Facilitating creation, (re)use, and interoperability for Knowledge Graphs in NFDI: the vision behind KGI4NFDI

Lozana Rossenova, Renat Shigapov, Moritz Schubotz, Fidan Limani, Benjamin Zapilko,
Till Sauerwein, Daniel Mietchen, Muhammad Elhossary, Konrad U. Förstner

1st Base4NFDI User Conference (UC4B2024), 2024-11-21

https://doi.org/10.5281/zenodo.14190374

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

- 125 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)

nfdi | Nationale Forschungsdaten Infrastruktur

Section (Meta)data, Terminology, Provenance

Working Group Charter Knowledge Graphs
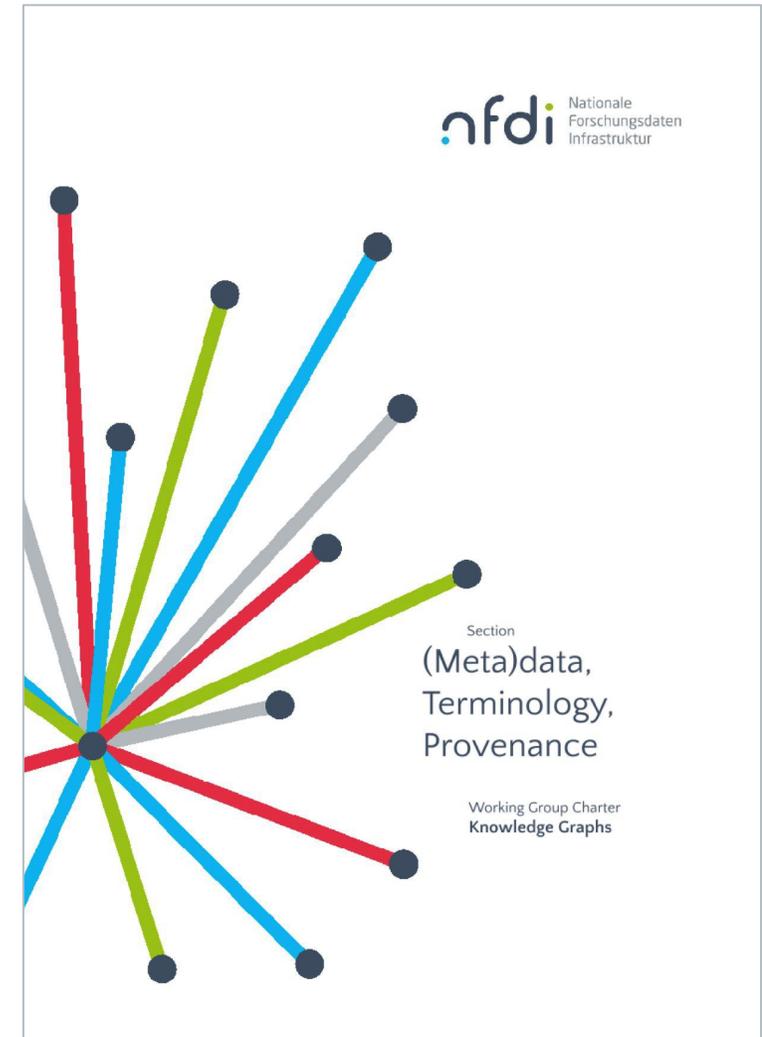
Why KGs and why KGI?

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

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

nfdi Nationale Forschungsdaten Infrastruktur

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
- Consortia, institutions and researchers need an easy-to-use, scalable and interoperable **KGI-as-a-Service**.

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)
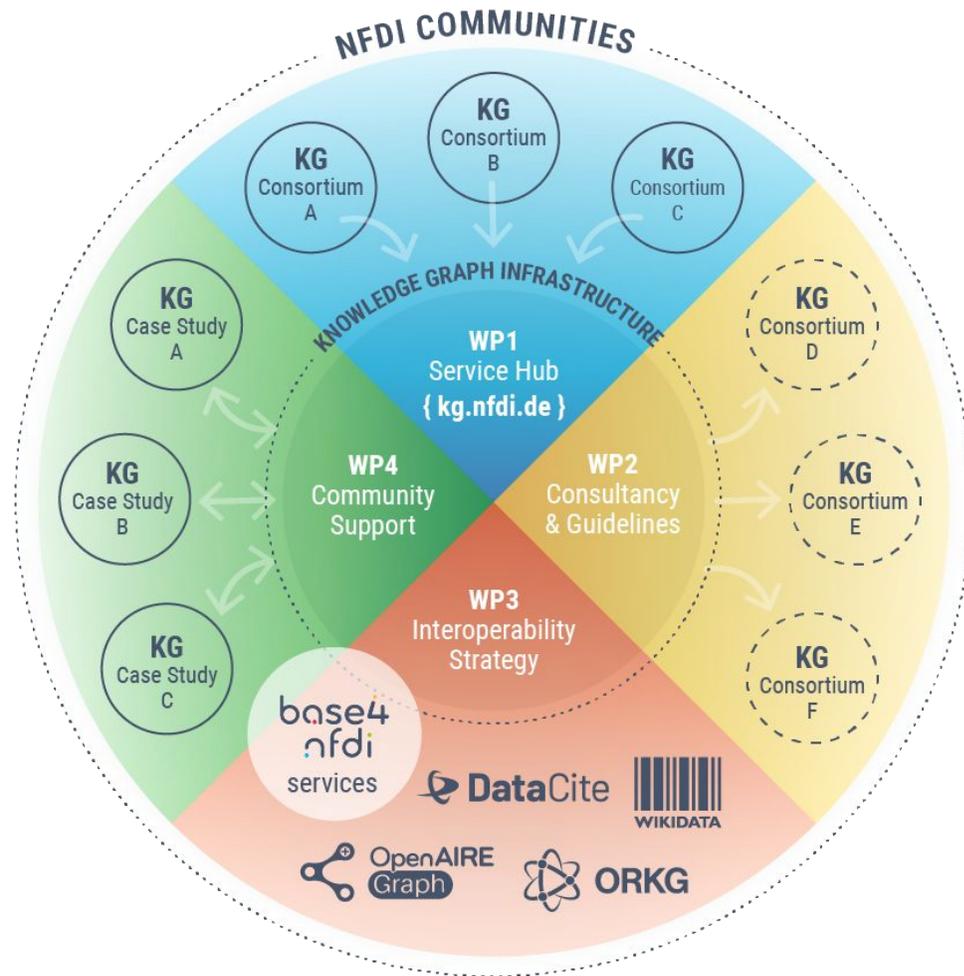- NFDI-MatWerk (KGs & KG Software)
- NFDIxCS

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

KGI4NFDI – KGI-as-a-Service

…will include a KG registry, which will aggregate information on all KGs contributed by NFDI consortia and by the research communities they represent, as well as a service to facilitate access to KGs across NFDI projects. Furthermore, the base service aims to empower research communities to create KGs by providing the necessary technologies and expertise for decentralised KG instances based on standards and tried-and-tested approaches.
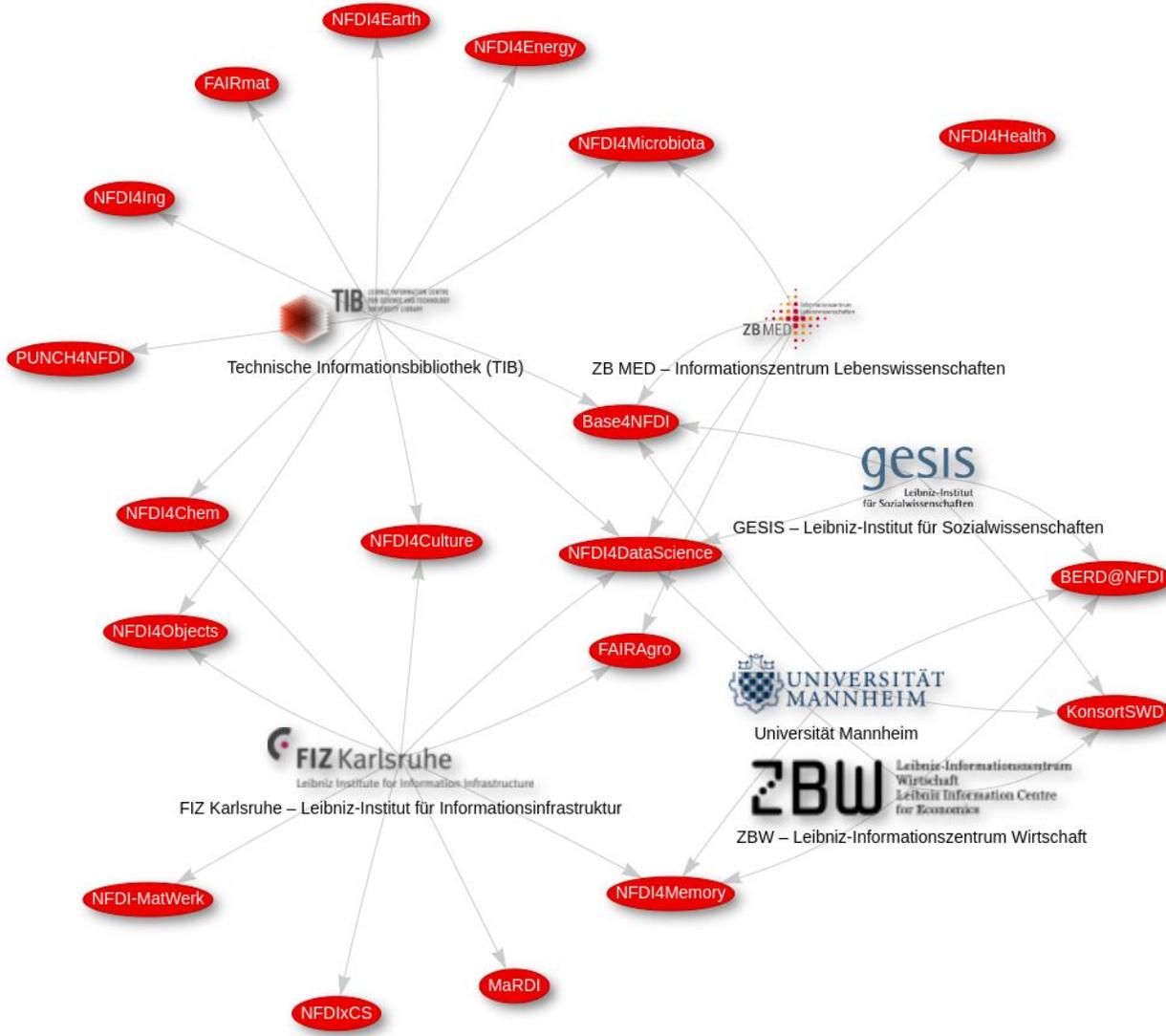
- FIZ Karlsruhe (MaRDI)
- GESIS (NFDI4DataScience)
- TIB (NFDI4Culture)
- Uni Mannheim Library (BERD@NFDI)
- ZB MED (NFDI4Microbiota)
- ZBW (KonsortSWD)

Co-applicant and participant affiliations to NFDI consortia. Data sourced from Wikidata: https://w.wiki/99YJ

Work Package 1 - Central services

- **D1.1 Central Registry of Knowledge Graphs in NFDI**
  - registry in the form of a knowledge graph to host metadata about NFDI services that rely on KGs
  - adoption of a simple ontology (based on ongoing work in several consortia)
  - editorial and curatorial process via common, easy-to-contribute-to and collaborative tool
  - RDF-based ETL workflow (based on NFDI4Culture)

- **D1.2 Platform for search and query across KGs**
  - central website
  - example queries

NFDI consortium	Knowledge Graph	URL
BERD@NFDI	Aktienführer KG	https://akf.kgi.uni-mannheim.de
BERD@NFDI	MaschinenBauIndustrie KG	https://mbi.kgi.uni-mannheim.de
MaRDI	MaRDI portal	https://portal.mardi4nfdi.de
NFDI4Cat	4Cat Meta Portal	https://meta4cat.fokus.fraunhofer.de/ (/sparql)
NFDI4Culture	Culture Knowledge Graph	https://nfdi4culture.de/resources/knowledge-graph
NFDI4Culture	Linked Stage Graph	https://slod.fiz-karlsruhe.de/about
NFDI4Culture	Semantic Kompakkt	https://wikibase.semantic-kompakkt.de
NFDI4DataScience	Open Research Knowledge Graph	https://orkg.org
NFDI4DataScience	TweetsKB	https://data.gesis.org/tweetskb
NFDI4DataScience	TweetsCov19	https://data.gesis.org/tweetscov19
NFDI4DataScience	ClaimsKG	https://data.gesis.org/claimskg
NFDI4DataScience	SoftwareKG	https://data.gesis.org/softwarekg
NFDI4DataScience	SoMeSci	https://data.gesis.org/somesci
NFDI4DataScience	Question Feature Sample	https://data.gesis.org/questionfeaturesample
NFDI4DataScience	4DS Meta Portal	https://meta4ds.fokus.fraunhofer.de/ (/sparql)
NFDI4DataScience	dblp computer science bibliography	https://sparql.dblp.org/dblp
NFDI4Earth	NFDI4Earth KG	https://nfdi4earth-knowledgehub.geo.tu-dresden.de/fuseki/dataset.html
NFDI4Ing	SciMesh	https://scimesh.org/about
NFDI4Memory	FactGRID	https://database.factgrid.de
NFDI4Microbiota	InteractOA (using Wikidata)	https://interactoa.toolforge.org
Text+	Gemeinsame Normdatei (GND)	https://www.dnb.de/EN/Professionell/Metadatendienste/Datenbezug/LDS/lds_node

Work Package 1 - Central services

https://sparql.kgi.services.base4nfdi.de/ (WIP!)

Work Package 2 – Consultancy and Guidelines

- **D2.1 Guidelines for creating and hosting Knowledge Graphs**
  - set of guidelines, step-by-step instructions, educational videos for:
    1. Installation & Configuration
    2. Data Modeling & Import
    3. Querying

  > Apache Jena and Virtuoso Draft guidelines published via GitHub

  - Ansible playbooks for deployment;

  > Wikibase deployment pipeline presented at MediaWiki Users and Dev Con 2024:
  info | video

- **D2.2 Consultancy service**
  - consultancy service with central contact point available via the main service hub
  - regular office hours as well as 1-to-1 consulting sessions on demand

https://kgi4nfdi.github.io/Guidelines/
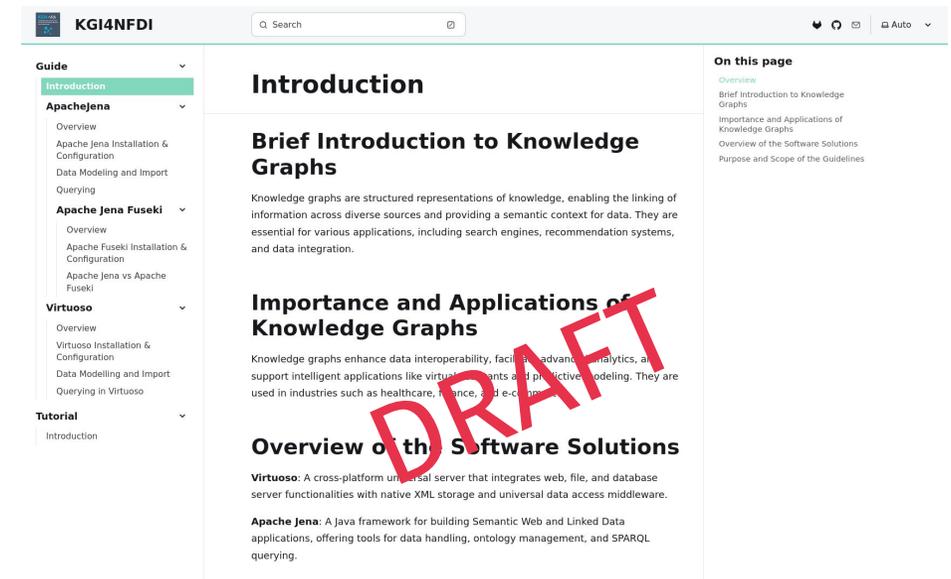
Work Package 3 – Interoperability Strategy

- **D3.1 Strategy for metadata mapping, linking, and integration**
  - based on a specification of interoperability dimensions considering:
    - existing and well-established standards & vocabularies – schema.org, Dublin Core, NFDIcore ontology, etc.
    - APIs for data sharing, such as OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting)
    - PIDs for consistent identification across all data
    - existing works on interoperability by other initiatives (e.g. EOSC interoperability framework)
  - strategy document which supports necessary consensus-building processes by providing guidelines and best practices

- **D3.2 Strategy for interoperability with national and international KG initiatives**
  - strategy to establish interoperability with: OpenAIRE Graph, PID Graph, ORKG, Helmholtz KG and Wikidata
  - identification of basis for potential collaboration and joint activities.
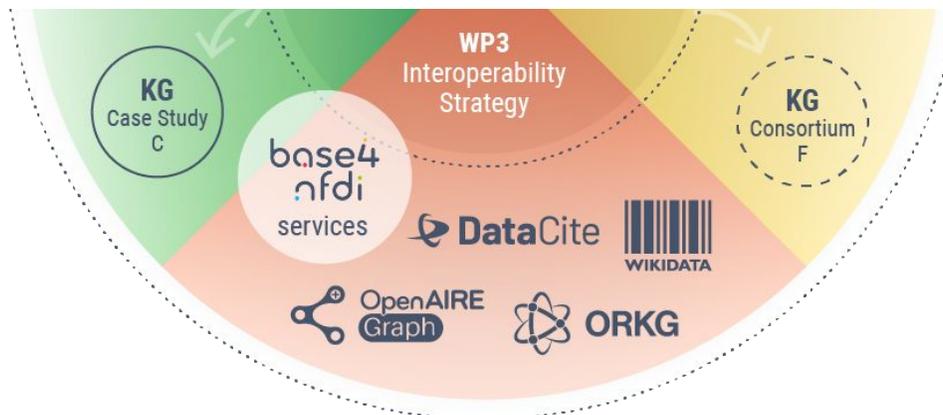
Work Package 4 - Community Support

- D4.1 Survey for the existing KG adoption practices in NFDI
  - two objectives:

    (a) identify requirements and feedback (benefits and challenges) relating to KG adoption

    (b) monitor, to the extent possible, KG adoption across the NFDI, including new use cases, technology preferences/solutions, and so on.
  - collaboration with Base4NFDI team and service steward
  - compare survey findings with international initiatives (EOSC, RDA) and monitor changes to the state of the art

- D4.2 Show cases
  - case studies to monitor and assess how the KGI basic service impacts consortia, and what synergies it offers
  - incubator process following the example of the NFDI AAI basic service to grow number over time
  - monitor development and/or adoption across consortia and facilitate exchange between KGI and the selected case studies
  - select the "query of the month" to show the value of connected NFDI KGs
  - examples of ontology harmonisation for KGs from different consortia with close topical proximity.
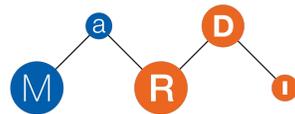
Working with the Base Services team
Developing User Personas for the KGI service

Aleena

Postdoc Research Scientist – and user of KGI4NFDI

Work: First year Postdoc Researcher in Bioinformatics

Aleena works remotely at the University of Cologne. Her scholarly interests are interdisciplinary and she enjoys contributing to citizen science projects like Wikidata. She has some experience using KGs but wants to learn more.

Goals

❑ Combine data from various sources (that the the researcher already has) into a KG.
❑ Model molecular interactions of proteins with KGs.
❑ Visualize protein-protein interactions.

Frequently Used Tools & Technologies

R, Python, APIs, relational databases

Tasks

(1) Look for existing knowledge graphs in the domain of Bioinformatics.
(2) Assess what's in a KG.
(3) Retrieve results from a KG query.
(4) Compare different entries in a KG.
(5) Decide whether to build a KG.

Pain Points

(1) Formulating correct and efficient queries.
(2) Lack of good default visualisation settings and possibilities to adjust them.
(3) Lack of good examples queries.
(4) Documentation that is incomplete or out of date.
(5) Lack of documentation suitable for beginners.
(6) Incomplete data.

Needs from KGI4NFDI

(1) A centralised registry of knowledge graphs.
(2) Good search query examples and documentation.
(3) Possibility to visualize results in useful ways.
(4) Overview of existing tools with clear indication of advantages / disadvantages.
(5) Easy tooling to merge or split entries.
(6) Autosuggestions when searching.

Nationale
Forschungsdaten
Infrastruktur

Lisa
Junior research software engineer – and user of KGI4NFDI

Work: Junior Developer in a research lab

Lisa is an MA graduate in CS, working in a research lab at a Leibniz Institute. She works closely with scientists and enjoys working on services that support innovation in science. She lacks industry experience, but is passionate about open source software and contributes to several projects in volunteer capacity.

Goals
- Gain an overview of available KG technologies.
- Learn how to set up KG environments for the scientist colleagues.
- Apply interoperability measures following scientific guidance.

Frequently Used Tools & Technologies
Git, Docker, container technologies, Python, JavaScript, relational databases, Elastic Search

Tasks
(1) Decide which KG technology to deploy.
(2) Learn about data interoperability.
(3) Deploy a new service.
(4) Deploy a specific KG solution.
(5) Troubleshoot installation and setup.

Pain Points
(1) Lack of documentation suitable for beginners.
(2) Lack of relevant case study examples.
(3) Only very general strategy documents.
(4) Unclear who runs a specific service and what is the best way to communicate with them.
(5) Comments on Git issues going unanswered.

Needs from KGI4NFDI
(1) Overview of existing tools with clear indication of advantages/ disadvantages.
(2) Guidance materials that share concrete issues arising in the use of KG technology
(3) Examples of similar services deployed by related projects.
(4) Contact point for questions and troubleshooting.
(5) Easy issue reporting.
(6) Easy access to the developer community.

Andreas
PhD Research Scientist – and user of KGI4NFDI

Work: Second year PhD Researcher in Sociology

Andreas is a PhD student at the University of Mannheim. He is interested in exploring large corpora of available sociological survey data. Andreas is comfortable using standard software for statistics and analysis, but wants to explore more innovative methods particularly in terms of discovery and data reuse.

Goals
- Innovate approaches to analysing political attitudes
- Explore highly cited publications using highly cited datasets (e.g. from existing surveys) in order to discover relevant variables for own research topic.
- Complete PhD degree.

Frequently Used Tools & Technologies
Statistics software, e.g. SPSS, Stata, R; Relational databases

Tasks
(1) Search for highly cited publications using highly cited datasets.
(2) Specify search criteria.
(3) Analyse search results.
(4) Troubleshoot result errors.
(5) Reuse search results in other data analysis software tools.

Pain Points
(1) Unclear where to begin searching and where to specify search parameters.
(2) Lack of experience with SPARQL and lack of examples or tutorials.
(3) Only one possible view of the results, e.g. linear result list.
(4) Errors whose source cannot easily be identified.
(5) No possibilities to download data in a format easy to reuse/process.

Needs from KGI4NFDI
(1) Easily accessible portal to interconnected publications and datasets.
(2) Support and tutorials for SPARQL queries.
(3) Visual and interactive ways to explore retrieved search results.
(4) Clear instructions on how to deal with errors.
(5) Possibility to download relevant survey datasets that appear in the results.

Luca
Senior research software engineer – and user of KGI4NFDI

Work: Senior Developer in a research lab

Luca is semantic web expert working in a research lab at a Fraunhofer Institute. He runs a big KG project and is eager to improve it and see how others use it. He is involved in various expert committees and wants to make knowledge and best practice sharing more effective.

Goals
- Adding the service they are running to the registry
- Ensuring interoperability between their service and other existing KG services
- Improve the visualisation of result output
- Facilitate federated search and optimise performance

Frequently Used Tools & Technologies
Git, CI/CD, OpenStack, Docker, container technologies, APIs, RDF, SPARQL, OWL, Elastic Search, Wikidata, Apache Jena, Fuseki, qlever, webvowl

Tasks
(1) Add own KG service to the KG registry.
(2) Improve visualisations for query results.
(3) Test the functionality and performance of federated queries.
(4) Resolve interoperability issues.
(5) Other improvements to existing service.
(6) Documentation and troubleshooting.

Pain Points
(1) Missing properties resulting in interoperability issues.
(2) Lack of good interdisciplinary use cases.
(3) Poor error messaging or timeouts during querying.
(4) No concrete actions after events/ conferences.
(5) Unclear issue reporting and who is responsible for issue handling.

Needs from KGI4NFDI
(1) Addition of properties when describing own service.
(2) Examples of how own KG service can be used.
(3) Information on the structure/schema of the involved KGs.
(4) Exchange with other experts in the field.
(5) Document and make service improvements easily available.
(6) Easy issue reporting.

Alexandra
Data steward – and user of KGI4NFDI

Work: Senior Data Steward for Engineering Data Domain

Alexandra works at a Leibniz Centre and is responsible for data collections in the Engineering domain. She is responsible for integrating the data in her institute to other national and international initiatives and is keen to explore the potential of KGs for this task.

Goals
- Integrate data into national and international aggregators
- Enrich own data collection with data from external KGs

Frequently Used Tools & Technologies
Git, KGs, Python, OpenRefine, RDF, SPARQL, SKOS, OWL, Ontology portals / Terminology services; Controlled Vocabulary software; relational databases;

Tasks
(1) Enrich the data from own institute.
(2) Search for and identify relevant sources for enrichment.
(3) Contribute data to national / international aggregators.
(4) Consider creating and maintaining a new KG for the data from own institute.
(5) Consider contributing to the KGI registry.

Pain Points
(1) Difficult to compare results to own source data.
(2) Lack of examples or mapping resources to support SPARQL queries across diverse sources.
(3) Lack of signposting how to contribute to aggregators.
(4) Inconsistent standards.
(5) No consultation or guidelines available
(6) No information on how to update data contribution over time.
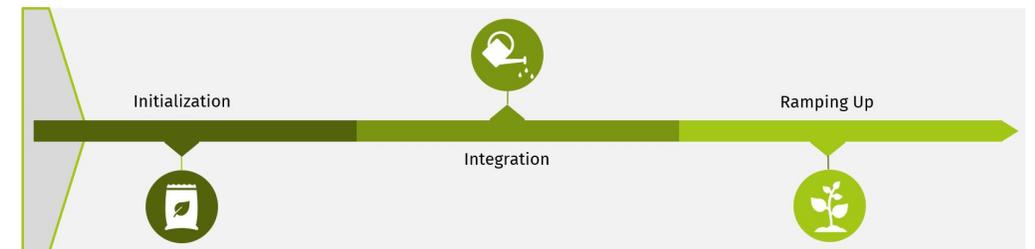
Needs from KGI4NFDI
(1) Find relevant datasets that have interconnecting points to own source data.
(2) Support with SPARQL queries, and ontology mappings that may be relevant across disciplines.
(3) Find out relevant portals to contribute to.
(4) Access to consulting service.
(5) Easy overview of currently supported ontologies and terminology sources.
(6) Easy path to contribute to the KGI registry.

Plans for the Integration Phase

- Integration of a Natural Language Interfaces using LLMs
- Tighten integration of / interaction with other Base4NFDI services
- Transitioning experimental KGs into stable stages
- Supporting ontology harmonization and mapping (according to strategy from D3.1)
- Extension of collection of Show Cases
- Interaction with EOSC services

Plans for the Ramp-up Phase

Initialization

Integration

Ramping Up

- Increase service performance and increase responsiveness
- Increase reliability and monitoring
- Reduce issue resolution time

Thank you!
Questions?