1st Conference on Research Data Infrastructure
Poster presentations II
https://doi.org/10.52825/CoRDI.v1i.391

Published: 07 Sept. 2023

NFDI4DS Gateway and Portal

Ricardo Usbeck[1] [https://orcid.org/0000-0002-0191-7211], Tilahun Abedissa Taffa[1] [https://orcid.org/0000-0002-2476-8335], Rudy Alexandro Garrido Veliz[1] [https://orcid.org/0009-0008-1582-2417], Rana Abdullah[1] [https://orcid.org/0009-0000-2652-5129] , Najeebullah Shams[1] [https://orcid.org/0000-0003-3725-2554] , Bianca Wentzel[2] [https://orcid.org/0000-0002-9218-5676], Zongxiong Chen[2] [https://orcid.org/0000-0003-2452-0572] , and Sonja Schimmler[2] [https://orcid.org/0000-0002-8786-7250]

[1]University of Hamburg, Germany
[2]Fraunhofer FOKUS, Germany

Abstract: NFDI4DataScience (NFDI4DS) is a consortium to support researchers in all stages of the research data lifecycle to conduct their research in line with the FAIR principles. The developed infrastructure targets researchers from a wide range of disciplines in data science and AI.

We present the ideas of the NFDI4DS gateway and the NFDI4DS portal. Two approaches to navigate digital objects (articles, data, machine learning models, workflows, scripts/code, etc.) from various NFDI4DS resources such as the ORKG, the DBLP database, and other research knowledge graphs (KGs). Transparency, reproducibility, and fairness will be fostered by a step-wise integration of existing and newly developed services into the overall system.

With this paper, we want to engage with the community and understand the needs and challenges of researchers in various disciplines regarding data science and AI. Therefore, we will discuss the currently developed prototypes and outline our plans for future development steps.

Keywords: Enabling RDM (incl. software), Linking RDM, FAIR, FDO

1 Introduction

NFDI4DS Gateway and Portal

The current paradigm shift towards data-driven and deep learning-based approaches in data science requires an expert-level understanding of available resources. Consequently, the community needs an integrated gateway and portal functioning as a search engine over multiple scholarly repositories and services. Through a unified search and exploration interface, users will be able to query a wide range of scientific databases. The results are currently mapped to Schema.org[1] and DCAT-AP[2]. The NFDI4DS gateway offers querying in an ad-hoc fashion, the NFDI4DS portal provides a harvesting-based approach to account for larger research data dumps.

[1] https://schema.org
[2] https://github.com/SEMICeu/DCAT-AP/

2 Approach

Architecture

The NFDI4DS gateway receives a search key, executes it against the repositories using their respective APIs, and outputs human-readable results.The heterogeneous API results are mapped to their respective schema. Since each entity has a different identifier in each repository, the background controller deduplicates results. In contrast, the NFDI4DS portal filters the underlying knowledge base of harvested metadata based on filters or by keyword utilizing the integrated search engine.

Currently, the ad-hoc based approach[3] is utilizing ten open-source scholarly repositories, namely DBLP[4], OpenAlex[5], CORDIS[6], European Language Grid[7], GEPRIS[8], GESIS[9], ORCiD[10], RESODATE[11], WIKIDATA[12], IEEE[13], and Zenodo[14]. Among these repositories, DBLP, OpenAlex, IEEE, GESIS, RESODATE, WIKIDATA, and Zenodo provide research resources like publications, datasets, software, etc. GEPRIS provides Deutsche Forschungsgemeinschaft (DFG) funded projects, likewise, CORDIS is a primary source for projects financed by the European Union (EU) commission. ELG is a platform to avail multi-lingual, cross-lingual, and mono-lingual language technologies in the EU. Unlike the others, to distinguish researchers uniquely, ORCiD provides a unique persistent researcher-owned and controlled digital identifier.

Similar to the gateway, the harvesting-based approach[15] retrieves metadata via APIs and harvesting interfaces of different repositories. The system is based on Piveau[1], a fully-fledged data management ecosystem. During the harvesting process, the metadata is transformed into the machine-readable format RDF using the DCAT-AP specification, which is based on W3C's Data Catalogue Vocabulary (DCAT) providing a plethora of properties, vocabularies and guidelines to express information about Open Data. This data is stored, indexed, and made available via a separate frontend and via APIs. In addition, the portal provides a SPARQL endpoint that enables direct querying of the underlying knowledge graph formed by the harvested metadata.

The plan is to integrate both systems so that the data from both approaches can be accessed via one entry point. In the future, we will not only be an aggregator but also run additional services like an assessment service. We also foresee using the RDF graphs for further downstream tasks as well as offering the architecture to other consortia.

Search Paradigms

Searching paradigms can be classified as keyword, structured/controlled, or natural language questions, depending on the technique used to process a given query. Key-

[3] https://nfdi-search.nliwod.org/
[4] https://dblp.org
[5] https://docs.openalex.org
[6] https://cordis.europa.eu
[7] https://live.european-language-grid.eu
[8] https://gepris.dfg.de/gepris/OCTOPUS
[9] https://www.gesis.org/home
[10] https://orcid.org
[11] https://resodate.org/resources/
[12] https://www.wikidata.org/wiki/Wikidata:Main_Page
[13] https://www.ieee.org
[14] https://zenodo.org
[15] https://meta4ds.fokus.fraunhofer.de

word searches match only lexical terms and do not consider structural or semantic mappings. To enhance precision, the controlled keyword search paradigm was developed. Structured searches require users to write a structured query like SQL or SPARQL and should be familiar with the querying interfaces. In contrast, the Question and Answering (QA) searching paradigm enables users to input natural language questions and receive answers by analyzing and reasoning over the underlying data source. The NFDI4DS gateway currently only utilizes keyword matching but foresees providing a Large Language Model-powered chatbot for search result page analysis. The NFDI4DS portal is based on keyword and faceted search. The faceted search enables users to refine search results by applying filters based on different attributes, enhancing the search experience and facilitating the exploration of specific subsets of data.

Frontend

Both systems provide a frontend showcasing the data in a direct way (see Figure 1). The interfaces are composed of a header menu as well as a search bar for exploring the data. The gateway menu provides links to events, the community and additional services[16] while the portal menu also includes a link to the SPARQL endpoint of the system. Both systems show a list of obtained results with basic information divided into tabs based on their content. While the portal currently only supports datasets and catalogues, the gateway provides a multitude of tabs covering researchers, articles or events. Each listed result links to a details page containing more detailed information on the data as well as links to additional NFDI4DS services. The NFDI4DS portal also provides links to the underlying linked semantic data. The gateway will enable searching via a chatbot through the user interface utilizing a different approach than the portal which offers filtering of results based on facets.

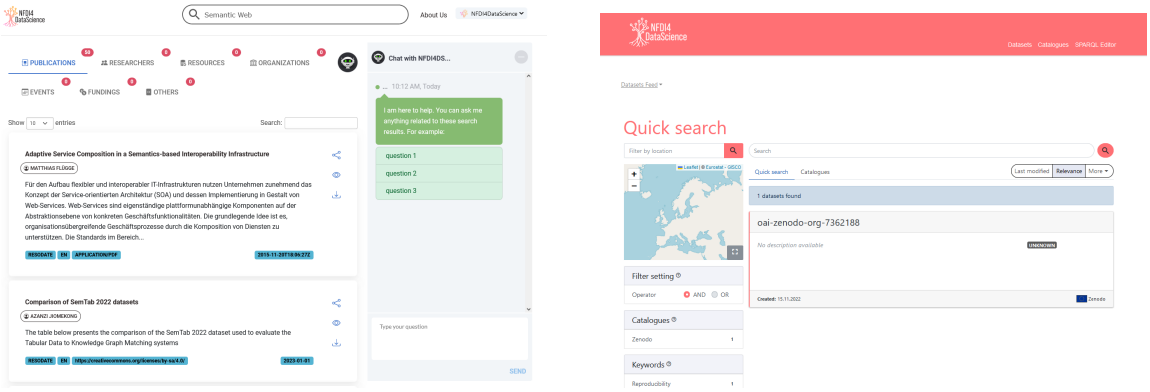

Figure 1. Interface of the gateway prototype (left) and the portal prototype (right).

3 Service Integration

The NFDI4DS gateway and portal integrate or will integrate different services in the future. Key to the success is a seamless integration of these services. In the following, we describe some central integrations that are currently under development:

[16] Using https://gitlab.com/TIBHannover/nfdi4ds/nfdi4ds-widget

NFDI4DS Research Knowledge Graph

NFDI4DS aims at providing well-structured public knowledge about scholarly resources and their adoption and relations to facilitate various use cases.The NFDI4DS research KG will entail automatically extracted metadata about resource relations, e. g. software mentions in scholarly publications, highly quality-controlled manual annotations of scholarly resources, and community annotations of scholarly publications from the ORKG.

NFDI4DS Registries

The consortium aims at providing registries for different digital objects, one of which is the DBLP computer science bibliography. DBLP and the ORKG started linking author and publication entities within their respective knowledge graphs. Using this linkage, ORKG can make use of DBLP's semantic organization and intellectual author disambiguation which, among other information, provides author identities even in the absence of schemes like ORCID [2]. At the same time, DBLP will make use of ORKG's deep semantic description of research content to enhance the information given on DBLP's website.

4 Conclusion

The NFDI4DS gateway and portal is currently under heavy development. Some features will be developed only after the finalization of this paper. Since NFDI is on the move, it is considered, that with this submission, the need to get in touch with and understand data scientists across consortia and beyond, can be satisfied.

Competing interests

The authors declare that they have no competing interests.

Funding

This joint project received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number: NFDI4DataScience (460234259).

References

[1] F. Kirstein, K. Stefanidis, B. Dittwald, S. Dutkowski, S. Urbanek, and M. Hauswirth, "Piveau: A large-scale open data management platform based on semantic web technologies," in *The Semantic Web*, A. Harth, S. Kirrane, A.-C. Ngonga Ngomo, *et al.*, Eds., Cham: Springer International Publishing, 2020, pp. 648–664, ISBN: 978-3-030-49461-2.

[2] L. L. Haak, M. Fenner, L. Paglione, E. Pentz, and H. Ratner, "ORCID: a system to uniquely identify researchers," *Learn. Publ.*, vol. 25, no. 4, pp. 259–264, 2012. DOI: 10.1087/20120404. [Online]. Available: https://doi.org/10.1087/20120404.