

The case for a common, reusable Knowledge Graph Infrastructure for NFDI

Lozana Rossenova¹, Moritz Schubotz² and Renat Shigapov³
14.09.2023, CoRDI

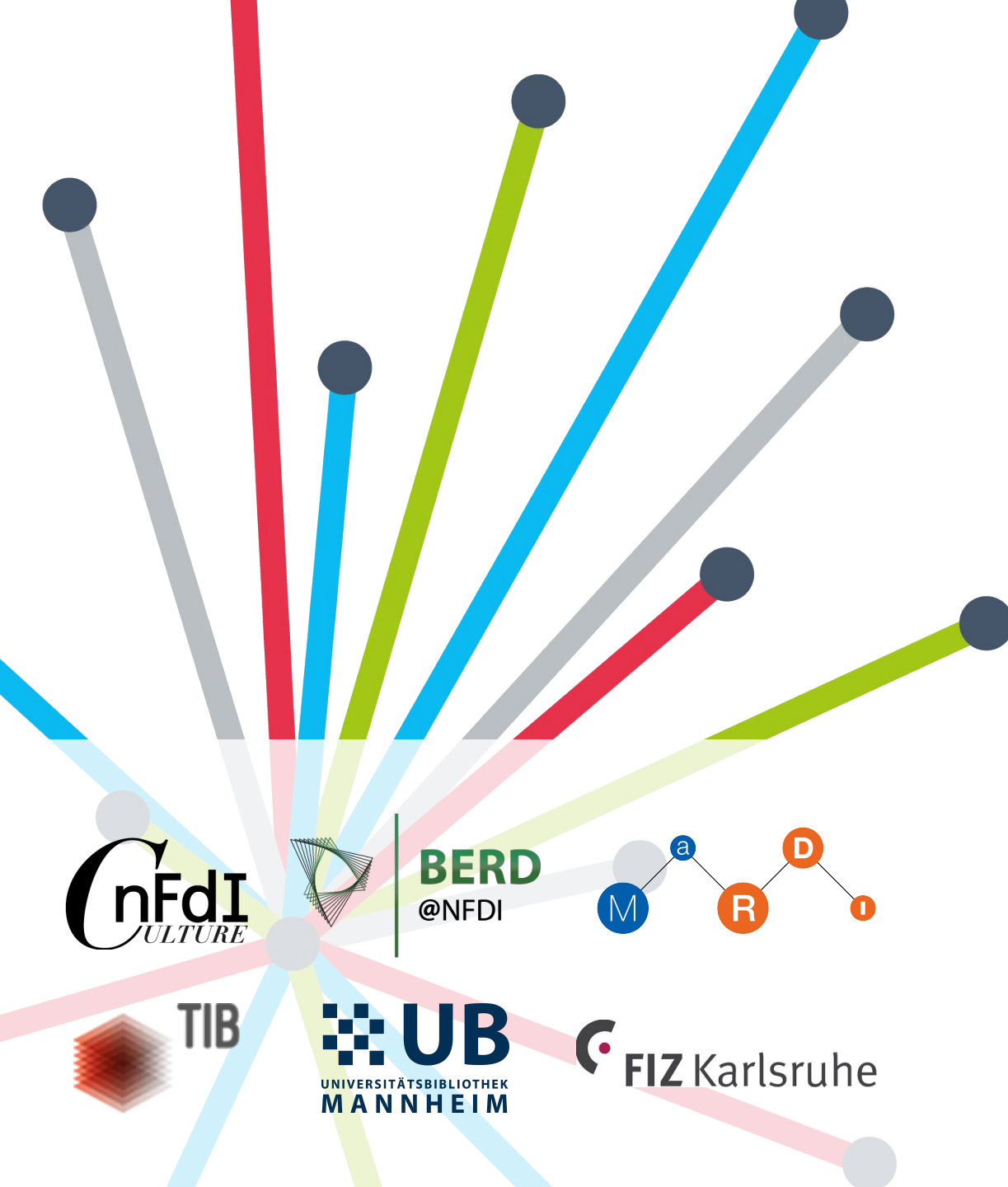
¹ TIB – Leibniz Information Centre for Science and Technology, Hannover;

² FIZ Karlsruhe - Leibniz Institute for Information Infrastructure, Berlin

³ University Library | University of Mannheim, Mannheim



This work is licensed under a [Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License](https://creativecommons.org/licenses/by-nc-nd/4.0/).



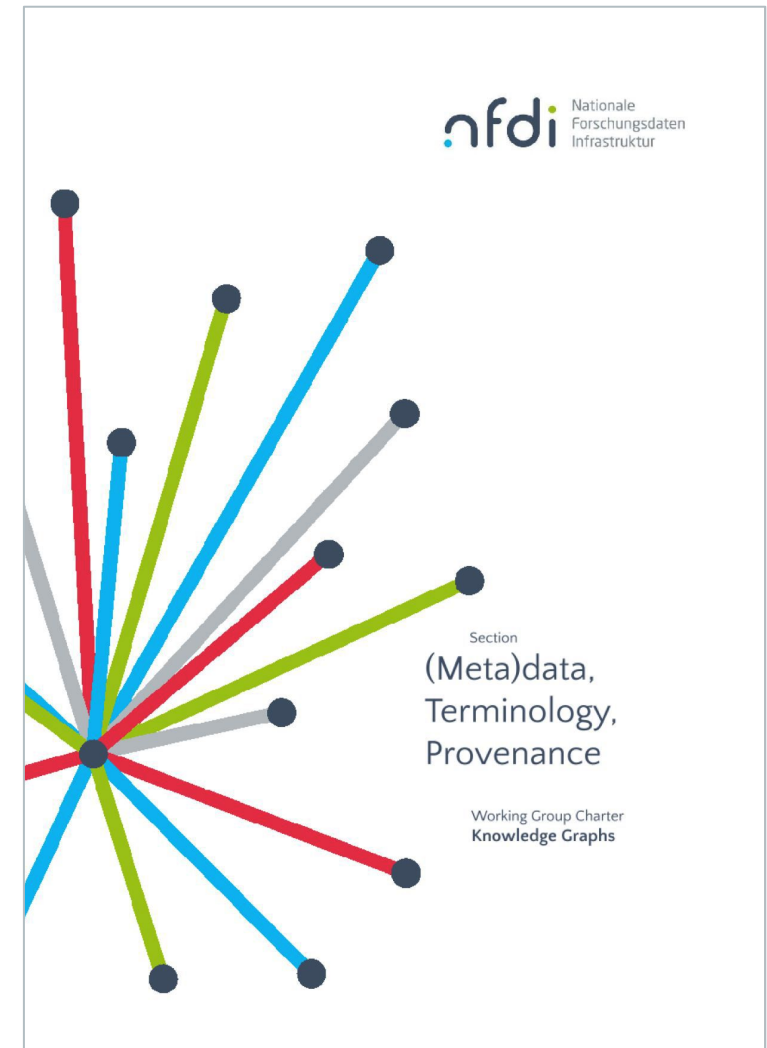
The Working Group “Knowledge Graphs” (KGs) in NFDI Section “(Meta)data, Terminologies, Provenance”

Motivation:

- Promoting the use of knowledge graphs by consortia, institutions and researchers;
- Improving FAIRness of NFDI and especially interoperability with national and international research data infrastructures;
- Contributing to development of KG tools and services.

Numbers

- 96 subscribers to the mailing list
- 56 members representing 22 consortia: the charter <https://doi.org/10.5281/zenodo.7515324>
- 3 coordinators: Renat Shigapov (BERD@NFDI), Lozana Rossenova (NFDI4Culture) & Moritz Schubotz (MaRDI)



Why KGs and why KGI?

Why KGs are an important technology for building an **interoperability framework** and enabling **data exchange**, as understood by our WG:

- KG is a **graph-structured knowledge base** containing a terminology (vocabulary or ontology) and data entities interrelated via the terminology;
- KGs are based on **semantic web technologies** (RDF, SPARQL, etc.) and often used for agile data integration;
- KGs are already **widely used** by research data producers and managers in Germany ([see poster](#));
- **Wikidata** as special connector linking between expert knowledge systems and world knowledge.

Invited talks:

1. **PID Graph & GraphQL** – Markus Stocker
2. **GESIS Search & KGI** – Benjamin Zapilko and Stefan Dietze
3. **Piveau & Data Europa** – Sonja Schimmler & Bianca Wentzel
4. **NFDI4DS Search at Uni Hamburg** – R. Usbeck, T. Taffa and A. Kraft
5. **OpenAIRE Research Graph** – Andreas Czerniak

Why KGs and why KGI?

Humanities and social sciences

- BERD@NFDI (KGs)
- KonsortSWD
- NFDI4Culture (KGs)
- NFDI4Memory (KGs)
- NFDI4Objects
- Text+ (KG)

Engineering sciences

- NFDI4DataScience (KGs & KG Software)
- NFDI4Energy (KG)
- NFDI4Ing (KG Software)
- NFDI-MatWerk (KGs & KG Software)
- NFDIxCS

Life sciences

- DataPLANT
- FAIRagro
- NFDI4Immuno
- GHGA
- NFDI4Biodiversity
- NFDI4BIOIMAGE (KG)
- NFDI4Health
- NFDI4Microbiota (KG)

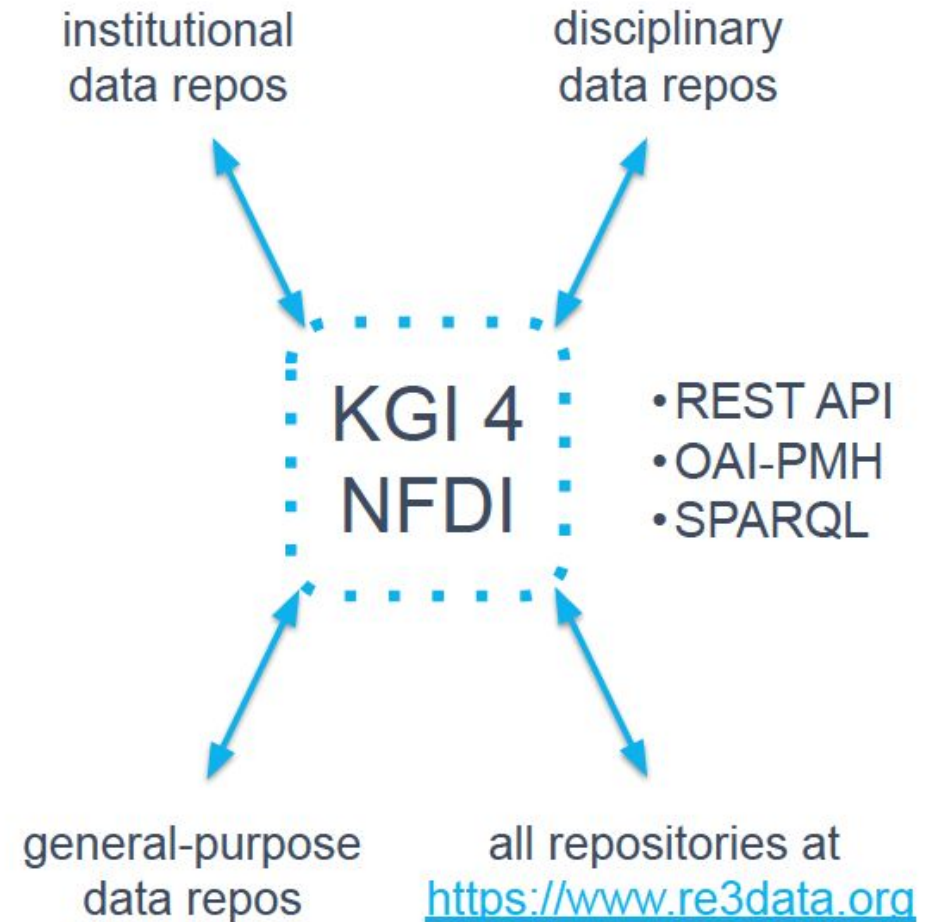
Natural sciences

- DAPHNE4NFDI
- FAIRmat
- NFDI4Cat (KG)
- MaRDI (KGs)
- NFDI4Chem (KGs)
- NFDI4Earth (KG)
- PUNCH4NFDI

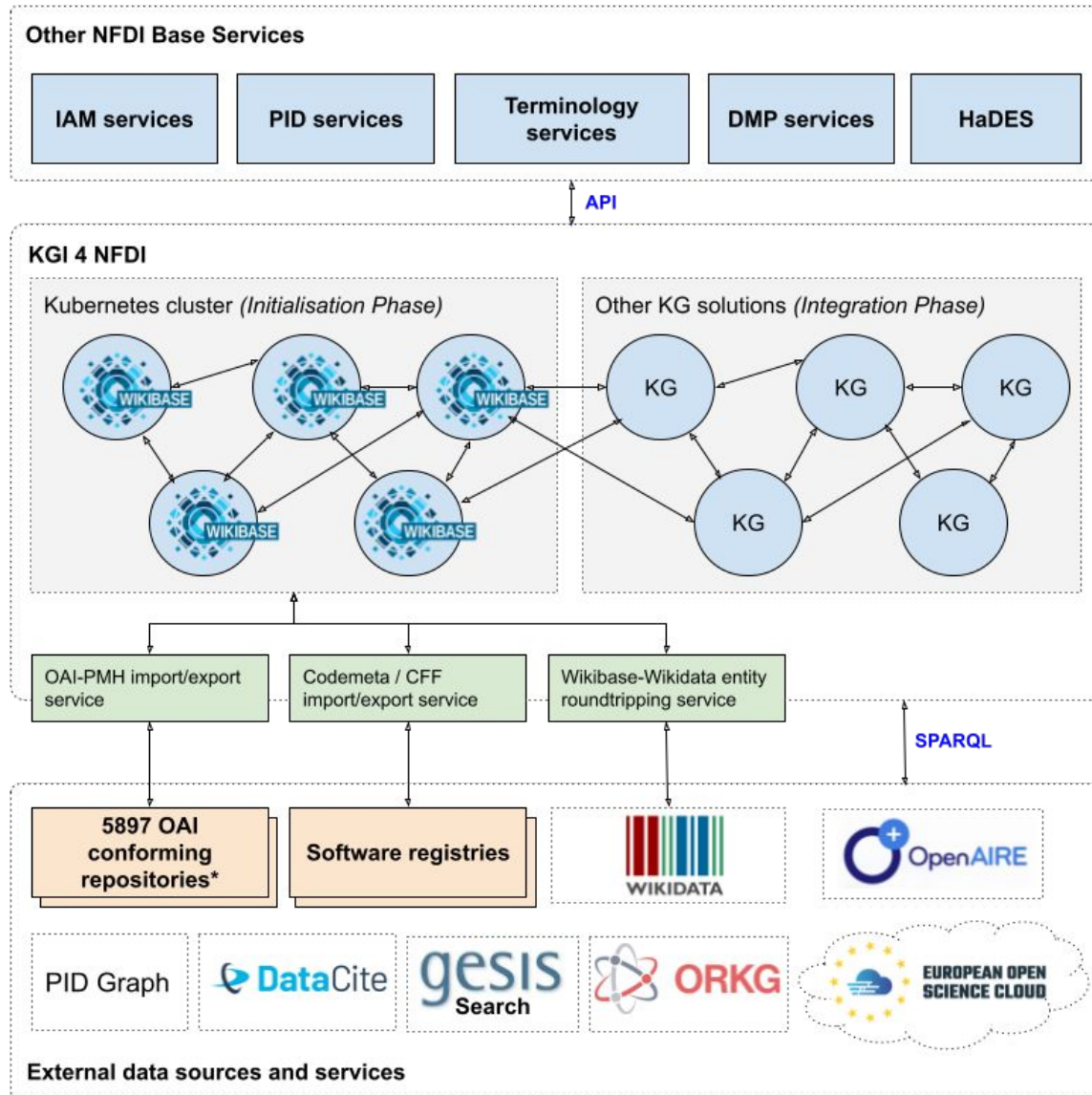
Why KGs and why KGI?

NFDI needs to be **interoperable internally and across national and international research data infrastructures** (as the section WGs testify):

- Individual solutions may be required to meet **domain-specific** requirements;
- NFDI needs an **interoperable network** of metadata knowledge graphs (RDF, SPARQL);
- Consortia, institutions and researchers need an easy-to-use, scalable and interoperable **KGI-as-a-Service**.



KGI-as-a-Service proposal



...an ecosystem of software, including tools for data import, validation and export, collaborative frontends, search APIs and SPARQL endpoints with result visualization, Extract-Transform-Load and data linking software...

* Source: <https://www.openarchives.org/Register/BrowseSites>

Original proposal to Base4NFDI

Proposal submitted on 15.02.2023:

- Combining the **ease-of-use** of software like Wikidata with research-backed data;
- Allowing NFDI stakeholders to **create KGs** without administrative overhead;
- Developing an **interoperability framework** for connecting KGs with research infrastructures;
- And establishing a **KGI-consultancy** to increase adoption of the KGI-service.

Pilot phase based on one specific tool suite as a **'minimum viable product'** (Wikibase):

- Landscape analysis (learning the needs of consortia and researchers; overview papers);
- Deployment scalability (Kubernetes cluster);
- Interoperability pipelines (OAI-PMH & Codemeta / CFF import/export to SPARQL);
- Consultancy (help with creating knowledge graphs).

Choice of pilot software suite and use cases

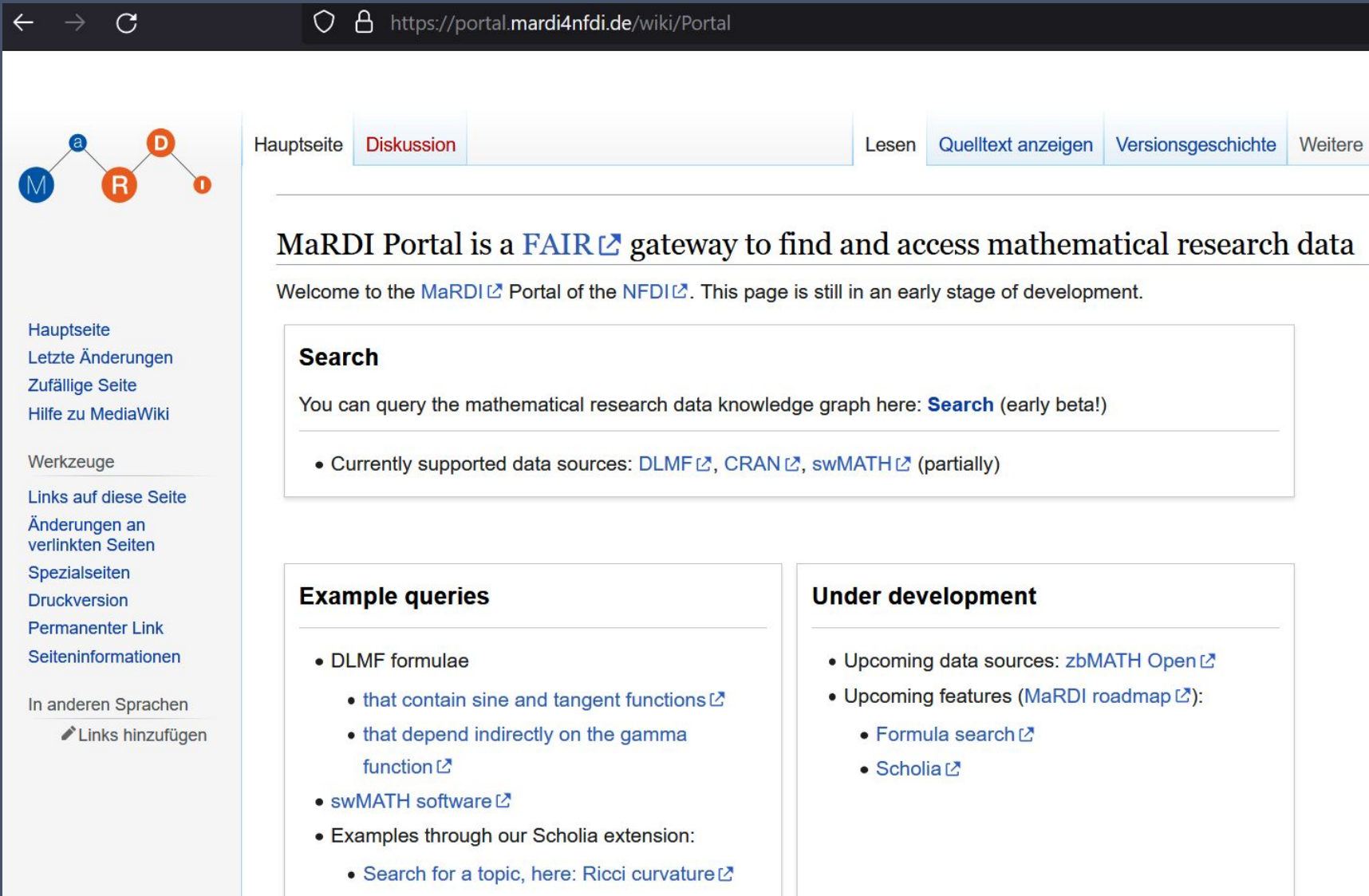
Wikibase and Wikidata adoption:

- Wikidata KG, already used by various consortia and participating institutions – both as a **repository** to upload data to, and a rich resource on the linked open data (LOD) cloud to **federate** with;
- **Growing adoption** of Wikibase and the popularity of Wikidata as proof-of-concept;
- Mix of human- and machine-readable interfaces can lower the barrier to **participation**.

Use cases:

- **MaRDI** and **BERD4NFDI** are using Wikibase instances as central portals for all research data;
- **NFDI4Culture** offer Wikibase instances to annotate digitized cultural objects with structured data;
- **NFDI4Memory** includes FactGrid, Wikibase instance hosted at the University of Erfurt, as central repository for data about historical persons and events.

MaRDI Portal and SPARQL endpoint



The screenshot shows the MaRDI Portal homepage. The browser address bar displays `https://portal.mardi4nfdi.de/wiki/Portal`. The page features a navigation menu with 'Hauptseite' and 'Diskussion' highlighted. Below the navigation, a main heading reads 'MaRDI Portal is a FAIR gateway to find and access mathematical research data'. A welcome message states: 'Welcome to the MaRDI Portal of the NFDI. This page is still in an early stage of development.' A search section includes the text 'You can query the mathematical research data knowledge graph here: Search (early beta!)' and a list of supported data sources: 'DLMF, CRAN, swMATH (partially)'. The page is divided into two columns: 'Example queries' and 'Under development'. The 'Example queries' column lists: 'DLMF formulae' (with sub-points for sine and tangent functions, and gamma function), 'swMATH software', and 'Examples through our Scholia extension' (with a link for Ricci curvature). The 'Under development' column lists: 'Upcoming data sources: zbMATH Open' and 'Upcoming features (MaRDI roadmap)' (with sub-points for Formula search and Scholia).

Hauptseite Diskussion Lesen Quelltext anzeigen Versionsgeschichte Weitere

MaRDI Portal is a FAIR gateway to find and access mathematical research data

Welcome to the MaRDI Portal of the NFDI. This page is still in an early stage of development.

Search

You can query the mathematical research data knowledge graph here: [Search](#) (early beta!)

- Currently supported data sources: [DLMF](#), [CRAN](#), [swMATH](#) (partially)

Example queries

- DLMF formulae
 - that contain sine and tangent functions
 - that depend indirectly on the gamma function
- swMATH software
- Examples through our Scholia extension:
 - Search for a topic, here: [Ricci curvature](#)

Under development

- Upcoming data sources: [zbMATH Open](#)
- Upcoming features (MaRDI roadmap):
 - Formula search
 - Scholia



The screenshot shows the MaRDIQueryService SPARQL endpoint interface. The browser address bar displays `https://query.portal.mardi4nfdi.de/#PRE`. The interface includes a 'MaRDIQueryService' header with 'Beispiele' and 'Abfragegenerator' buttons. A SPARQL query is entered in the main text area:

```
1 PREFIX wdt: <https://portal.mardi4nfdi.de/prop/direct/>
2 PREFIX wd: <https://portal.mardi4nfdi.de/entity/>
3 SELECT ?item ?d1mfid ?formula
4 WHERE {?item wdt:P4 wd:Q1750, wd:Q1754 .
5         OPTIONAL{?item wdt:P2 ?d1mfid .}
6         OPTIONAL{?item wdt:P14 ?formula .}
7     }
8 }
```

Below the query editor, a table displays the results of the query:

item	d1mfid	formula
Q <https://portal.mardi4nfdi.de/entity/Q1799>	4.14.E7	$\cot z = \frac{\cos z}{\sin z} = \frac{1}{\tan z}$

BERD@NFDI: KGs of German enterprises (Books-to-KG data integration) & reconciliation service

os://mbi.kgi.uni-mannheim.de/wiki/Main_Page 67%


MaschinenBauIndustrie

Aus MBI

Quelltext anzeigen Versionsgeschichte Diskussion

Overview

The MaschinenBauIndustrie (MBI) Knowledge Graph contains structured data from the book "Die Maschinen-Industrie im Deutschen Reich" written by Herbert Patschan in 1937. The book was scanned, OCR-ed, structured and semantified at [UB Mannheim](#). The data includes basic data for the German companies from machine industry.



Access to data

- The MBI data is accessible by both humans and machines
- The MBI data is currently not available openly
- [List of properties](#)
- [List of items](#)
- [Advanced search for items](#)
- [SPARQL endpoint](#)
- [API](#)
- [Export to CSV](#)
- Every entity can be downloaded in .json, .rdf, .ttl, .nt and .jsonld formats. An example for Daimler-Benz A.-G.: <https://mbi.kgi.uni-mannheim.de/entity/Q707.json>, <https://mbi.kgi.uni-mannheim.de/entity/Q707.rdf>, <https://mbi.kgi.uni-mannheim.de/entity/Q707.ttl>, <https://mbi.kgi.uni-mannheim.de/entity/Q707.nt> and <https://mbi.kgi.uni-mannheim.de/entity/Q707.jsonld>.

Data model

- Properties
 - [Properties with capitalized labels](#) are initial properties having datatypes "string" and used to model raw extracted data.
 - [Properties with non-capitalized labels](#) are properties having various datatypes (e.g., "item", "time") and used to model "semantified data".

127.0.0.1:3333/project?project=1988867990814 90%

OpenRefine manufacturing companies wikidata csv

Facet / Filter Undo / Redo 18 / 18 140 rows





Refresh Reset all Remove all Show all

Facet: **CompanyName: judgment** change
1 choice Sort by: name count
none 140
Facet by choice counts

Facet: **CompanyName: best candidate's score** change reset
4 — 51
 Numeric Non-numeric Blank Error
50 0 50 0

Manage Wikibase instances

Click on an item below to select it as the target Wikibase to work against. This will clear any existing schema. After switching to the target Wikibase, you should reconcile your data against it before editing the schema.

	Aktienführer https://akf.kgi.uni-mannheim.de/wiki/ Delete
	MBI active https://mbi.kgi.uni-mannheim.de/wiki/
	Wikidata Delete https://www.wikidata.org/wiki/
	Wikimedia Commons Delete https://commons.wikimedia.org/wiki/

10. <http://www.wikidata.org/entity/> search for match
Hauni Maschinenbau AG

← → ↻ https://wikibase.semantic-kompakt.de/wiki/Main_Page

CnFdI
3D DATA
ENRICHMENT

Hauptseite Diskussion

Main Page

Inhaltsverzeichnis [Verbergen]

- 1 Semantic annotation for 3D cultural artefacts
- 2 About the original case study
- 3 Adding new data in the archive
- 4 Example item pages for different types of data
- 5 Data model reference
- 6 Some example data queries
 - 6.1 Federated Queries
- 7 Indexes for quick reference

Semantic annotation for 3D cultural artefacts

A suite of tools for semantic annotation of 3D cultural artefacts is being developed by the [Open Science lab at TIB, Hannover](#). Operating within Task area 1: Digital Cultural Heritage within a knowledge graph environment, so that 3D objects' geometry, attributes and metadata are not lost. The project builds on several existing FOSS tools:

- [OpenRefine](#), a data cleaning, reconciliation and batch upload tool;
- [Wikibase](#) (the tool behind the interface you are viewing now), a suite of tools for semantic annotation of digital culture.

Links auf diese Seite
Änderungen an verlinkten Seiten
Spezialseiten
Druckversion
Permanenter Link
Seiteninformationen

In anderen Sprachen
Links hinzufügen

```
14 #Query wikidata
15 SERVICE wdgqs: {
16
17 #Find castles with renaissance architectural style
18 ?castle wdt:P31 wd:Q751876.
19 ?castle wdt:P149 wd:Q236122.
20
21 #Look for those castles in a radius of 100km around our castle
22 SERVICE wikibase:around {
23   ?castle wdt:P625 ?location .
24   bd:serviceParam wikibase:center ?coordinates .
25   bd:serviceParam wikibase:radius "100" .
26 }
27
28 #Get labels from Wikidata
29 ?castle rdfs:label ?castleLabel.
30 OPTIONAL { ?castle wdt:P18 ?image. }
31 FILTER((LANG(?castleLabel)) = "de")
32 }
33 }
```

Map 8 results in 2769 ms <> Code Download Link

NFDI4Memory: FactGrid

https://database.factgrid.de/wiki/Main_Page

Welcome!


on the FactGrid database, a project of the Gotha Research Centre operated by the data lab of the University of Erfurt. With the support of Wikimedia Germany we are using a MediaWiki with Wikidata's "wikibase" extension. Our main product are data which we collect on "items" such as these:

- Adam Weishaupt
- Paris

You will need to be logged in to see the "add statement" link with which you can add information in the form of triple based machine readable claims — which everyone can now explore with the SPARQL data mining language, at our Query Service. All FactGrid data are CC0-licensed. You can download any search in various data formats with the aim to explore FactGrid data in other software environments or visualise searches with various tools on our site.

Visit our

- FactGrid FAQ for more information on why you might love to use FactGrid as your research platform
- Help Section for further assistance



https://database.factgrid.de/query/#%23defaultView%3ATimeline%0ASELECT%3FReid

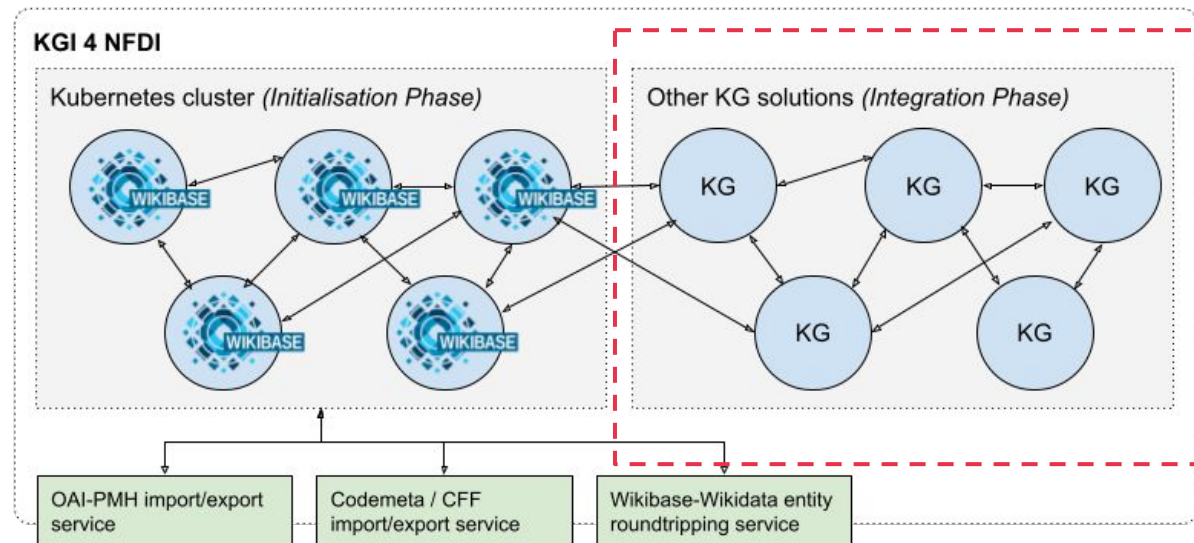


1750	1760	1770	1780	1790	1800	1810	1820
			●	●			
			12. Februar 1784				
			Adolph Freiherr Knigge zu Gast bei Christoph Bode, Arbeit an der gemeinsamen Erklärung, Weimar, 1784-02-12				
			11. Februar 1784				
			Adolph Freiherr Knigge zu Gast bei Christoph Bode's, Gespräche über Freimaurerei und Illuminaten, Weimar, 1784-02-11				
			16. Juli 1782				
			Wilhelmsbader Konvent, 15. Juli bis 1. Sept. 1782				
			1. Mai 1776				
			Gründung des Perfectibilisten-Ordens (1778 in Illuminatenorden umbenannt)				
			13. Februar 1784				
			Freiherr Knigge zu Gast bei Christoph Bode, Gedanken über einen Nachfolger der Strikten Observanz, Weimar, 1784-02-13				
			13. Februar 1784				
			Knigge und Bode Abends zu Gast bei Goethe, Über den Ursprung des Illuminatenordens, Weimar 1784-02-13				

Beyond initialisation phase

Development and operation phases:

- Extending the KGI service to non-Wikibase KGs;
- Growing adoption & support for computational methods (e.g. NLP or ML models) enabled by such an infrastructure;
- Gathering use cases of the KGI service from consortia, institutions and researchers;
- Demonstrating national and international interoperability of NFDI.



Outcome and feedback on the proposal

Unsuccessful as basic services, suggested changes:

- Include **use cases** from more consortia;
- Better explain how the **different software solutions** already in place can be integrated;
- Gather support from more consortia at **voting** stage (especially important for later funding phases);

Lessons learned:

1. Natural and life sciences have **other data workflows**, not accounted for in case studies we considered for pilot phase.
2. **Ontology and terminology service** questions need to be solved independently from concrete KG infrastructure solutions.
3. **Service-orientation** of Base4NFDI doesn't provision for implementation of one specific open source solution.

Outlook and next steps

- Reformulating the base service proposal as a dedicated DFG proposal focused on further development of Wikibase suite, matching the research and expertise of the co-applicants;
- Focus WG activities around **more cooperation with other WGs and the Sections** in order to work jointly on issues with the application of KGs independent of specific software solutions (e.g. ontologies and ontology alignment, terminology services, etc.).

More at:

<https://doi.org/10.5281/zenodo.7515324>

<https://doi.org/10.5281/zenodo.8337431>

Join us!

<https://lists.nfdi.de/postorius/lists/section-metadata-wg-kg.lists.nfdi.de>