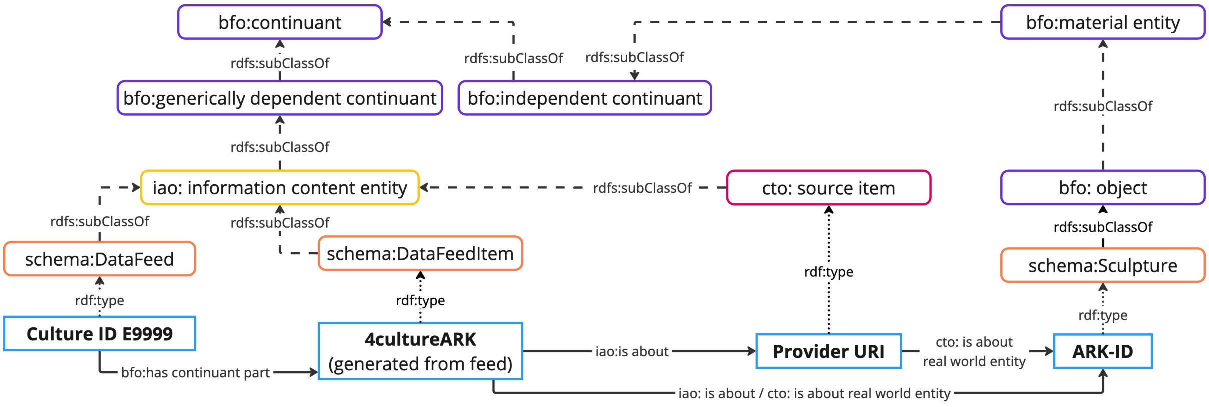


Figure 2: The core structure of the NFDI4Culture Ontology (CTO)

3.3. CTO and the NFDI4Culture Domains

CTO is designed to represent CH research data within the framework of a unified data index. Its primary scope is to enable centralized access to decentralized resources from the domains of musicology, performing arts, media studies, architecture, and art history. Thereby, CTO focuses on three core areas: (1) the representation of the consortium and its infrastructure, including persons, organizations, services, standards, and events; (2) the content of cultural heritage research data, such as objects, persons, locations, and events referenced in the metadata, along with associated media and external vocabulary links; and (3) access and reuse aspects, including legal statements, licenses, contact information, and supported export formats.

The ontology's structure comprises four main components, illustrated in figure 2. A `schema:DataFeed` represents a data feed created and maintained via the Culture Information Portal once a data provider is ready to integrate their data. Each feed is described by metadata such as contact persons, licenses, export formats, and institutional affiliations. For each record within a feed, a persistent ARK-ID is issued as a `schema:DataFeedItem`. This stable reference entity contains minimal metadata (e.g., license and timestamps) and functions as a durable anchor in the KG, even if the underlying content is altered or deleted.

Content-related metadata are associated with a `cto:CTO_0001005` (source item), representing the provided research resource. This includes media references, external identifiers, temporal metadata, and subject-specific details such as musical incipits (usually the first few bars of a musical piece, presented in standard music notation). A source item may also be linked to the real-world entity it describes (e.g., sculpture, building, person, or text); however, this linkage is only materialized when explicitly stated in the source data.

Figure 3 illustrates the use of CTO in the musicology domain. The source item *Frühlingsgruß* is published by the organization RISM Online[29], which aggregates musical records from international collections. The associated person *Robert Schumann* is referenced in the original provider data. To preserve the lightweight design (REQ1), the specific relationship type is not modeled in detail but represented using the `cto:CTO_0001009` (has related person) property, because in the index it is merely relevant to know "*which persons are related to this data feed*". The same modeling approach applies to related events, organizations, and locations.

An ARK-ID[30] is assigned to the entity *Robert Schumann* in the KG due to its relation to *Frühlingsgruß* in the provider metadata. Since the RISM identifier is provided in the source metadata, and such identifiers must be queryable (e.g., "*Select all entities with a RISM identifier*"), the class `nfdicore:NFDI_00001016` (nfdicore: rism identifier) is introduced as a subclass of `iao:IAO_0000578` (iao: centrally registered

[29] https://rism.online/
[30] https://arks.org/about/ark-overview/

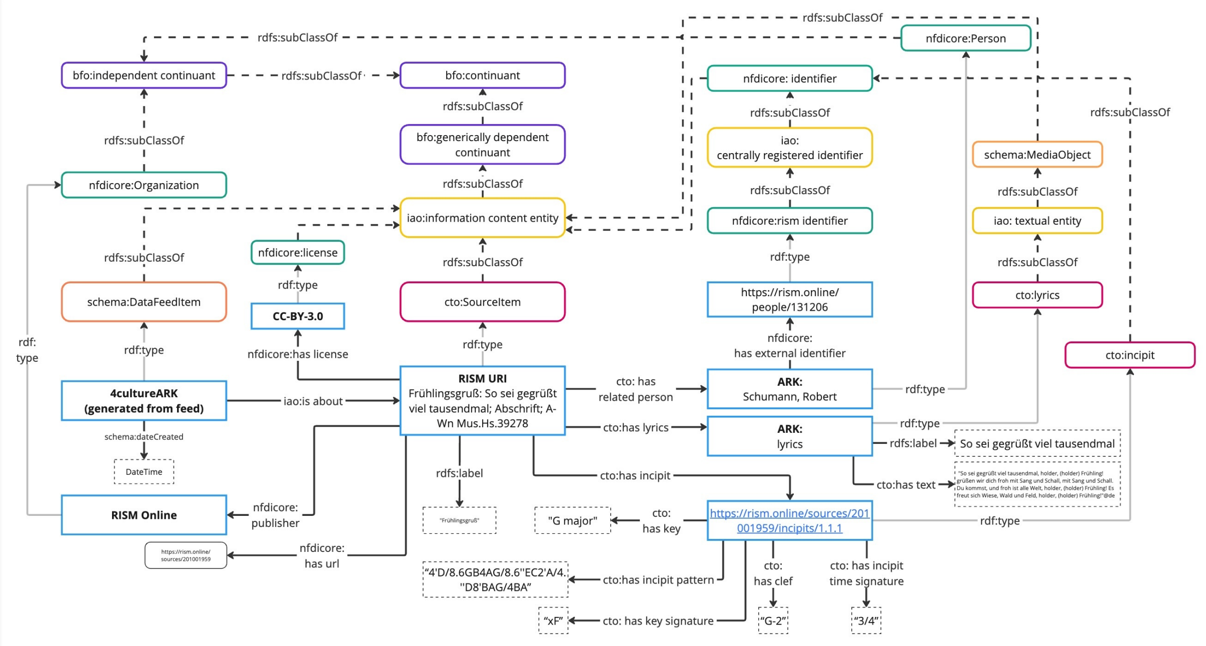


Figure 3: Modeling example in the musicology domain

identifier), as specified in REQ3. Additionally, the representation of lyrics and incipits addresses a subject-specific requirement in the musicology domain (REQ2).

Likewise, the modeling example for the performing arts domain is shown in figure 6 in Appendix A. In this case, the provider URI references the `schema:TheaterEvent` which was integrated as a subclass to `bfo:BFO_0000003` (occurrent). The source metadata includes the AAT classification[31] "*plays*"[32]. To support such classifications, the class `cto:CTO_0001027` (cto:classifier) was introduced, following the design pattern to the `nfdicore:NFDI_0000015` (nfdicore: identifier) classes. In addition to AAT, CTO incorporates various classification systems, including Iconclass [25], UNESCO[33], and CERL[34]. The source item also references images in its metadata, each of which is represented in CTO by its content URL and license statement.

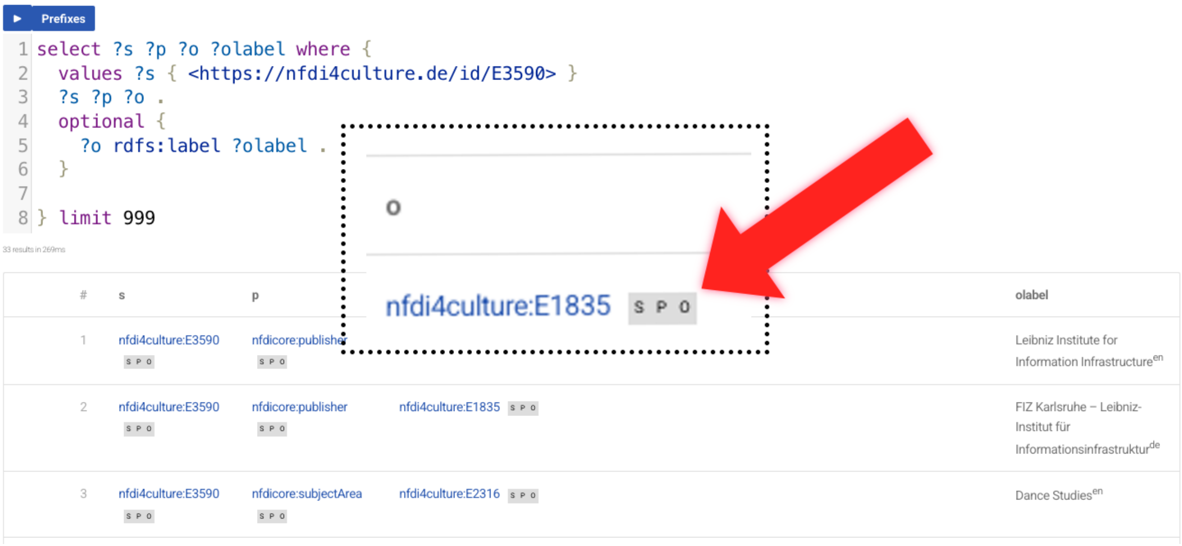
3.4. Bridging Community Needs and Cross-Domain Integration

The presented approach addresses RQ1 by providing a consistent representation framework that supports data harvested from multiple domains within the cultural heritage landscape. Through its structured integration with the Culture Information Portal, CTO facilitates centralized access while maintaining the decentralized nature of data storage and ownership. RQ2 is addressed by showing how CTO meets the specific needs of the cultural heritage community while remaining interoperable with other NFDI consortia and tasks through its alignment with NFDIcore, BFO, and the reuse of core ontologies such as IAO and SWO. The modular design ensures compatibility across domains without sacrificing domain-specific expressivity.

CTO demonstrates its generalizability through the direct reuse by the NFDI4Memory consortium [23]. Its development and application have also been actively discussed in various working groups beyond the core development team.

To meet REQ7, the Ontology Development Kit (ODK) was adopted as the foundation for the CTO

[31] https://www.getty.edu/research/tools/vocabularies/aat/
[32] http://vocab.getty.edu/aat/300201028
[33] https://vocabularies.unesco.org/browser/thesaurus/en/
[34] https://www.cerl.org/resources/cerl_thesaurus/main



```
select ?s ?p ?o ?olabel where {
  values ?s { <https://nfdi4culture.de/id/E3590> }
  ?s ?p ?o .
  optional {
    ?o rdfs:label ?olabel .
  }

} limit 999
```

33 results in 269ms

#	s	p	o	olabel
1	nfdi4culture:E3590 S P O	nfdicore:publisher S P O		Leibniz Institute for Information Infrastructure^{en}
2	nfdi4culture:E3590 S P O	nfdicore:publisher S P O	nfdi4culture:E1835 S P O	FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur^{de}
3	nfdi4culture:E3590 S P O	nfdicore:subjectArea S P O	nfdi4culture:E2316 S P O	Dance Studies^{en}

Figure 4: A screenshot of browsing triples with SHMARQL. Portion of the screen zoomed in to show clickable links choosing s,p,o position

development workflow [26]. Although originally created for the OBO Foundry[35] and hence, the biomedical domain, ODK offers a structured environment that supports transparent, reproducible, and collaborative ontology engineering for all domains. ODK is especially useful for CTO, because its quality control capabilities combine the ROBOT tool[36] for validation, and reporting with automated workflows implemented through GitHub Actions. This helps to ensure consistency, supports continuous testing, and contributes to the long-term maintainability and reusability of CTO. Finally, ODK facilitates interdisciplinary ontology development by promoting modularity, alignment with upper ontologies, and reusable design patterns.

Section 5 further describes the application of CTO together with the SHMARQL tool in the NFDI4Culture infrastructure and provides examples for its applicability in the NFDI4Culture consortium.

4. SHMARQL as Linked Data Publishing Platform

SHMARQL is a novel Linked Data publishing platform that enables semantic web professionals to disseminate data effectively. Similar to LODView[37], LODE [27], or YASGUI[38] and integrating functionality from these systems into a cohesive whole.

SHMARQL is developed in Python[39] and designed as a building-block for web application development. It is built using FastHTML[40], a modern Python web application framework enabling rapid development and extension. Originally developed to support the NFDI4Culture project's data publishing requirements, SHMARQL has since been generalized into a versatile platform and adopted by multiple other research projects. A core feature of SHMARQL is the ability to easily browse the structure of data published in a knowledge graph by navigating properties and utilizing IRIs in either subject (head) or object (tail) positions, as demonstrated in Figure 4. This feature is commonly supported in commercial KG publishing tools.[41] However, when exploring a new KG that has not been published in a system exposing this feature for the first time, it typically requires cumbersome manual SPARQL query creation.

[35] https://obofoundry.org/
[36] https://robot.obolibrary.org/
[37] https://lodview.it/
[38] https://yasgui.org/
[39] https://www.python.org/
[40] https://www.fastht.ml/
[41] For example in the Triply https://triply.cc/ system

SHMARQL can be used ”out-of-the-box” by any user with the goal to explore existing KGs published via a SPARQL endpoint, or it can be extended and customized as a web application for publishing new knowledge graphs from RDF data.

A distinctive feature of SHMARQL is its integrated data stories functionality, which combines markdown-based narrative documentation with dynamic charting capabilities to create compelling, interactive presentations of knowledge graph content. Data Stories work by allowing users to create Markdown[42] files that seamlessly embed interactive visualizations, charts, and graphs directly within explanatory text, transforming static documentation into engaging and exploratory experiences. This integrated approach enables researchers and data publishers to craft coherent narratives around their datasets while providing immediate access to the underlying RDF data through embedded SPARQL queries and visualizations. As the KG evolves, the associated charts and data representations within these stories can be automatically updated, ensuring consistency between the narrative context and the current state of the data.

An ongoing challenge in SPARQL-based KG querying is the frequent requirement for full-text search capabilities, which are not standardized within the SPARQL specification itself. Although researchers and practitioners routinely need to perform textual searches across literal values, such as finding entities by partial name matches, searching within descriptions, or identifying concepts through keyword queries, the absence of standardized full-text operators forces implementations to rely on vendor-specific extensions or cumbersome workarounds.

To address this challenge of interoperability, SHMARQL has been designed with the explicit goal of enabling effortless KG deployment with integrated full-text search capabilities through minimal configuration requirements. Rather than forcing users to navigate complex triplestore-specific implementations or invest significant effort in setting up custom search infrastructures, SHMARQL abstracts away these technical complexities by providing a unified interface that seamlessly combines SPARQL querying with full-text search functionality.

Given a collection of n-triple or turtle files in a directory on disk, a single command is needed for a user to launch a SHMARQL instance with a fully-functional triplestore:

```
docker run --rm -it -v $(pwd):/data -e DATA_LOAD_PATHS=/data -p 8000:8000 ghcr.io/epoz/shmarql
```

This approach not only eliminates the traditional barriers to implementing text-based discovery in semantic web applications but also supports the platform’s broader mission of simplifying RDF data publishing with custom styling and enabling compelling data storytelling through integrated access to underlying data for exploration and analysis.

4.1. SHMARQL Platform Architecture

SHMARQL is publicly available[43] and offers deployment flexibility through its containerized Docker architecture. Users can run SHMARQL as a standalone solution to ingest and publish various RDF data formats (including n-triples, n-quads, turtle, and other standard serializations) using its built-in triplestore of high performance powered by Oxigraph[44], a Rust-based engine known for its speed and efficiency. Alternatively, SHMARQL can be configured as a sophisticated front-end interface to existing triplestore infrastructures, seamlessly integrating with popular solutions such as Virtuoso[45], Qlever[46], Apache Jena Fuseki[47], GraphDB [48], or any commercial provider that exposes a SPARQL 1.1 compliant endpoint.

When using the built-in Oxigraph triplestore, there is no need to configure or manage an additional server, and importing RDF files is both fast and straightforward. The Oxigraph software is packaged

[42] https://en.wikipedia.org/wiki/Markdown

[43] https://shmarql.com/

[44] https://github.com/oxigraph/oxigraph

[45] https://www.virtuoso.com/

[46] https://github.com/ad-freiburg/qlever/

[47] https://jena.apache.org/documentation/fuseki2/

[48] https://graphdb.ontotext.com/

as a library directly integrated into the SHMARQL system. This integration allows users to specify the disk location of their data files, enabling automatic data ingestion during system startup. For moderately sized knowledge graphs containing up to several hundred thousand triples, this ingestion process is sufficiently fast to be performed dynamically. For larger knowledge graphs, the system can be configured to persist imported data, eliminating the need for re-ingestion upon application restart.

Alternatively, when using SHMARQL as a front-end to an existing triplestore, users need only specify the target SPARQL endpoint, and all queries are proxied and cached to the existing system. The queries are executed server-side using Python code, which resolves incompatibility issues caused by Cross-Origin Resource Sharing (CORS)[49] security restrictions that occur when SPARQL queries are executed from front-end JavaScript libraries against endpoints lacking proper CORS header configuration.

To extend standard SPARQL syntax with full-text and similarity search capabilities, SHMARQL leverages the FIZzysearch[50] SPARQL rewriter and the bikiDATA[51] RDF indexing and querying library. The FIZzysearch module employs a SPARQL syntax-aware parser[52] to parse queries into concrete syntax trees and provides contextual replacements of configurable query components. This approach enables the addition of features such as full-text searching through ”magic triples” to existing triplestores, even those not originally published with SHMARQL. For example:

```
select ?s where {
    ?s fizzy:fts "something" .
}
```

The system executes a full-text query to retrieve all triples containing literal strings that match the term ”something”. It also supports wildcard patterns and boolean query operations for enhanced search capabilities. Beyond full-text search functionality, the system can be extended with additional enhancements such as RDF2Vec-based knowledge graph embeddings[53] or sentence embeddings[54] to enable similarity-based searching.

For data story creation, SHMARQL utilizes the widely adopted MkDocs[55] documentation system. A custom plugin was developed for MkDocs to enable ’shmarql’ syntax code blocks within markdown documents, which display SPARQL query results directly in the narrative rather than exposing the raw query source. This approach facilitates compelling data storytelling while preserving analytical transparency by allowing readers to access the underlying queries for further analysis or data verification (cf. Section 5).

4.2. Cross-Disciplinary Adoption of SHMARQL

The platform has already seen a broad uptake across diverse research initiatives, with active deployments in projects including NFDI4Memory[56], RADAR[57], and NFDI-Matwerk[58], while generating substantial interest on integrating the system from additional projects such as Graceful 17[59]. This dual-mode capability makes SHMARQL particularly valuable for organizations that want to enhance their existing semantic web infrastructure with improved user interfaces and data storytelling capabilities, while also serving as a complete out-of-the-box solution for new projects. The platform’s architecture ensures that, regardless of the underlying triplestore configuration, users benefit from consistent full-text search integration, custom styling options, and the integrated data stories functionality, making it suitable for both rapid prototyping and production deployment scenarios.

[49] https://developer.mozilla.org/en-US/docs/Web/HTTP/Guides/CORS
[50] https://github.com/ISE-FIZKarlsruhe/fizzysearch
[51] https://github.com/ISE-FIZKarlsruhe/bikidata
[52] https://tree-sitter.github.io/
[53] http://rdf2vec.org/
[54] https://sbert.net/
[55] https://www.mkdocs.org/ and specifically the https://squidfunk.github.io/mkdocs-material/ theme
[56] https://4memory.de/
[57] https://www.radar-service.eu/radar/en/home
[58] https://nfdi-matwerk.de/
[59] https://graceful17.hypotheses.org/

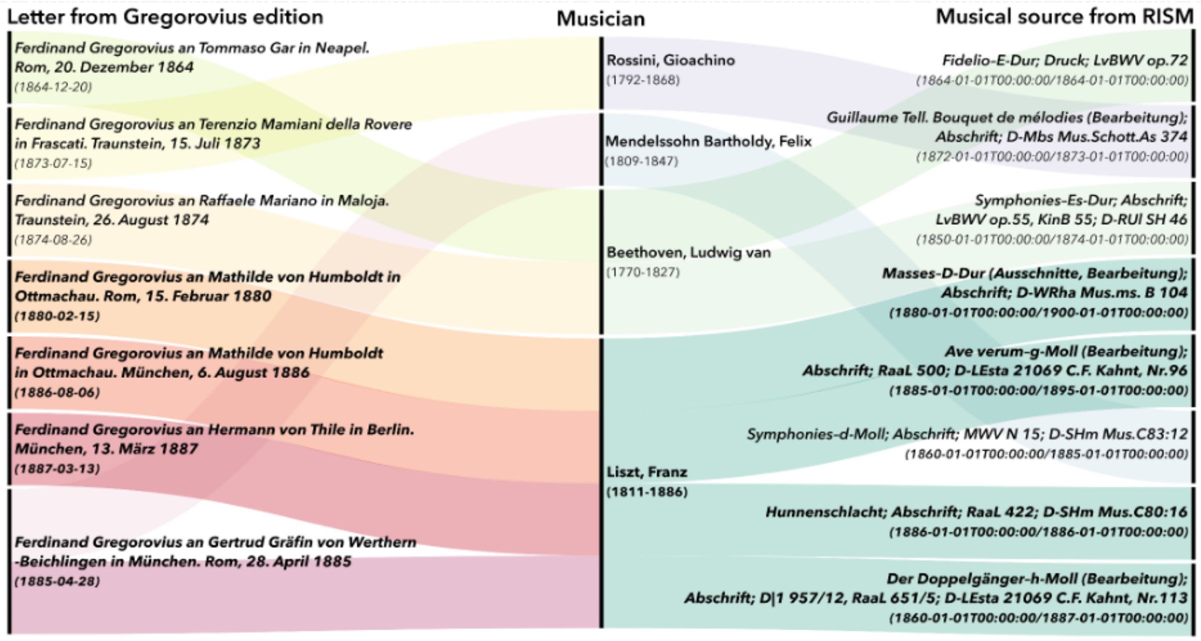


Figure 5: Interconnections of persons in the Gregorovius letter edition and RISM Online

4.3. Effective Dissemination and Discovery of Cultural Heritage Resources

The platform achieves the goals expressed in RQ3 by reducing the steps required to publish RDF-modeled data and lowering the cognitive load needed to configure features that support discovery and analysis tools. With a single Docker container command, a complete Linked Data environment becomes operational, providing integrated documentation, queries, analyses, and visualizations. Feedback from end-users who have deployed the SHMARQL platform has been consistently positive, with users expressing satisfaction regarding the platform's ease of use and flexibility for data exploration.

The SHMARQL development team welcomes community contributions and is actively working to add new features that will extend and enhance the system during the next phase of NFDI4Culture consortium activities.

5. Use Cases of CTO and SHMARQL in the NFDI4Culture Portal

The application and the benefit of the presented contributions within the infrastructure of NFDI4Culture can be illustrated on the example of the digital letter edition *Ferdinand Gregorovius: Poesie und Wissenschaft. Gesammelte deutsche und italienische Briefe*[60] and their interconnections to musicological data from RISM Online[61]. The Gregorovius edition contains 1,093 annotated pieces of correspondence from the historian Ferdinand Gregorovius, whose heritage testifies a rich engagement with intellectual-historical movements and musicians of his time. Although the data from the Gregorovius edition and RISM Online are related through shared content and authority data, no common access point existed prior to their integration into the NFDI4Culture-KG. This fragmentation prevented researchers from performing unified queries, conducting cross-dataset searches, or discovering meaningful interconnections essential for comprehensive research and data reuse. Figure 5 showcases connections on the level of persons between both data sets, which were revealed through a SPARQL query[62] [28].

Furthermore, SHMARQL and CTO are employed in the data stories platform provided by NFDI4Cul-

[60] https://gregorovius-edition.dhi-roma.it/

[61] https://rism.info/

[62] https://nfdi4culture.de/go/kg-gregorovius-rism-example-musical-sources-letters

ture[63], illustrated in Figure 7 in Appendix A. This application combines semantic technologies with storytelling to promote data literacy and showcase the potential of LOD in the CH domain. Guided by narrative structures and the integration of visualizations, and interactive elements such as SPARQL queries, the data stories enable innovative exploration modes for semantically connected CH datasets within the NFDI4Culture-KG. A practical example is the data story *An Italian Data Journey* [29], developed within NFDI4Culture. It explores the 18th-century opera holdings of the Doria Pamphilj Archive in Rome using data from the Partitura project at the German Historical Institute in Rome and illustrates how historical musicological data can be enriched with external resources such as Wikidata[64], GeoNames[65], ICONCLASS[66] and RISM. By integrating the dataset into the NFDI4Culture-KG and leveraging computational resources from the European Open Science Cloud (EOSC)[67], the data story demonstrates how federated infrastructures can improve the interoperability and reusability of CH research data [30].

6. Conclusion

This paper contributed the NFDI4Culture Ontology (CTO) and the SHMARQL Linked Data platform as integral components for the semantic representation and discovery of CH research data. By aligning with BFO and extending the mid-level NFDIcore ontology, CTO provides a lightweight yet expressive framework for integrating decentralized CH metadata into a centralized, queryable infrastructure. SHMARQL complements this by enabling intuitive exploration and publication of Linked Data, enhanced with full-text search and narrative-based interfaces through its Data Stories feature. The contributions demonstrate how CTO and SHMARQL jointly address the challenges of cross-domain interoperability, semantic richness, and accessible knowledge graph interaction. Their adoption within and beyond NFDI4Culture, ranging from musicology and digital editions to interdisciplinary data stories, shows their potential for reuse across diverse research domains.

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

Declaration on Generative AI: The authors used GPT-4 and Claude Sonnet 4 for grammar and spelling check. The authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. Y. Jaradeh, A. Oelen, K. E. Farfar, M. Prinz, J. D'Souza, G. Kismihók, M. Stocker, S. Auer, Open Research Knowledge Graph: Next Generation Infrastructure for Semantic Scholarly Knowledge, in: Proceedings of the 10th International Conference on Knowledge Capture, 2019, pp. 243–246.

[2] O. Karras, J. Göpfert, P. Kuckertz, T. Pelser, S. Auer, Organizing Scientific Knowledge From Energy System Research Using the Open Research Knowledge Graph, arXiv preprint arXiv:2401.13365 (2024).

[3] T. Tietz, et al., NFDI4Culture Ontology (CTO), 2025. URL: https://nfdi4culture.de/ontology/3.0.0.

[4] J. Waitelonis, et al., NFDIcore Ontology, 2025. URL: https://nfdi.fiz-karlsruhe.de/ontology/3.0.1.

[5] R. Arp, B. Smith, A. D. Spear, Building ontologies with Basic Formal Ontology, MIT Press, 2015.

[6] J. N. Otte, J. Beverley, A. Ruttenberg, BFO: Basic Formal Ontology, Applied Ontology 17 (2022) 17–43.

[7] R. Arp, B. Smith, A. D. Spear, Building ontologies with Basic Formal Ontology, MIT Press, 2015.

[63] https://datastories.nfdi4culture.de/

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

[65] https://www.geonames.org/

[66] https://iconclass.org/

[67] https://open-science-cloud.ec.europa.eu/

[8] S. Borgo, R. Ferrario, A. Gangemi, N. Guarino, C. Masolo, D. Porello, E. M. Sanfilippo, L. Vieu, Dolce: A descriptive ontology for linguistic and cognitive engineering, Applied ontology 17 (2022) 45–69.

[9] J. Corson-Rikert, S. Mitchell, B. Lowe, N. Rejack, Y. Ding, C. Guo, The VIVO ontology, in: VIVO: A Semantic Approach to Scholarly Networking and Discovery, Springer, 2012, pp. 15–33.

[10] B. Inostroza, R. Cid, A. Hogan, Rdf playground: an online tool for learning about the semantic web, in: Companion Proceedings of the ACM Web Conference 2023, 2023, pp. 111–114.

[11] R. Verborgh, M. Vander Sande, O. Hartig, J. Van Herwegen, L. De Vocht, B. De Meester, G. Haesendonck, P. Colpaert, Triple pattern fragments: a low-cost knowledge graph interface for the web, Journal of Web Semantics 37 (2016) 184–206.

[12] D. V. Camarda, S. Mazzini, A. Antonuccio, Lodlive, exploring the web of data, in: Proceedings of the 8th International Conference on Semantic Systems, 2012, pp. 197–200.

[13] A. Micsik, S. Turbucz, Z. Tóth, Exploring publication metadata graphs with the lodmilla browser and editor, International Journal on Digital Libraries 16 (2015) 15–24.

[14] O. Bruns, T. Tietz, L. Söhn, J. J. Steller, S. R. Ondraszek, E. Posthumus, T. Schrade, H. Sack, 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.

[15] O. Bruns, T. Tietz, L. Söhn, J. J. Steller, E. Poshumus, T. Schrade, H. Sack, Gotta Catch'em All: From Data Silos to a Knowledge Graph, in: Proceedings of 21st Extended Semantic Web Conference, ESWC 2024, Poster & Demos, 2024.

[16] J. J. Steller, L. C. Söhn, J. Tolksdorf, O. Bruns, T. Tietz, E. Posthumus, H. Fliegl, S. Pittroff, H. Sack, T. Schrade, Communities, Harvesting, and CGIF: Building the Research Data Graph at NFDI4Culture, in: DHd 2024 Quo Vadis DH (DHd2024), Passau, Deutschland, 2024. doi:10.5281/zenodo.10698301.

[17] T. Tietz, O. Bruns, L. Söhn, J. Tolksdorf, E. Posthumus, J. J. Steller, H. Fliegl, E. Norouzi, J. Waitelonis, T. Schrade, H. Sack, 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.

[18] S. Arabandi, et al., Information artefact ontology (iao), http://purl.obolibrary.org/obo/iao.owl, 2024. Accessed: 2025-04-27.

[19] J. Malone, et al., The Software Ontology (SWO): a Resource for Reproducibility in Biomedical Data Analysis, Curation and Digital Preservation, Journal of Biomedical Semantics 5 (2014) 25. URL: https://doi.org/10.1186/2041-1480-5-25. doi:10.1186/2041-1480-5-25.

[20] M. Black, et al., EDAM: the Bioscientific Data Analysis Ontology, Poster presented at F1000Research, 11 (ISCB Comm J): 1, 2022. doi:10.7490/f1000research.1118900.1, version 1; not peer-reviewed. Open access.

[21] O. Bruns, T. Tietz, J. Waitelonis, E. Posthumus, H. Sack, Nfdicore 2.0: A bfo-compliant ontology for multi-domain research infrastructures, 2024. URL: https://arxiv.org/abs/2410.01821. arXiv:2410.01821.

[22] H. B. Nasrabadi, J. Waitelonis, E. Norouzi, K. Hubaiev, H. Sack, Nfdi matwerk ontology (mwo), https://github.com/ISE-FIZKarlsruhe/mwo, 2024. Revision: v3.0.0, accessed 2025-07-30.

[23] S. Ondraszek, et al., NFDI4Memory Ontology (MemO), 2024. URL: https://nfdi.fiz-karlsruhe.de/4memory/ontology/.

[24] G. A. Gesese, J. Waitelonis, Z. Chen, S. Schimmler, H. Sack, NFDI4DSO: Towards a BFO Compliant Ontology for Data Science, in: Proceedings of the 20th International Conference on Semantic Systems (SEMANTICS 2024), 2024.

[25] L. D. Couprie, Iconclass: an iconographic classification system, Art libraries journal 8 (1983) 32–49.

[26] N. Matentzoglu, D. Goutte-Gattat, S. Z. K. Tan, J. P. Balhoff, S. Carbon, A. R. Caron, W. D. Duncan, J. E. Flack, M. Haendel, N. L. Harris, et al., Ontology Development Kit: a Toolkit for Building, Maintaining and Standardizing Biomedical Ontologies, Database 2022 (2022) baac087.

[27] S. Peroni, D. Shotton, F. Vitali, The live owl documentation environment: a tool for the automatic generation of ontology documentation, in: International Conference on Knowledge Engineering and Knowledge Management, Springer, 2012, pp. 398–412.

[28] L. C. Söhn, T. Tietz, J. J. Steller, P. Kehrein, A. Büttner, O. Bruns, E. Posthumus, J.-P. Grünewälder, J. Hörnschemeyer, C. Sander, V. Grund, H. Fliegl, H. Sack, T. Schrade, Nfdi4culture integration stories: Bridging gaps between isolated research resources, in: Digital Humanities 2025: Accessibility & Citizenship (DH2025), Zenodo, Lisbon, Portugal, 2025. URL: https://doi.org/10.5281/zenodo.16570076. doi:10.5281/zenodo.16570076.

[29] T. Schrade, An Italian Data Journey: Analysing Research Data about 18th-century Italian Opera Using the Culture Knowledge Graph and Federated European Research Infrastructures, NFDI4Culture Data Story (Story ID E6263), 2025. URL: https://datastories.nfdi4culture.de/story/E6263, persistent Identifier: https://nfdi4culture.de/id/E6263.

[30] L. C. Söhn, T. Tietz, T. Schrade, J. J. Steller, A. Büttner, O. Bruns, E. Posthumus, H. Fliegl, H. Sack, Telling Data Stories: Linking Infrastructure, Semantics and Cultural Heritage Data in NFDI4Culture, 2025. Manuscript under review at DHd2026.

A. Addidional Visualizations

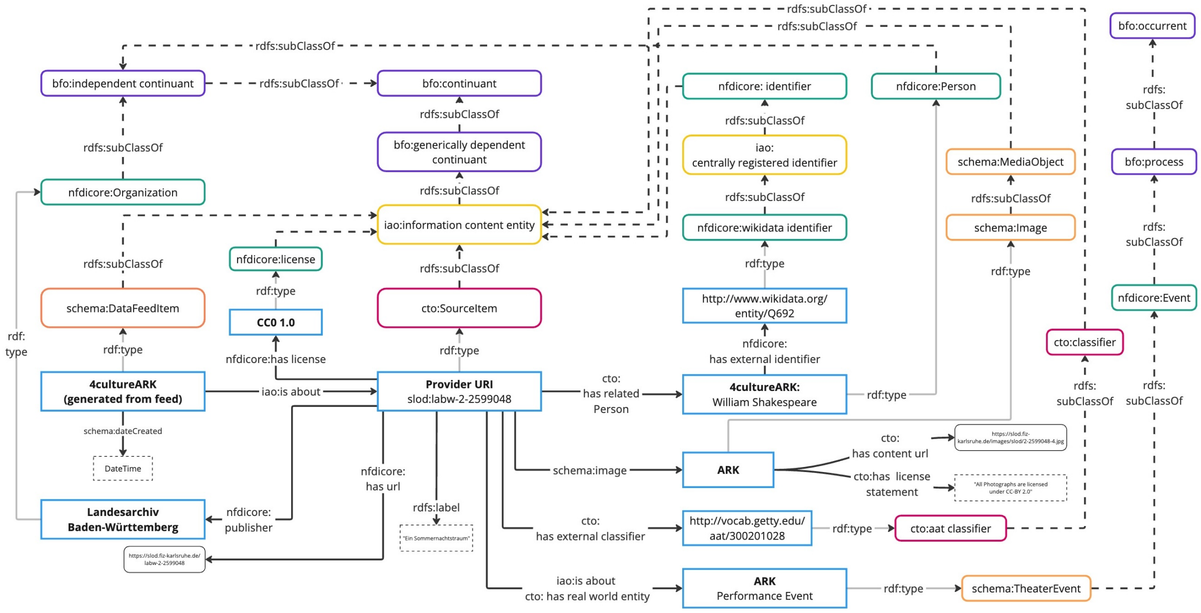
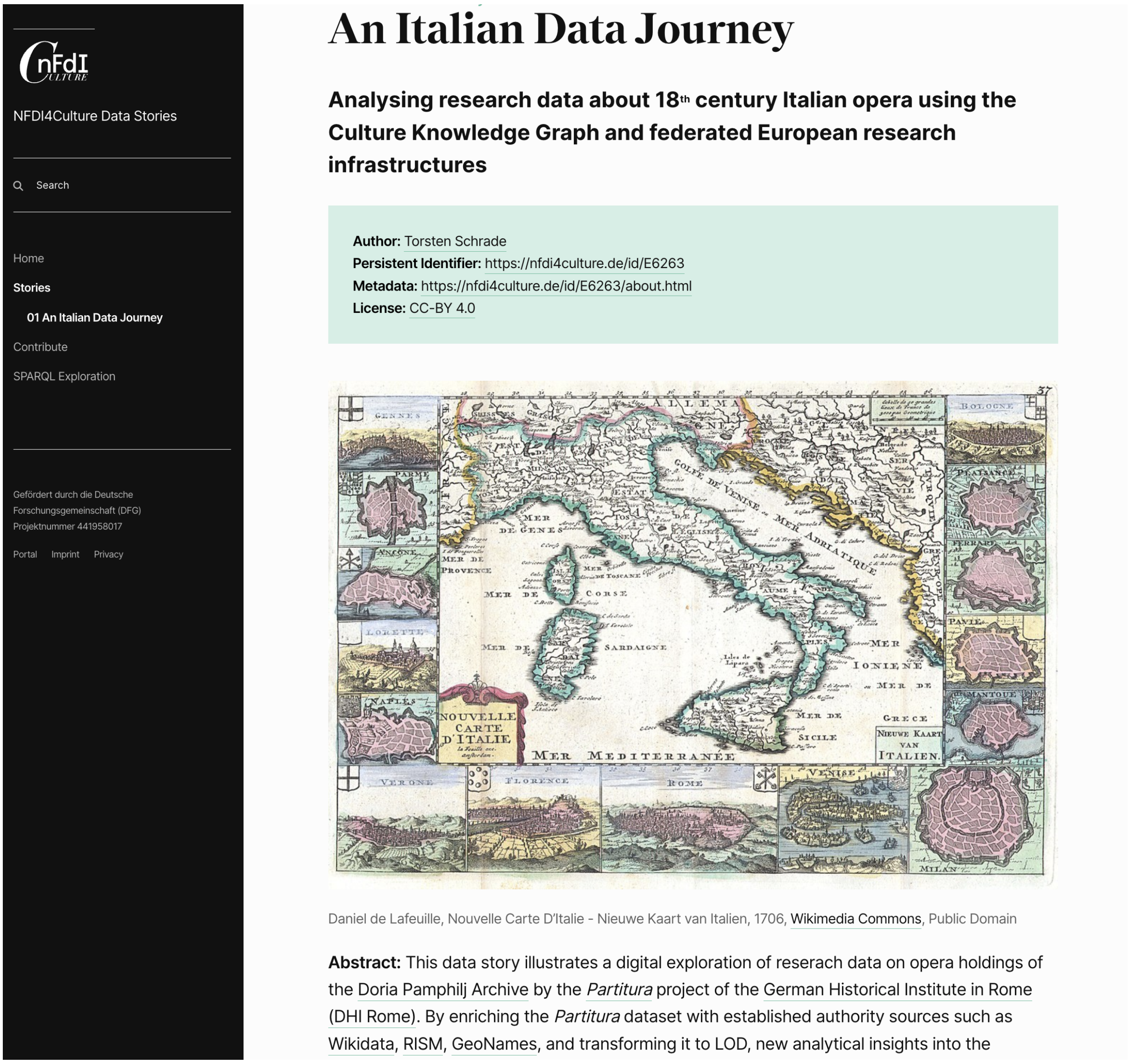


Figure 6: Modeling example of the NFDI4Culture Ontology (CTO) in the performing arts domain



NFDI4Culture Data Stories

Search

Home

Stories

01 An Italian Data Journey

Contribute

SPARQL Exploration

Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) Projektnummer 441958017

Portal Imprint Privacy

An Italian Data Journey

Analysing research data about 18th century Italian opera using the Culture Knowledge Graph and federated European research infrastructures

Author: Torsten Schrade
Persistent Identifier: https://nfdi4culture.de/id/E6263
Metadata: https://nfdi4culture.de/id/E6263/about.html
License: CC-BY 4.0

Daniel de Lafeuille, Nouvelle Carte D'Italie - Nieuwe Kaart van Italien, 1706, Wikimedia Commons, Public Domain

Abstract: This data story illustrates a digital exploration of reserach data on opera holdings of the Doria Pamphilj Archive by the *Partitura* project of the German Historical Institute in Rome (DHI Rome). By enriching the *Partitura* dataset with established authority sources such as Wikidata, RISM, GeoNames, and transforming it to LOD, new analytical insights into the

Figure 7: A data story on the NFDI4Culture platform "An Italian Data Journey" by Torsten Schade