A Knowledge Graph-Based Data Integration Workflow for NFDI4Culture and Beyond

Authors:
Jonatan Jalle Steller[1]*, Linnaea Söhn[1], Torsten Schrade[1], Oleksandra Bruns[2,3], Tabea Tietz[2,3], Etienne Posthumus[2], Sarah Rebecca Ondraszek[2,3], and Harald Sack[2,3]
*Lead presenter
[1]jonatan.steller@adwmainz.de, Academy of Sciences and Literature Mainz, Geschwister-Scholl-Straße 2, 55131 Mainz, Germany
[2]FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany
[3]Karlsruhe Institute of Technology (AIFB), Kaiserstr. 89, 76133 Karlsruhe, Germany

Abstract:
Despite covering specific scientific disciplines, all NFDI consortia share similar concepts such as organizations, people, institutions, areas of expertise, data repositories, projects, data sets, and research (meta)data, just to name a few [1, 2]. When interconnected, such information has the potential to open up new research horizons. To achieve this, however, the data needs to be available as linked open data (LOD). Hence a comprehensive workflow, which includes data discovery, harvesting, preprocessing, mapping, and integration into a knowledge graph (KG) is required, which is the topic of the present work [2, 4].

Taking the NFDI4Culture KG as an example, we present a workflow for data ingestion as LOD, which is, in principle, also applicable to other NFDI consortia. Here, the NFDI4Culture KG acts as a single point of access to various decentralized research data resources, and aggregates diverse and isolated data from the research domain, enabling discoverability, interoperability and reusability of cultural-heritage data [3].

The NFDI4Culture KG consists of the Research Information Graph (RIG), which describes metadata such as publishers, contact points, standards, licenses, and data portals, and the Research Data Graph (RDG), which interconnects the content metadata provided by data portals to make granular items accessible for search. Taking into account the challenges and objectives of NFDI4Culture to aggregate a diverse landscape of cultural-heritage research data for improved interoperability, we designed a Python package of reusable LOD components, harvesters using these components, a SPARQL endpoint explorer (*shmarql*), and an ETL (Extract, Transform, Load) pipeline. The latter consists of six modular workflow components, adaptable for independent use or within a comprehensive, automated ingest routine.

Step 1: run harvest routines. This works through a set of RDF-based action files with schema.org-based step definitions to scrape remote data, connect the feed to its metadata in the RIG, and generate persistent identifiers for imported resources. To ensure harmonization and interoperability across harvested data, transformations available in our Python package are applied to generate triples according to the *nfdicore*/*cto* ontology from various common data formats in the cultural-heritage domain, if necessary.

Step 2: clean harvested data. To ensure harmonization between the harvested data feed and its associated action file, triples representing the harvesting state are added or deleted.

Step 3: commit harvest state. Changes made by a harvesting run are pushed to the pipeline's own repository to ensure up-to-date action files.

Step 4: prepare and index data. If there are changes in a data feed, data directories are automatically updated or created and search indexes are produced.

Step 5: build a new endpoint. To prevent downtimes, a new SPARQL endpoint container is built while the previous version remains available. Once the new endpoint becomes operational, the old container is stopped and removed.

Step 6: publish statistics. In a last step, statistics about the integrated data feeds are pushed to a public dashboard. It supports data analysis and visualizations based on the execution of provided SPARQL queries.

Deletions and alterations in the KG are handled by the same routine as data feeds are periodically re-harvested and included in the next endpoint.

References:

[1] Sack, H., Schrade, T., Bruns, O., Posthumus, E., Tietz, T., Norouzi, E., Waitelonis, J., Fliegl, H., Söhn, L., Tolksdorf, J., Steller, J.J., Azocar Guzman, A., Fathalla, S., Zainul Ihsan, A., Hofmann, V., Sandfeld, S., Fritzen, F., Laadhar, A., Schimmler, S., Mutschke, P.: Knowledge Graph Based RDM Solutions: NFDI4Culture - NFDIMatWerk - NFDI4DataScience. In: 1st Conf. on Research Data Infrastructure (2023).
[2] Bruns, O., Tietz, T., Söhn, L., Steller, J.J., Ondraszek, S.R., Posthumus, E., Schrade, T., Sack, H.: What's Cooking in the NFDI4Culture Kitchen? A KG-based Research Data Integration Workflow. In: 4th Workshop on Metadata and Research (objects) Management for Linked Open Science (DaMaLOS), co-located with ESWC (2024).
[3] Tietz, T., Bruns, O., Fliegl, H., Posthumus, E., Schrade, T., Sack, H.: Knowledge Graph-basierte Forschungsdatenintegration in NFDI4Culture. In: Busch, A., Trilcke, P. (eds.) DHd2023: Open Humanities, Open Culture (2023).
[4] Bruns, O., Söhn, L., Tietz, T., Steller, Jalle., J., Posthumus, E., Schrade, T., Sack, H., Gotta Catch'em All: From Data Silos to a Knowledge Graph. In: ESWC Satellite Events: Poster and Demo Track (2024).

Keywords: knowledge graph, ontology, semantic web, research data management, FAIR