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

Lozana Rossenova[1][https://orcid.org/0000-0002-5190-1867], Moritz Schubotz[2][https://orcid.org/0000-0001-7141-4997], and Renat Shigapov[3][https://orcid.org/0000-0002-0331-2558]

[1] TIB – Leibniz Information Centre for Science and Technology, Germany

[2] FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany

[3] Universitätsbibliothek Mannheim, Germany

Introduction

The Strategic Research and Innovation Agenda (SRIA) of the European Commission identifies *Knowledge Graphs (KGs)* as one of the most important technologies for building an interoperability framework and enabling data exchange among users across countries, sectors, and disciplines [1]. KG is a graph-structured knowledge base containing a terminology (vocabulary or ontology) and data entities interrelated via the terminology [2]. KGs are based on semantic web technologies (RDF, SPARQL, etc.) and often used for agile data integration. KGs also play an essential role within Germany as a vehicle to connect research data and research-related entities and make those accessible – examples include the GESIS Knowledge Graph Infrastructure, TIB Open Research Knowledge Graph, and GND.network. Furthermore, the Wikidata knowledge graph, maintained by Wikimedia Germany, contains a large number of research-related entities and is widely used in scientific knowledge management in addition to being an important advocacy tool for open data [3]. Extending domain-specific ontology-supported KGs with the multidisciplinary, crowdsourced knowledge in Wikidata KG would enable significant applications. The linking between expert knowledge systems and world knowledge empowers lay persons to benefit from high-quality research data and ultimately contributes to increasing confidence in scientific research in society.

Motivation

To date, several NFDI consortia have started building individual KG solutions or providing KG-compatible data using different formats and endpoints [4, 5, 6]. In addition, many of the working groups within the cross-disciplinary Sections of NFDI also include KG technologies as part of their goals and planned activities, e.g. with regards to the provision of persistent identifiers [7], ontology mappings [8], data integration and exchange endpoints [9], or training materials for RDM [10]. In some cases, individual solutions are required to meet domain-specific requirements [11]. However, in many cases, the technical and organizational overhead to run KG services can be hard to justify and is a burden to individual consortia or individual institutional members of consortia. To start discipline-specific KG work easily, NFDI consortia, participant institutions and researchers need reusable, scalable *Knowledge Graph Infrastructure (KGI)*. KGI in the context of NFDI would not only include a triplestore or graph database. KGI should encompass a whole ecosystem of software allowing to create a KG, including tools for data

import, validation and export, collaborative frontends, search APIs and SPARQL endpoints with result visualization widgets, Extract-Transform-Load and data linking software adapted to the technology stack.

Service proposal

Developed in the context of base services for the NFDI, the Knowledge Graphs Working Group from Section "Metadata, Terminologies, Provenance" is proposing a pilot KGI which provides KG infrastructure-as-a-service, combining the ease of use of software like Wikidata with research-backed data. This includes 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.

The pilot KGI will be developed in an agile way, starting from one specific tool suite as a 'minimum viable product', and after an initialization phase will be expanded to meet the needs and requirements established through consultation with relevant NFDI stakeholder communities. The starting tool suite will be Wikibase, the open source software behind Wikidata KG, which is already being used by various consortia and participating institutions across a range of use cases. For example, MaRDI and BERD4NFDI are using Wikibase instances as central portals for all research data generated by their respective consortia participants [4, 5]. In the context of NFDI4Culture, TIB's Open Science Lab deploys Wikibase instances to structure data about digitized cultural objects entered and edited via a 3D-Viewer and Annotation tool [3]. NFDI4Memory have recently also partnered with an existing major Wikibase project, Fact-Grid, run by the Gotha Research Centre and hosted at the University of Erfurt, which provides a central repository for data about historical persons and events [12].

The growing adoption of Wikibase and the popularity of Wikidata itself – both as a repository to upload data to, and a rich resource on the linked open data (LOD) cloud to federate with – can serve as proof-of-concept that an approach to KGs involving a mix of human- and machine-readable interfaces can lower the barrier to participation across a wide range of disciplinary fields and foster the creation of complex, cross-disciplinary connections. Given the increasingly interdisciplinary use of resources (e.g., datasets, methods or models), ease-of-access to participating in the creation and exchange of FAIR research data in different fields is crucial and lays the foundation for multi- and cross-disciplinary discovery.

The identical approach to interfacing with data, modelling and querying data between Wikidata and Wikibase provides native federation capabilities across these services, which NFDI consortia that use Wikibase are taking advantage of [3, 4, 12]. Developing the Wikibase tool suite further in the context of a pilot KGI for NFDI will streamline deployment and customization, and improve interoperability at scale. This work is intended to lessen the burden on individual consortia, use existing synergies, and provide ready-to-use infrastructure for KGs.

Outlook

Beyond an initialization phase, the success of KGI in NFDI will depend on growing adoption of the services and computational methods (e.g. NLP or ML models) enabled by such an infrastructure. The interoperability framework will have to extend across Wikibase, Wikidata and other KG tool suites towards a unified NFDI with a EOSC compatibility layer. Regular consulting with and feedback-seeking from consortia partners will also be important for the expansion of the infrastructure service with relevant new offerings and for adapting to new use cases.
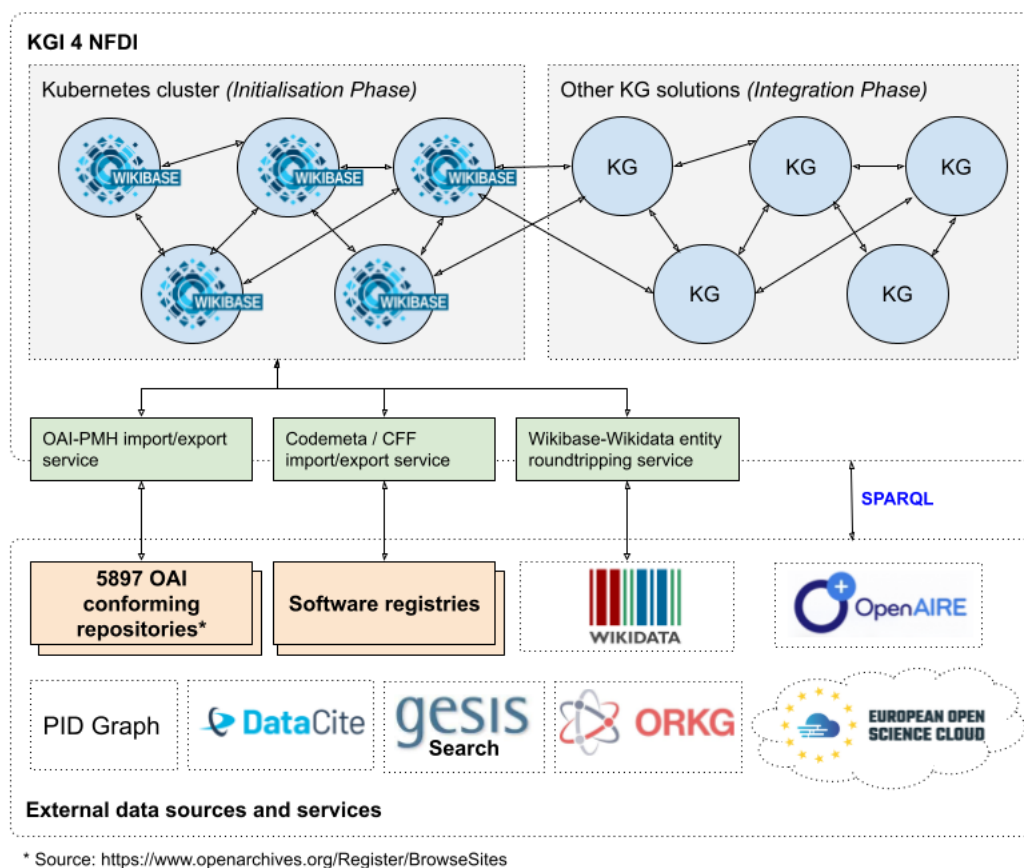
Figure 1. Overview diagram of the working concept for the KGI service.

Competing interests

The authors declare that they have no competing interests.

References

1. European Commission, Directorate-General for Research and Innovation, "Strategic Research and Innovation Agenda (SRIA) of the European Open Science Cloud (EOSC)," Publications Office of the European Union, 2022. [Online]. Available: https://data.europa.eu/doi/10.2777/935288
2. [A. Hogan, C. Gutierrez, M. Cochcz, G. de Melo, S. Kirranc, A. Pollcrcs, et al, *Knowledge Graphs.* Springer Cham, 2022. [Online]. Available: https://doi.org/10.1007/978-3-031-01918-0
3. L. Rossenova, P. Duchesne, and I. Blümel, "Wikidata and Wikibase as complementary research data management services for cultural heritage data," in *Proc. of the 3rd Wikidata Workshop 2022, co-located with the 21st International Semantic Web Conference (ISWC2022),* Virtual Event, Hangzhou, China, October 2022. [Online]. Available: https://ceur-ws.org/Vol-3262/paper15.pdf
4. The MaRDI consortium, "MaRDI: Mathematical Research Data Initiative Proposal," *Zenodo*, 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6552436
5. R. Shigapov and I. Schumm. (2021). BERD: The knowledge graph of German companies. Presented at Wikibase in Knowledge Graph based Research Data Management (NFDI) Projects. [Online]. Available: https://madoc.bib.uni-mannheim.de/58793
6. S. Bruns, H. Fliegel, E. Posthumus, H. Sack, T. Schrade, and T. Tietz. (2023). Knowledge Graph-basierte Forschungsdatenintegration in NFDI4Culture. Presented at

DHd2023 - Open Humanities Open Culture, Belval and Trier. [Online]. Available: https://doi.org/10.5281/zenodo.7748740

7. S. Bingert, J. Brase, F. Burger, B. Dreyer, S. Hagemann-Wilholt, P. Vierkant, and P. Wieder, "Concept for Setting up the Persistent Identifier Services Working Group in the NFDI Section "Common Infrastructures" (1.0)," *Zenodo,* 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6507760

8. I. Anders,T. Arera-Rütenik, S. Arndt, R. Baum, N. Betancort, I. Blümel, C. Busse, A. Daniel, F. Engel, L. Ghiringelli, S. Hachinger, H.r Israel, N. Karam, A. Kranz, R. Lenz, D. Linke, T. Petrenko, L. Rossenova, D. Schulz, and N. Kockmann, "Ontology Harmonization and Mapping - Working Group Charter (NFDI section-metadata) (1.0)," *Zenodo,* 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6726519

9. M. Stocker, L. Rossenova, R. Shigapov, N. Betancort, S. Dietze, B. Murphy, C. Bölling, M. Schubotz, and O. Koepler, "Knowledge Graphs – Working Group Charter (NFDI Section-Metadata) (1.2)." *Zenodo*, 2023. [Online]. Available: https://doi.org/10.5281/zenodo.7802304

10. P. Pelz, S. Herres-Pawlis, J. Liermann, F. Strauß, D. Ohse, N. Kockmann, M. Liebau, D. Tschink, J. Dierkes, J. Vandendorpe, R. Müller, K. Förstner, T. Hamann, C. Keßler, J. O. Heuer, D. Hausen, K. Sauerland, A. Bonn, A. Münzmay, and T. Hörner, "Working Group Charter Training Infrastructures (1.0)," *Zenodo,* 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6478698

11. P. Strömert and O. Koepler. (2022). Ontologies4Chem: A use case to build a NMR research data knowledge graph. Presented at ACS Spring Meeting. 23.03.2022. [Online]. Available: https://docs.google.com/presentation/d/1qr2OiFVW4u-KFjtD71D08zOeEdH1q-nh/edit#slide=id.p1

12. O. Simons. "FactGrid Goes NFDI." FactGrid Blog. https://blog.factgrid.de/archives/3104 (accessed April 20, 2023).