

**BERD**  
@NFDI

Focused Tutorial on Capturing, Enriching, Disseminating Research Data Objects

# Knowledge graphs in BERD and in NFDI

Dr. Renat Shigapov

UBM - Universitätsbibliothek Mannheim

25.11.2022

## Knowledge graphs for capturing, enriching and disseminating

Knowledge graphs in BERD

Knowledge graphs in NFDI

Summary

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 460037581

**is**

flexible

graph database

RDF + SPARQL

semantic network

interlinked entities

ontology + entities

semantic

alive

## Knowledge Graph

data linking

data science

connecting data silos

**for**

data integration

semantic machine-readable data

data unification

data deduplication

data quality

data search

data disambiguation



## Capturing

- ✓ Recognising and linking entities from unstructured texts to create a new structured dataset
- ✓ Capturing data subgraphs via API, SPARQL endpoints & bulk files
- ✓ Asking complex questions and getting answers to them via SPARQL endpoints

Christian Drosten works in Charité, Germany.



PERSON

ORGANIZATION

LOCATION

Christian Drosten PERSON Q1079331 works in Charité ORG Q162684 , Germany LOC Q183 .

<https://www.wikidata.org/wiki/Q1079331>

<https://www.wikidata.org/wiki/Q183>

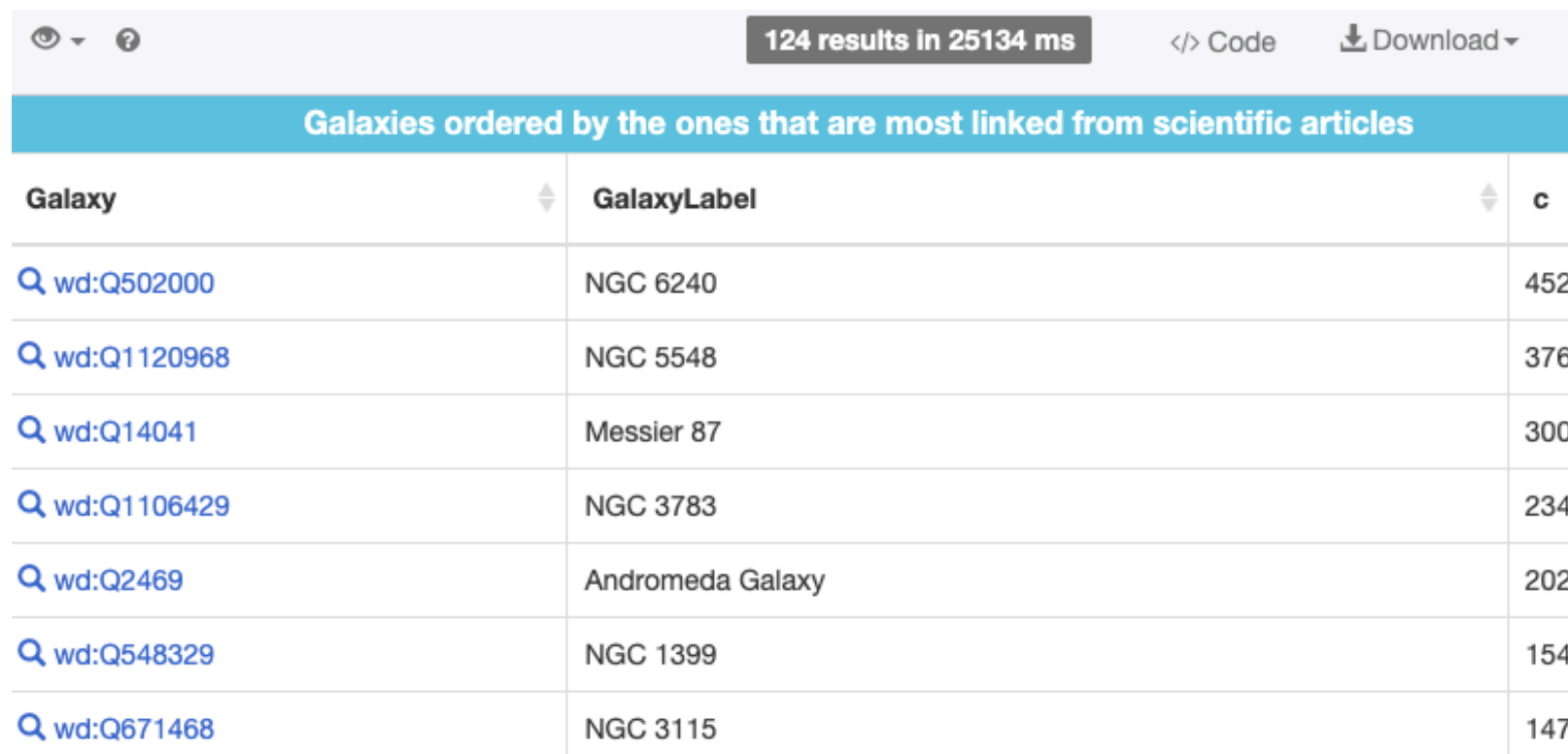
<https://www.wikidata.org/wiki/Q162684>



## Capturing

- ✓ Recognising and linking entities from unstructured texts to create a new structured dataset
- ✓ Capturing data subgraphs via API, SPARQL endpoints & bulk files
- ✓ Asking complex questions and getting answers to them via SPARQL endpoints

✓ Galaxies ordered by the ones that are most linked from scientific articles:  
[https://w.wiki/5\\$UB](https://w.wiki/5$UB)



124 results in 25134 ms

</> Code Download

Galaxies ordered by the ones that are most linked from scientific articles		
Galaxy	GalaxyLabel	c
<a href="#">wd:Q502000</a>	NGC 6240	452
<a href="#">wd:Q1120968</a>	NGC 5548	376
<a href="#">wd:Q14041</a>	Messier 87	300
<a href="#">wd:Q1106429</a>	NGC 3783	234
<a href="#">wd:Q2469</a>	Andromeda Galaxy	202
<a href="#">wd:Q548329</a>	NGC 1399	154
<a href="#">wd:Q671468</a>	NGC 3115	147

## Enriching

- ✓ **Semantic enrichment of metadata and data via named entity recognition and linking (texts & tables)**
- ✓ Data (enrichment) augmentation using a knowledge graph (e.g., for tabular data)
- ✓ Data integration of multiple datasets

Christian Drosten works in Charité, Germany.

PERSON

ORGANIZATION

LOCATION

Christian Drosten PERSON Q1079331 works in Charité ORG Q162684 , Germany LOC Q183 .

<https://www.wikidata.org/wiki/Q1079331>

<https://www.wikidata.org/wiki/Q183>

<https://www.wikidata.org/wiki/Q162684>

## Enriching

✓ Semantic enrichment of metadata and data via named entity recognition and linking (texts & tables)

✓ **Data (enrichment) augmentation using a knowledge graph (e.g., for tabular data)**

✓ Data integration of multiple datasets

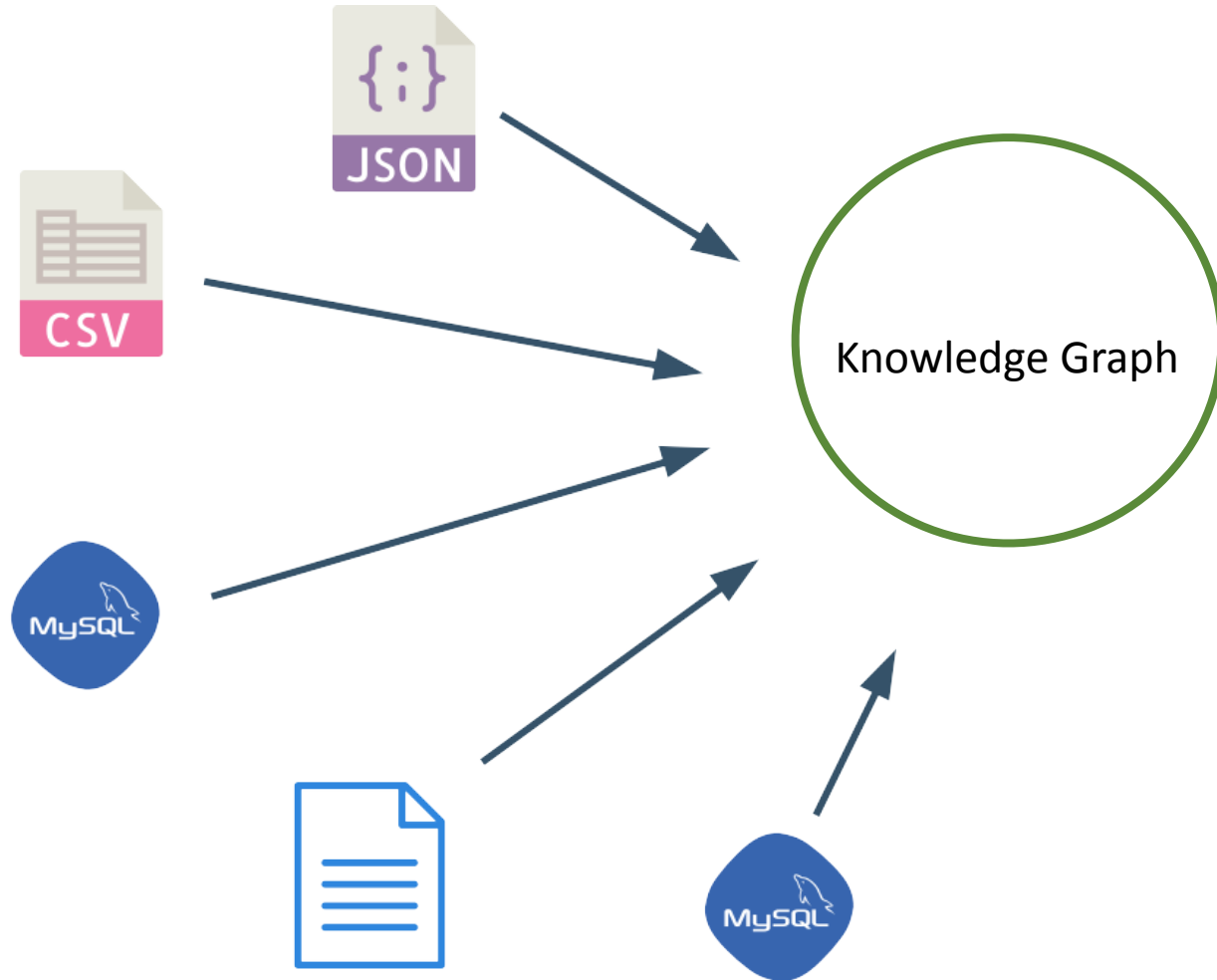
1	Deutsche Telekom	mobile phone industry	Bonn	1994-01-01
2	Lufthansa	air transport	Cologne	1953-01-01
3	SAP SE	software development	Walldorf	1972-04-01

1	<a href="#">Deutsche Telekom</a>	<a href="#">mobile phone industry</a>	<a href="#">Bonn</a>	1994-01-01	<a href="https://www.telekom.com/">https://www.telekom.com/</a>
2	<a href="#">Lufthansa</a>	<a href="#">air transport</a>	<a href="#">Cologne</a>	1953-01-01	<a href="https://www.lufthansa.com">https://www.lufthansa.com</a>
3	<a href="#">SAP SE</a>	<a href="#">software development</a>	<a href="#">Walldorf</a>	1972-04-01	<a href="https://www.sap.com">https://www.sap.com</a>
property		<a href="#">industry</a>	<a href="#">headquarters location</a>	<a href="#">inception</a>	<a href="#">official website</a>
type	<a href="#">business</a>	<a href="#">industry</a>	<a href="#">urban municipality of Germany</a>		
datatype		<a href="#">WikibaseItem</a>	<a href="#">WikibaseItem</a>	<a href="#">Time</a>	<a href="#">Url</a>



## Enriching

- ✓ Semantic enrichment of metadata and data via named entity recognition and linking (texts & tables)
- ✓ Data (enrichment) augmentation using a knowledge graph (e.g., for tabular data)
- ✓ **Data integration of multiple datasets**



## Dissemination

✓ Releasing data as a knowledge graph is enabling data science applications and data reuse



Knowledge graph = FAIR and linked data

<https://www.reading.ac.uk/research-services/research-data-management/about-research-data-management/the-research-data-lifecycle>

**Use knowledge graphs for capturing, enriching and disseminating research data objects. This is the best invitation for humans and machines to reuse your data.**

Knowledge graphs for capturing, enriching and disseminating  
**Knowledge graphs in BERD**  
Knowledge graphs in NFDI  
Summary

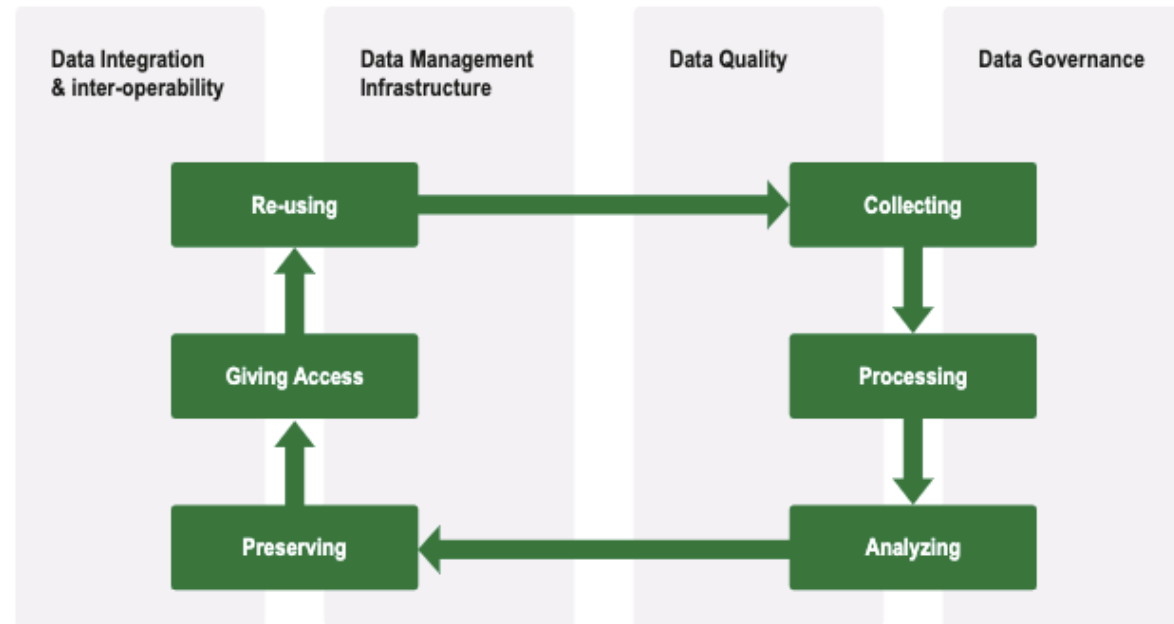
Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 460037581

<https://www.linkedin.com/school/berd-nfdi>

<https://gepris.dfg.de/gepris/projekt/460037581>

<https://www.wikidata.org/wiki/Q108542181>

## NFDI Consortium for Business, Economic and Related Data

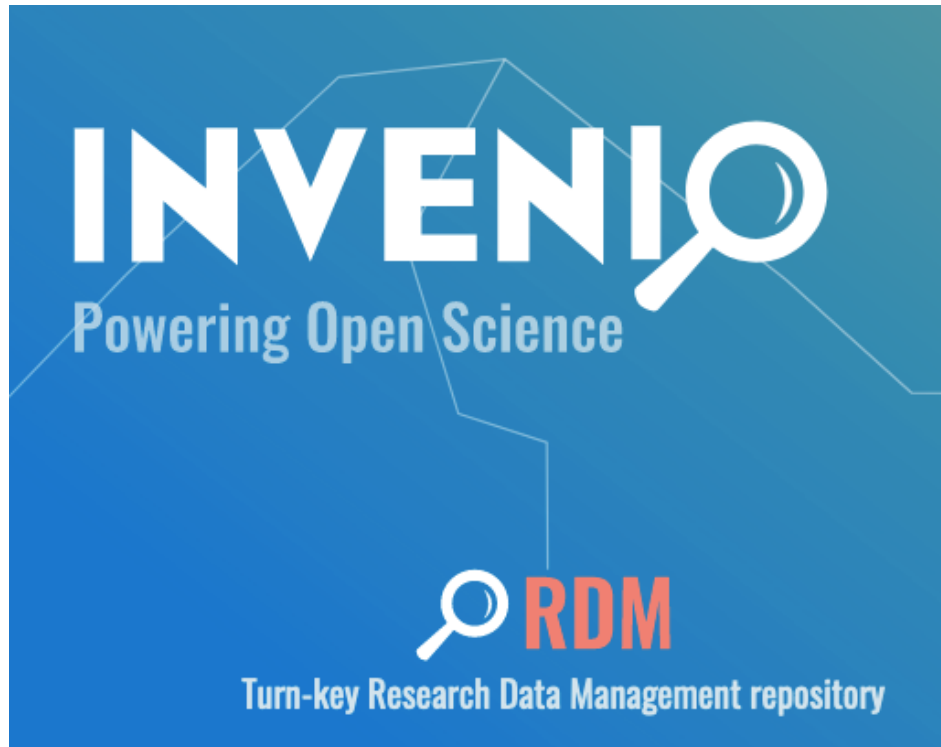


<https://www.berd-nfdi.de>

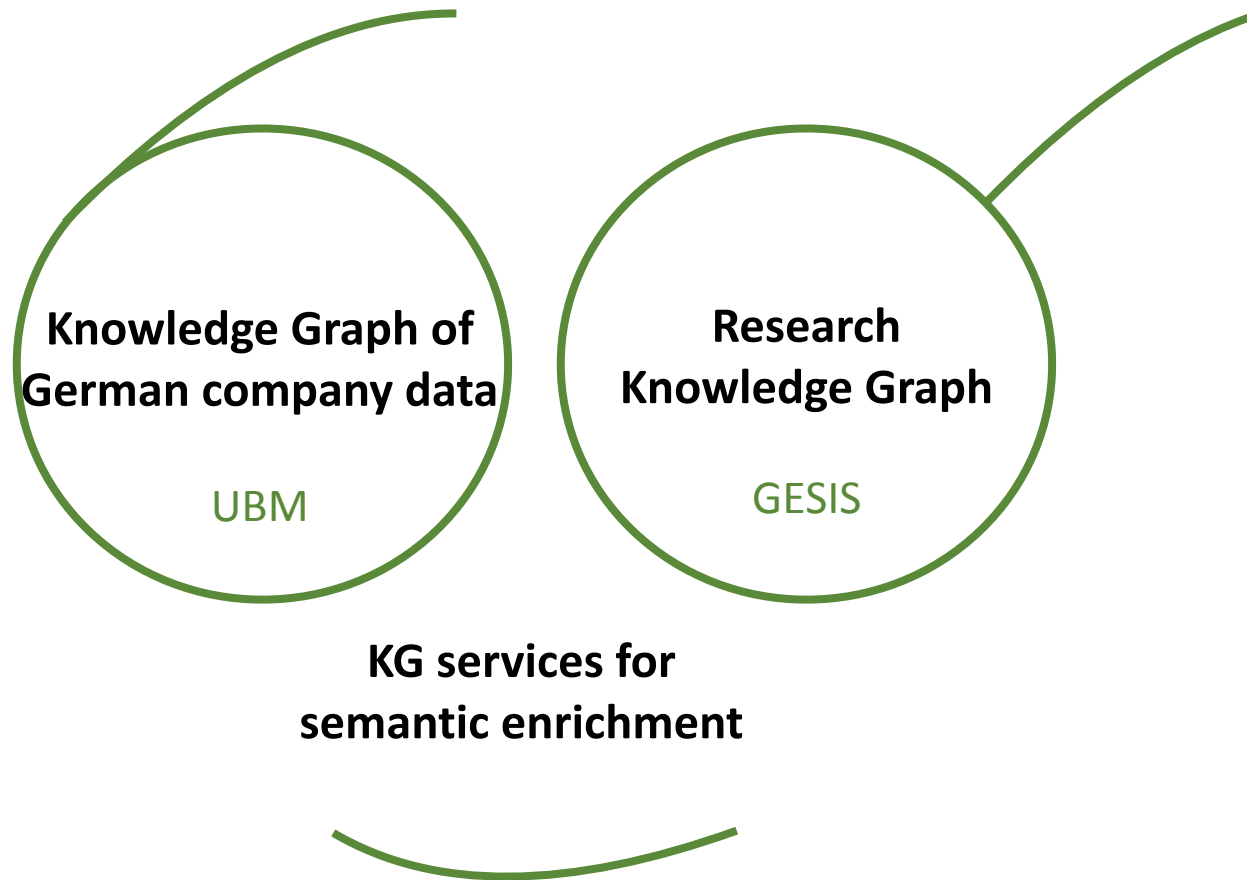
[https://twitter.com/BERD\\_NFDI](https://twitter.com/BERD_NFDI)



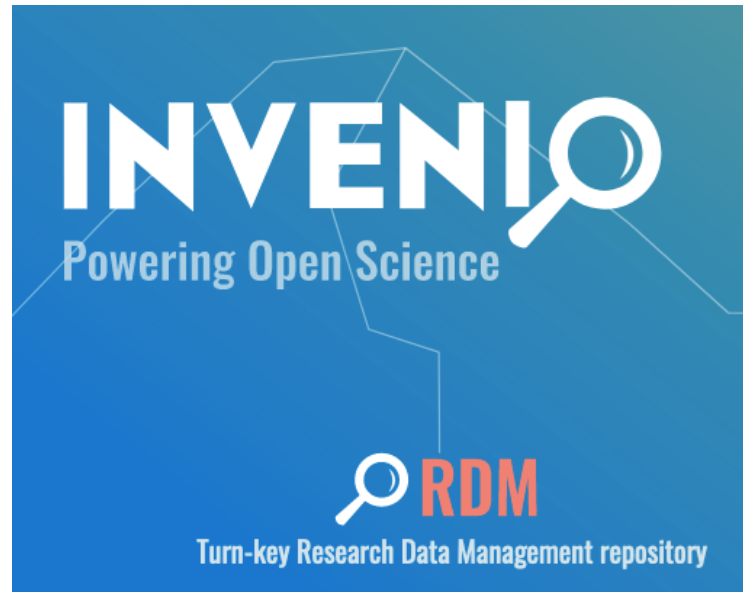
## Single point of access to BERD (meta)data



ZBW + collaboration between many partners



Open Source RDM software  
developed by CERN



<https://inveniosoftware.org>

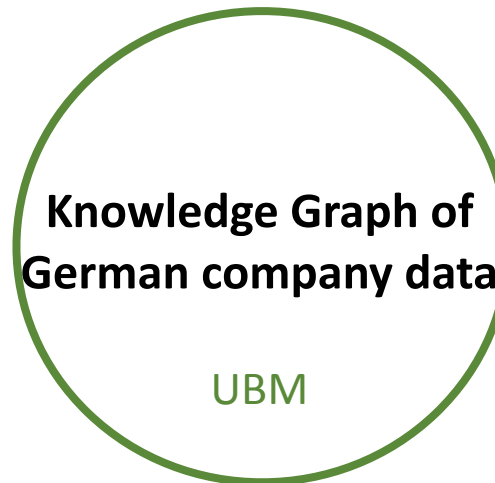
based on PostgreSQL and Elasticsearch/OpenSearch

- ✓ Python application
- ✓ **REST API**
- ✓ Export formats: JSON, Citation Style Language JSON, DataCite JSON/XML, and Dublin Core
- ✓ **OAI-PMH server**
- ✓ DataCite-based metadata
- ✓ DOI registration via DataCite
- ✓ **Scalable (an example is Zenodo)**

- ✓ Customisable faceted search
- ✓ Advanced query syntax
- ✓ Previewers of PDFs, images, CSV, Markdown, XML and JSON
- ✓ Auto-complete as you type
- ✓ Login via institutional account
- ✓ Restricted records
- ✓ Multilingual support

➔ Many distributed datasets, repositories and databases with German company data, both modern and historical

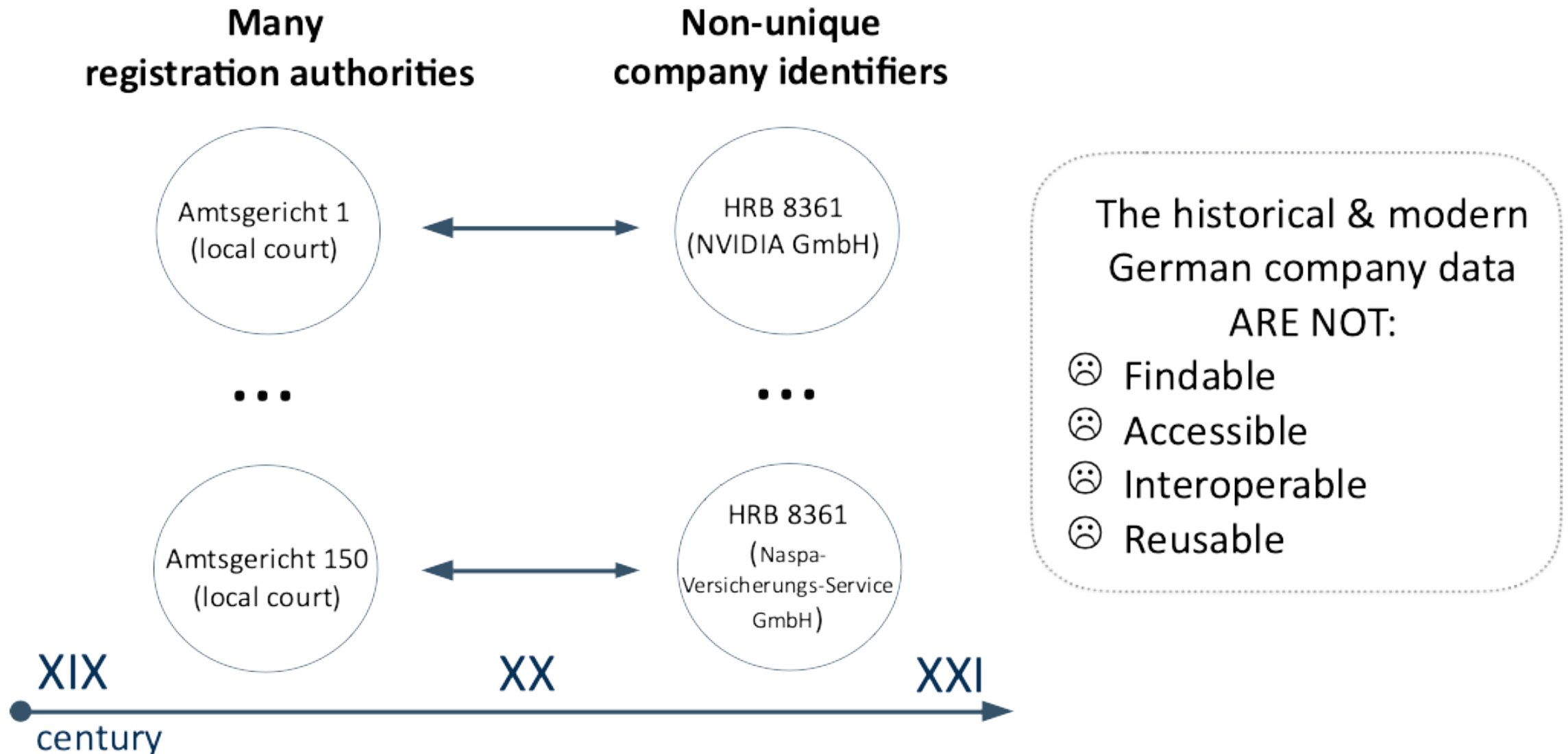
➔ There are no FAIR and linked data services for German company data



➔ Some of the data sources (reports, books and government notices) are not even digitised and OCR-ed

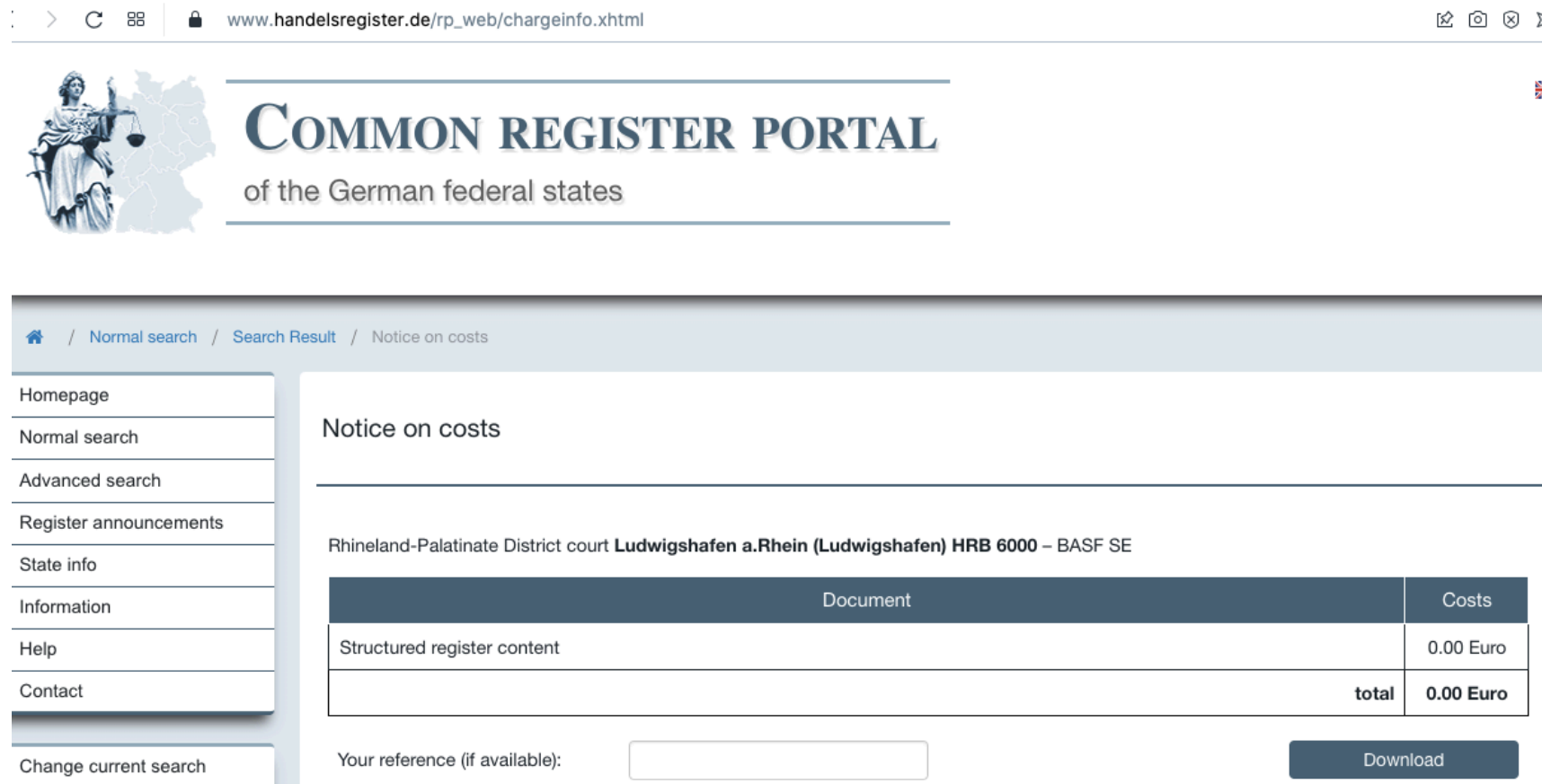
➔ Reuse of German company data is shamefully low. Applications are very limited.

# Motivation: The curse of German company data



➔ since 01.08.2022 the data is free

[https://www.handelsregister.de/rp\\_web/normalesuche.xhtml](https://www.handelsregister.de/rp_web/normalesuche.xhtml)



The screenshot shows the 'COMMON REGISTER PORTAL of the German federal states' website. The breadcrumb trail is: Home / Normal search / Search Result / Notice on costs. A sidebar on the left contains links for Homepage, Normal search, Advanced search, Register announcements, State info, Information, Help, Contact, and Change current search. The main content area is titled 'Notice on costs' and shows search results for 'Rhineland-Palatinate District court Ludwigshafen a.Rhein (Ludwigshafen) HRB 6000 – BASF SE'. A table displays the costs for document retrieval:

Document	Costs
Structured register content	0.00 Euro
<b>total</b>	<b>0.00 Euro</b>

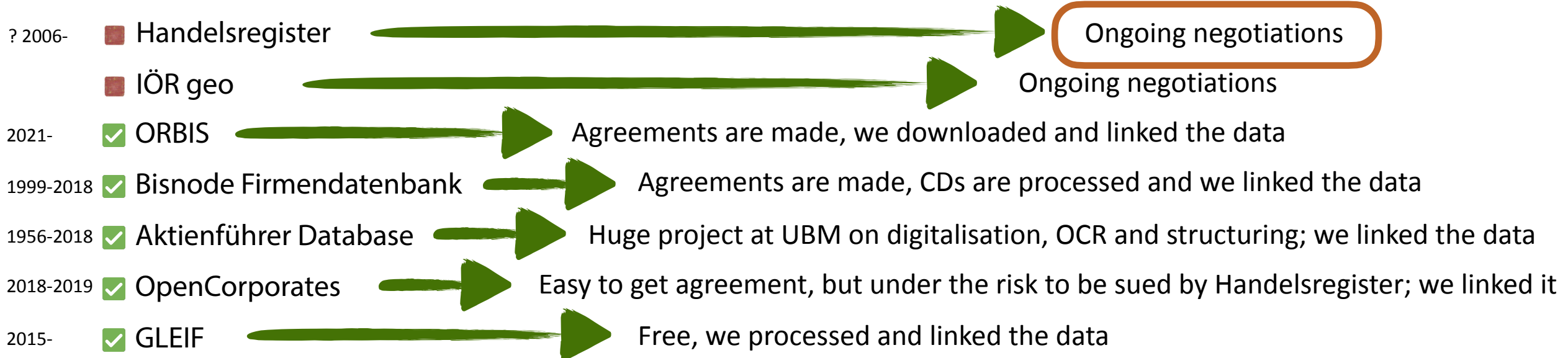
Below the table, there is a field for 'Your reference (if available):' and a 'Download' button.

➔ but there is no API and a user cannot request more than 60 legal entities per hour



## structured

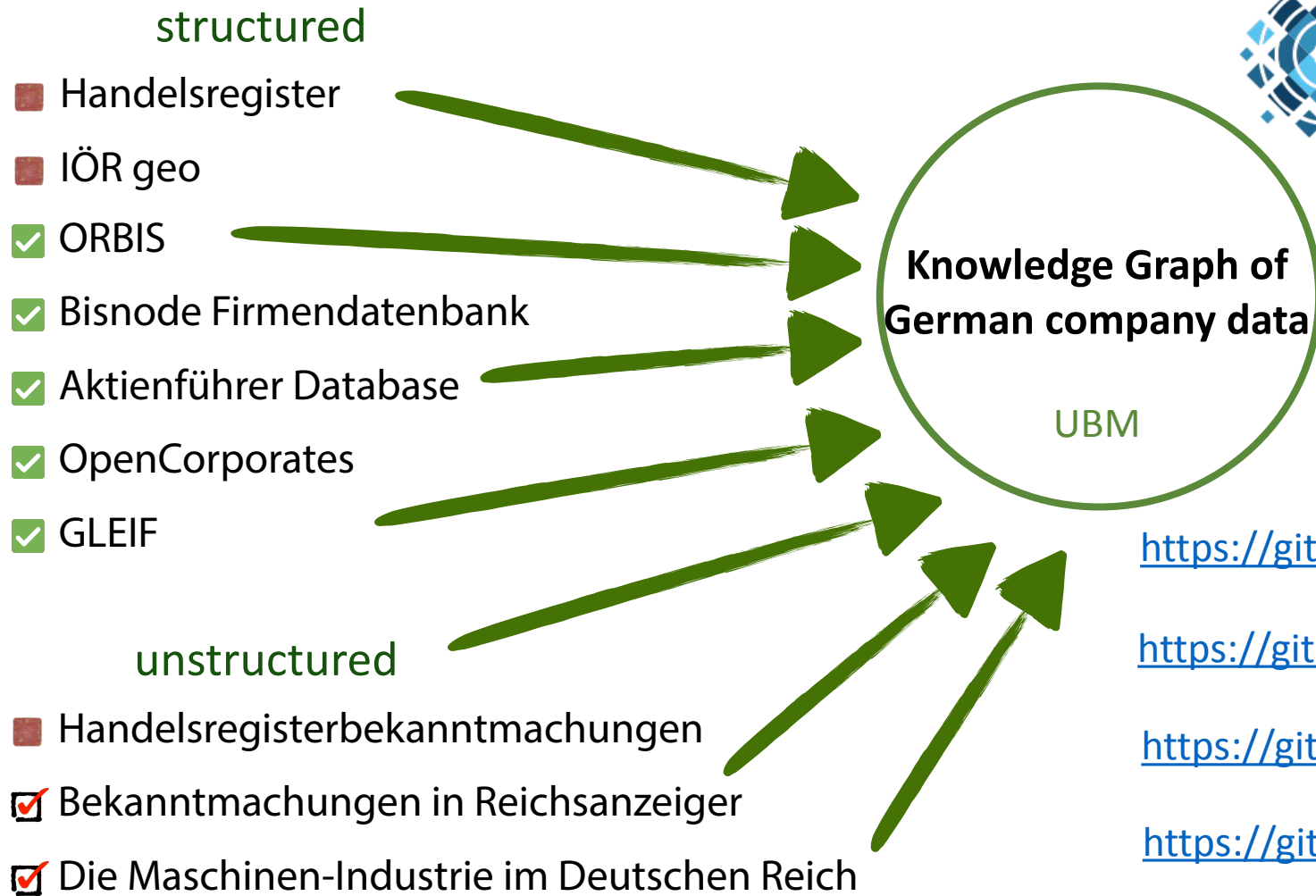
main PROBLEM



## unstructured

main PROBLEM





API + SPARQL + bulks RDF files

Export-Extension for non-technical users is under development

We use and develop open-source software for data integration, knowledge graph construction and entity linking:

<https://github.com/UB-Mannheim/bbw>

<https://github.com/UB-Mannheim/RaiseWikibase>

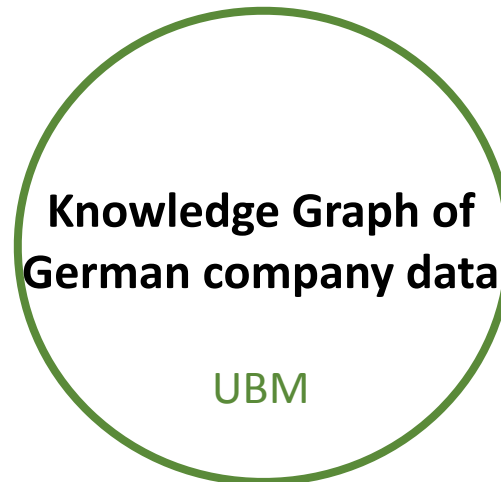
<https://github.com/UB-Mannheim/spacyopentapioca>

<https://github.com/UB-Mannheim/blatt>

<https://github.com/UB-Mannheim/reichsanzeiger-nlp>

## AS A SEPARATE SERVICE

- ✓ Free and open access to modern and historical German company data (à la Wikidata) via API, SPARQL and bulk files. The data is linked with many other databases and ontologies
- ✓ Reconciliation (linking) services and many other data science applications
- ✓ Openness and flexibility to add more German company data



## AS A KG-SERVICE FOR BERD INVENIO SINGLE POINT OF ACCESS

- ✓ Search over extended metadata automatically extracted via NEL pipelines
- ✓ Improving search experience in Invenio using aliases of entities from the knowledge graph
- ✓ Autocompleting the metadata fields using the data from the knowledge graph

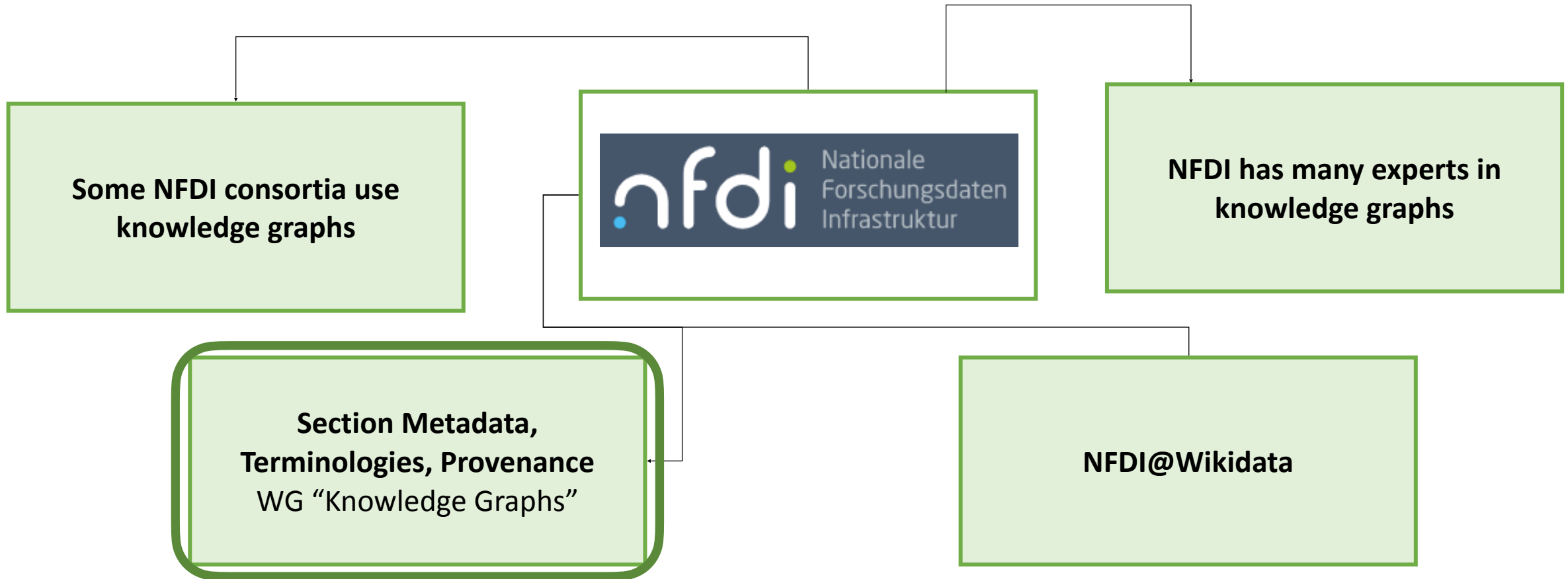
**The knowledge graphs in BERD are used for semantic enrichment.**

**The BERD knowledge graph of German company data is a separate long-awaited service which will enable many data-driven applications. The main challenge is getting data from Handelsregister.**

## Knowledge graphs for capturing, enriching and disseminating Knowledge graphs in BERD **Knowledge graphs in NFDI** Summary

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 460037581





Check out the folder of the WG KG at the Google Drive and join the group. See the Onboarding-file.

[https://www.wikidata.org/wiki/Wikidata:WikiProject\\_NFDI](https://www.wikidata.org/wiki/Wikidata:WikiProject_NFDI)



Section “(Meta)data, Terminologies, Provenance”

## WG “Knowledge Graphs” (KG)

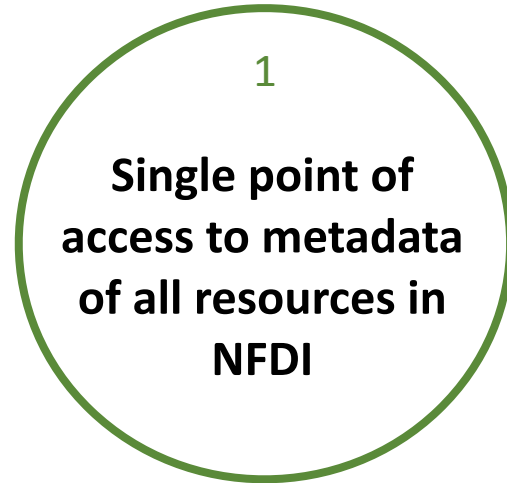
Coordinators: Renat Shigapov, Christian Bölling & Moritz Schubotz

Initiators: Lozana Rossenova & Markus Stocker

Charter: <https://doi.org/10.5281/zenodo.7228955>

Number of members: 25. Mailing list with 64 subscribers:

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



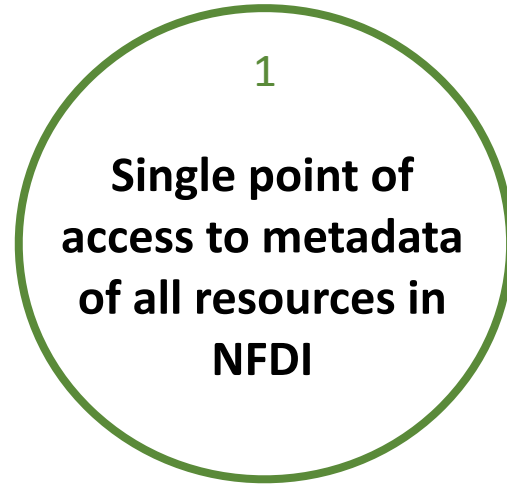
See more in the charter

<https://doi.org/10.5281/zenodo.7228955>

+ at the Google Drive  
of the WG KG

+ at <https://base4nfdi.de>

- NFDI is a distributed infrastructure. It would be nice to send one search query to get results from all consortia



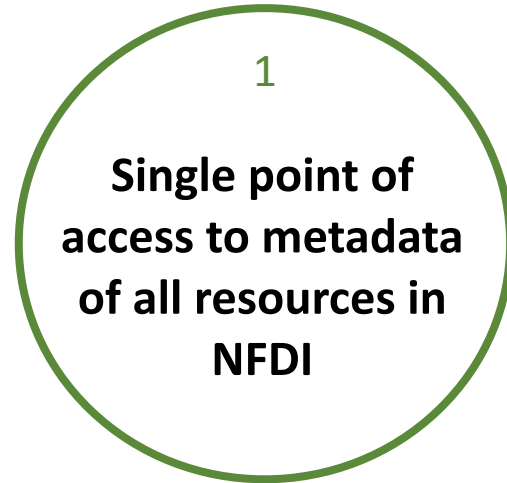
See more in the charter

<https://doi.org/10.5281/zenodo.7228955>

+ at the Google Drive  
of the WG KG

+ at <https://base4nfdi.de>

- NFDI is a distributed infrastructure. It would be nice to send one search query to get results from all consortia
- Single point of access in NFDI will enable research on NFDI itself and help multidisciplinary researchers



See more in the charter

<https://doi.org/10.5281/zenodo.7228955>

+ at the Google Drive  
of the WG KG

+ at <https://base4nfdi.de>

- NFDI is a distributed infrastructure. It would be nice to send one search query to get results from all consortia
- Single point of access in NFDI will enable research on NFDI itself and help multidisciplinary researchers
- There might be a need for multiple single points of access for the clusters of NFDI consortia working with similar types of resources

See more in the charter

<https://doi.org/10.5281/zenodo.7228955>

+ at the Google Drive  
of the WG KG

+ at <https://base4nfdi.de>

2

**Knowledge graph  
infrastructure with  
tools for linking,  
mapping, etc.**

- KGs are great, but NFDI consortia need a ready-to-use software with many related tools to solve data integration problems



See more in the charter

<https://doi.org/10.5281/zenodo.7228955>

+ at the Google Drive  
of the WG KG

+ at <https://base4nfdi.de>

2

**Knowledge graph  
infrastructure with  
tools for linking,  
mapping, etc.**

- KGs are great, but NFDI consortia need a ready-to-use software with many related tools to solve data integration problems
- NFDI consortia will be able to create metadata knowledge graphs for their resources

See more in the charter

<https://doi.org/10.5281/zenodo.7228955>

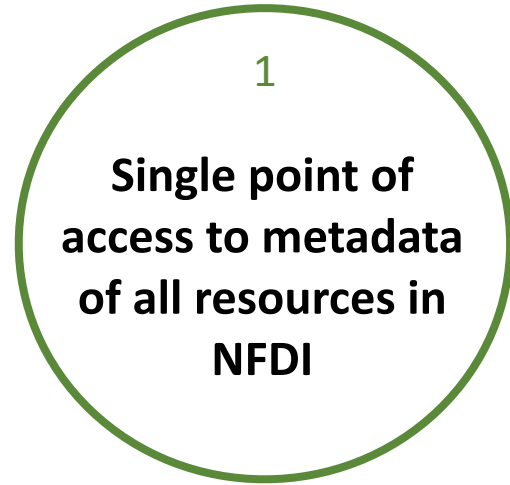
+ at the Google Drive  
of the WG KG

+ at <https://base4nfdi.de>

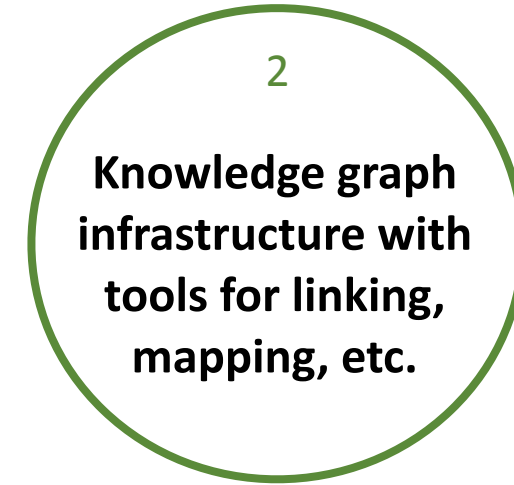
2

**Knowledge graph  
infrastructure with  
tools for linking,  
mapping, etc.**

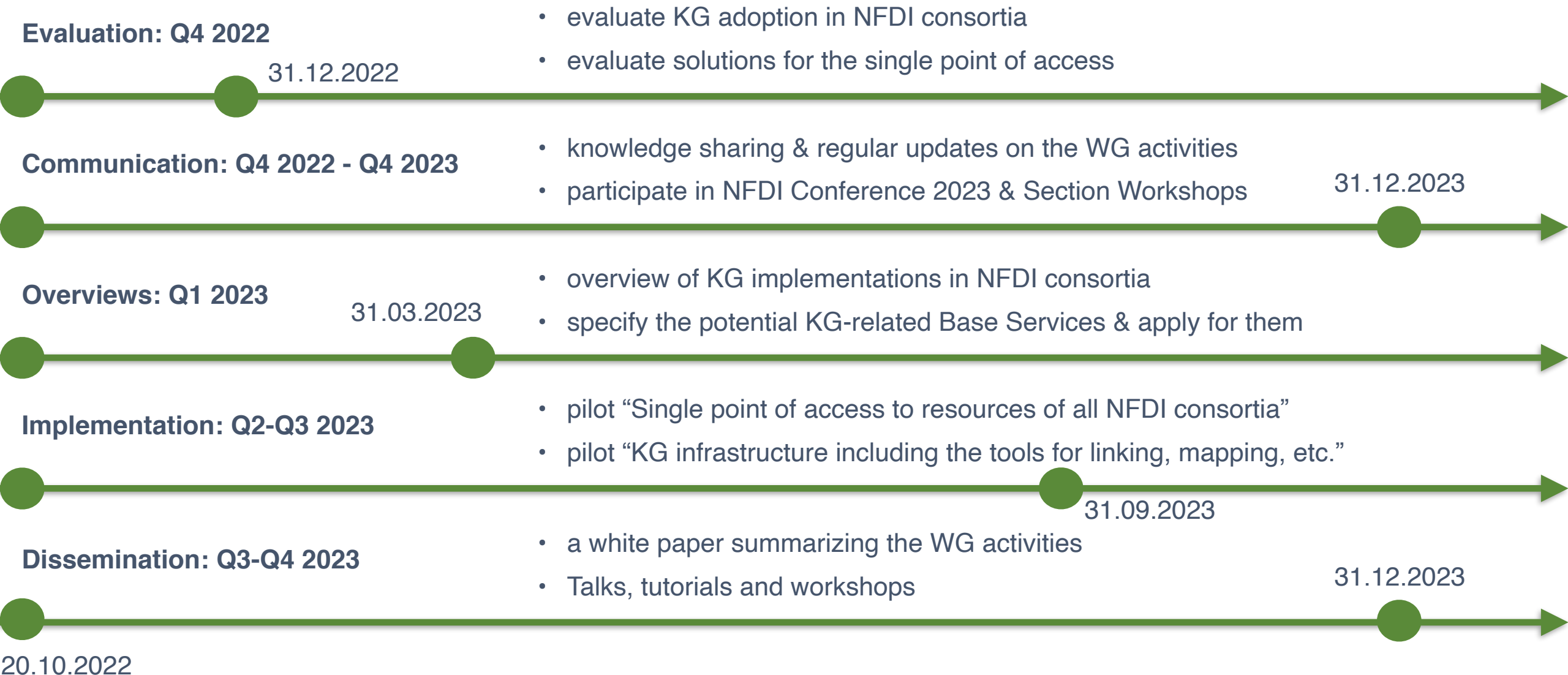
- KGs are great, but NFDI consortia need a ready-to-use software with many related tools to solve data integration problems
- NFDI consortia will be able to create metadata knowledge graphs for their resources
- Researchers will get a simple setup to release their research data objects as knowledge graphs



See more in the charter  
<https://doi.org/10.5281/zenodo.7228955>  
+ at the Google Drive of the WG KG  
+ at <https://base4nfdi.de>



- NFDI is a distributed infrastructure. It would be nice to send one search query to get results from all consortia
- Single point of access in NFDI will enable research on NFDI itself and help multidisciplinary researchers
- There might be a need for multiple single points of access for the clusters of NFDI consortia working with similar types of resources
- KGs are great, but NFDI consortia need a ready-to-use software with many related tools to solve data integration problems
- NFDI consortia will be able to create metadata knowledge graphs for their resources
- Researchers will get a simple setup to release their research data objects as knowledge graphs



**The NFDI working group “Knowledge graphs” aims to apply for two Base Services: 1) A single point of access to metadata of resources in NFDI, and 2) Knowledge graph infrastructure with tools for linking, mapping, etc. The goal is to help both consortia and researchers.**

Subscribe to our mailing list and receive an invitation to our next WG KG meeting on 01.12.2022 at 15:00:

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

## Knowledge graphs for capturing, enriching and disseminating Knowledge graphs in BERD Knowledge graphs in NFDI **Summary**

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 460037581



## Knowledge graphs for capturing, enriching and disseminating research data objects

- Knowledge graphs are great technology to make your research data object FAIR and linked
- Release your research data object as a knowledge graph
- This enables both humans and machine to reuse your data

### Knowledge graphs in BERD

- Knowledge graphs are used for semantic enrichment in BERD
- Knowledge graph of German company data is an ongoing data integration effort at UBM
- The main challenge is getting data from Handelsregister

### Knowledge graphs in NFDI

- Join the NFDI working group “Knowledge graphs” to learn more
- We plan to apply for two base services in 2023: 1) A single point of access to metadata of resources in NFDI, and 2) Knowledge graph infrastructure with tools for linking, mapping, etc.