The Helmholtz Knowledge Graph: driving the transition towards a FAIR data ecosystem in the Helmholtz Association

Jens Bröder[1] [0000−0001−7939−226X], Gabriel Preuß[2] [0000−0002−3968−2446], Fiona D'Mello[1] [0000−0002−0465−1009], Said Fathalla[1] [0000−0002−2818−5890], Volker Hofmann[1] [0000−0002−5149−603X], and Stefan Sandfeld[1] [0000−0001−9560−4728]

[1] Forschungszentrum Jülich GmbH, https://ror.org/02nv7yv05, Institute of Advanced Simulation - Materials Data Science and Informatics (IAS-9), Germany
{j.broeder, f.dmello, s.fathalla, v.hofmann, s.sandfeld}@fz-juelich.de
[2] Helmholtz-Zentrum Berlin für Materialien und Energie, Germany
gabriel.preuss@helmholtz-berlin.de

Abstract. The Helmholtz knowledge graph aggregates digital assets and research output from the various institutional and siloed digital infrastructures within the Helmholtz association. It is part of the technical backbone of a FAIR data space that is established by the "Helmholtz Metadata Collaboration" (HMC). There, it is used to drive change towards better metadata practices, increase visibility of data and provide useful data-based services. In this paper, we present how metadata used to describe Helmholtz's digital assets and research output is harvested and uplifted. The data is made publicly accessible to both humans and machines through text search and a SPARQL endpoint respectively.

Keywords: Knowledge graph · Linked-Data · FAIR · Helmholtz-Metadata-Collaboration · schema.org · jsonld · data-mining · Metadata

1 Introduction

Research in the Helmholtz Association is carried out in inter- and multidisciplinary collaborations that span between its 18 independently operating non-university research centers across Germany. Helmholtz digital infrastructure is institutional, and thus Helmholtz's research data and other digital assets are stored and maintained in independent silos, lack visibility and accessibility and their full value remains unavailable to scientists, managers, strategists, and policy makers. Metadata on the web is typically used to track citations not data. It often lacks completeness and semantic quality and therefore, published research data often fails to satisfy FAIR principles [1] resulting in a lack of interoperability and re-usability. The "Helmholtz Metadata Collaboration (HMC)[3]" is taking on this challenge by developing innovative technologies and tools for a sustainable handling of research data through high-quality metadata. Consequently, we

[3] https://helmholtz-metadaten.de/en

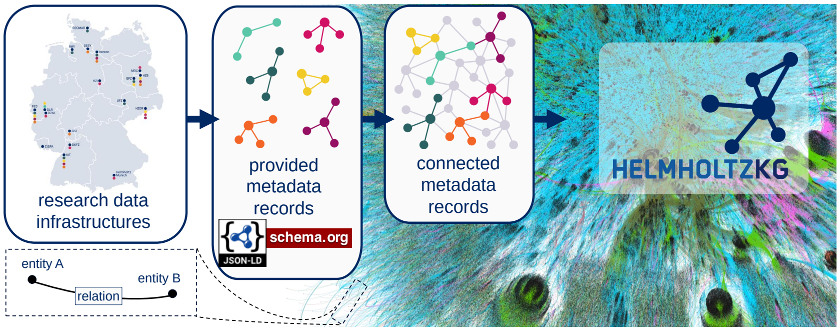
research data infrastructures
provided metadata records
schema.org
JSON-LD
connected metadata records
HELMHOLTZKG
entity A
entity B
relation

Fig. 1. Aggregation of the Helmholtz KG:
Data records (schema.org / JSON-LD) are continuously harvested from Helmholtz data providers and run through our data pipelines for initial uplifting, de-duplication and integration into the Helmholtz KG.

launched the "unified Helmholtz Information and Data Exchange (unHIDE)[4] - an initiative to network and harmonize Helmholtz digital infrastructure, and connect Helmholtz data through a lightweight interoperability layer in form of the Helmholtz Knowledge Graph. With this, we envision to (1) provide a better cross-organizational access to Helmholtz's (meta)data and information assets on an upper semantic level, (2) harmonize and optimize the related metadata across the association, and (3) form a basis from where semantic quality and depths of metadata descriptions can be improved and extended into domain and application levels. The institutional focus and defined domain boundaries within the Helmholtz's research fields differentiate the Helmholtz KG from other graphs with wider scopes, such as e.g. the OpenAIRE graph[2], which we approach as partners for graph-graph exchange of data and developed technologies.

2 Data aggregation and statistics

To aggregate the Helmholtz KG, metadata records are harvested from Helmholtz data providers and integrated (Fig. 1): we developed a library of harvesters[5] that harvests records recurrently through common web standards such as OAI-PMH, sitemaps, feeds, or from the APIs of established data providers (DataCite, GitHub, GitLab). The data in the Helmholtz KG is aligned along `https://schema.org` semantics, for which exposed metadata records are preferably provided as linked-data serialized (e.g. JSON-LD) documents. All harvested records are processed through a data pipeline utilizing the workflow manager Prefect version 2.15.0, during which records are initially uplifted with inferable semantic annotations and then de-duplicated. Then, records are uploaded into an OpenLink Virtuoso triple store, and indexed into an Apache Solr database

[4] `https://helmholtz-metadaten.de/en/unhide_helmholtz-kg`

[5] Software project: `https://codebase.helmholtz.cloud/hmc/hmc-public/unhide`

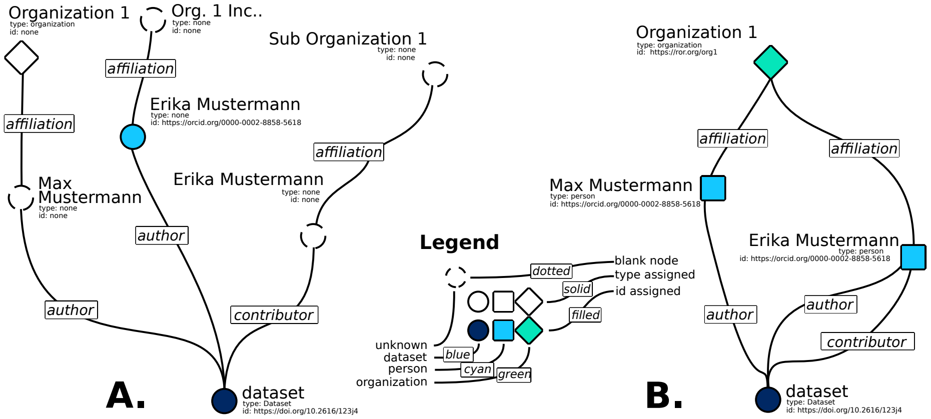

Fig. 2. Uplifting data records:
A. Often, metadata is not systematically typed or assigned with persistent identifiers (PIDs). The resultant connections show that the same entities appear as duplicated blank nodes in several instances. **B.** The same data with assigned types and PIDs allows resolution of entities leading to a higher connectivity of the data in the graph.

to support full federated text search. The graph is exposed as a set of triples through a SPARQL endpoint[6]. In addition, users can search the graph data through a user-friendly web front end[7] as well as an API[8]. The aggregation and deployment design was inspired by the Ocean InfoHub (OIH) project[3] whose open code base kick-started our development. The graph is deployed on the HDF Cloud at the Jülich Supercomputing Centre[4]. The first release of the Helmholtz KG contains 2.15 mil. metadata records which were harvested from 32 Helmholtz data providers. At the graph level this results in 72 mil. RDF triples. Of these, 16.35 mil. entities are associated with a semantic type (approx. 7 typed entities per record). Of these, 793k entities are associated to an persistent identifier (PID) or URL. Currently, the most frequent types of entities are persons, organizations, documents, datasets, software and events. PIDs as well as correct semantic annotation (i.e. of entity types) are important to increase the connectivity in the graph by resolving entities as shown in Fig.2. Semantically poor data often lacks PIDs (Fig.2A) resulting in duplicated instances of the same nodes within the same record. This impairs search queries up to a level where duplicated instances might not be recognized and found for a given query. By assigning PIDs (Fig.2B) entities can be resolved leading to an increased number of connections to a single node. PID usage with data varies with entity types: orcid.org and ror.org identifiers are found to be used to refer to persons respectively organizations exclusively. In contrast doi.org identifiers are used to refer

[6] SPARQL endpoint: `https://sparql.unhide.helmholtz-metadaten.de`
[7] Web front end: `https://search.unhide.helmholtz-metadaten.de`
[8] Web API: `https://api.unhide.helmholtz-metadaten.de`

to a number of different research outputs including research data and scholarly communications.

3 Outlook

In the future, we plan to continuously grow the graph by connecting more infrastructures as data providers from within Helmholtz. We further look to integrate data from Helmoltz FAIR digital objects. Through consulting and assisting data providers to expose high-quality metadata on the web we will (1) increase their search engine optimization and (2) harmonize the way how top-level metadata in our association is used. Further, we will use the graph data to uplift and semantically enrich the provided data records. This will be achieved by type inference and entity resolution through logic and the application of machine learning methods. This uplifted data will be contrasted with the original data and can be provided back to the authoritative source of the metadata. We aim to keep the graph semantics and technology interoperable with other scientific knowledge graphs - such as the semantic pedigree ODIS [3] - to allow graph-graph interaction of data exchange and federated querying.

Data and software availability All software related to the Helmholtz knowledge graph is open source and freely available. The graph data can be fully extracted via API and the SPARQL endpoints. Versioned dumps of the graph data are planned in the future.

Acknowledgments. This project was funded by the Helmholtz Metadata Collaboration (HMC), an incubator-platform of the Helmholtz Association within the framework of the Information and Data Science strategic initiative. The authors acknowledge the Helmholtz Data Federation (HDF) for providing services and compute on the HDF Cloud at the Jülich Supercomputing Centre (JSC). We thank our colleagues Pier Luigi Buttigieg, Thomas Jejkal, Oonagh Mannix, Anton Pirogov, Silke Gerlich and Mustafa Soylu for contributions to the unHIDE initiative.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Wilkinson, Mark D., et al. "The FAIR Guiding Principles for scientific data management and stewardship." Scientific data 3.1 (2016): 1-9. `https://doi.org/10.1038/sdata.2016.18`
2. Manghi P., et al. (2022). "OpenAIRE Research Graph Dataset", Dataset, Zenodo. `https://doi.org/10.5281/zenodo.3516917`
3. FILS, Douglas, et al. Ocean InfoHub: A Global Knowledge Network for the Ocean Data and Information System (ODIS). In: AGU Fall Meeting Abstracts. 2021. S. IN45H-0523.
4. B. Hagemeier: HDF Cloud – Helmholtz Data Federation Cloud Resources at the Jülich Supercomputing Centre. JLSRF., 5, A137. (2019) `https://doi.org/10.17815/jlsrf-5-173`.