NFDI4Energy Conference

1. NFDI4Energy Conference

DOI: 10.5281/zenodo.10610501

Published: 02.02.2024

The Leibniz Data Manager: Supporting Researchers in the Lifecycle of Research Data

Abstract for the 1. NFDI4Energy Conference

Philipp D. Rohde[1,2,3] [https://orcid.org/0000-0002-9835-4354],
Ahmad Sakor[2,3] [https://orcid.org/0000-0001-8007-7021],
Mauricio Brunet[1] [https://orcid.org/0000-0001-9576-8845],
Enrique Iglesias[2,3] [https://orcid.org/0000-0002-8734-3123],
Mazen Bechara[3] [https://orcid.org/0009-0009-7554-3935],
Susanne Arndt[1] [https://orcid.org/0000-0002-1019-9151],
Mathias Begoin[1] [https://orcid.org/0000-0003-3922-8638],
Angelina Kraft[1] [https://orcid.org/0000-0002-6454-335X], and
Maria-Esther Vidal[1,2,3] [https://orcid.org/0000-0003-1160-8727]

[1]TIB Leibniz Information Centre for Science and Technology, Hannover, Germany
[2]L3S Research Center, Hannover, Germany
[3]Leibniz University of Hannover, Hannover, Germany

Keywords: Research Data Management, Knowledge Graphs, FAIR Principles

1 Introduction and Motivation

Research Data Management (RDM) involves the systematic organization, storage, preservation, and sharing of data throughout a research project's lifecycle. Its primary goal is to ensure effective handling, maintenance, and accessibility of data, supporting the reproducibility, transparency, and integrity of research outcomes, crucial aspects for the reliability and trustworthiness of scientific research [1]. The surge in digital research data in the late 20th century necessitated standardized practices. Recognizing data's pivotal role in advancing science, funding agencies formalized policies for data sharing and management. Exemplary initiatives, such as those led by the National Institutes of Health (NIH) in the USA[1] and the National Research Data Infrastructure (NFDI) in Germany[2], underline the commitment to transparency and collaboration. These initiatives highlight the pivotal role of research data management in scientific reproducibility, transparency, and excellence, fostering collaboration across disciplines to establish legally compliant, interoperable, and sustainable data infrastructures [1], [2].
The FAIR principles (Findable, Accessible, Interoperable, Reusable) and reproducible guidelines guide the publication and exchange of research digital objects and their metadata. While research data repositories are valuable, they fall short in holistically managing scientific objects and data, hindering computational transparency over

[1] https://data.library.arizona.edu/data-management/nih-data-management-sharing-policy-2023
[2] https://www.nfdi.de/?lang=en

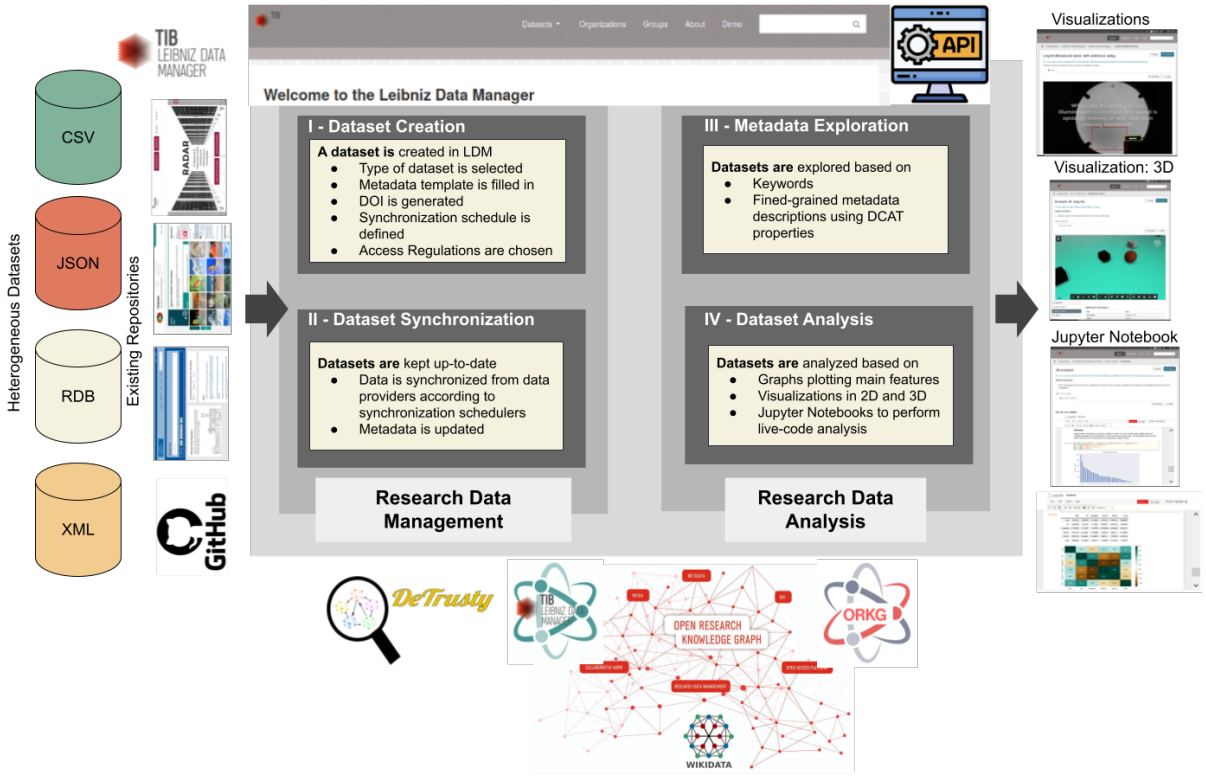

Figure 1. Overview of the Leibniz Data Manager

time. The resulting heterogeneity in data and metadata standards, APIs, file formats, licensing, archival guidelines, and more, makes searching across repositories time-consuming for researchers seeking data, necessitating a strategic approach. Geisler et al. [3] outline requirements for an effective data ecosystem, positioning knowledge-driven data ecosystems as frameworks for enhancing transparency in data exchange. Despite these efforts and relevant initiatives, a study published at Nature Index in 2019[3] reveals that scientists are not totally familiar with FAIR principles, and data management plans are complicated to apply. Similarly, a survey conducted within the German neuroscience community shed light on prevailing challenges in research data management [4]. Barriers include a lack of standardized data and metadata, limited provenance tracking, insufficient infrastructure for sensitive data, low literacy, and constrained resources. The results underscore the need for systematic development of standards, tools, and infrastructure, coupled with training and support initiatives, emphasizing the role of effective data management in facilitating data sharing in research communities.

2 The Leibniz Data Manager

The *Leibniz Data Manager* (LDM) [5] aims to address these challenges and support researchers in the lifecycle of research data management. With features designed to streamline data organization, enhance metadata management, and ensure compliance with FAIR principles, the Leibniz Data Manager serves as a resource for researchers seeking to overcome barriers in effective research data management, see Figure 1.
LDM publishes various types of research digital objects (RDO) following the FAIR principles, including datasets in different formats (e.g., CSV, JSON, or XML), data services

[3] https://www.nature.com/nature-index/news/what-scientists-need-to-know-about-fair-data

demonstrated as live code using Jupyter notebooks, and data visualizations (e.g., 2D and 3D support). The LDM dashboard allows researchers not only to manage research digital objects (e.g., collecting and searching) but also to analyze them using analytical methods implemented as Jupyter notebooks or using integrated visualization tools. Additionally, to ensure a persistent and global identification of RODs, LDM provides the option to generate a Digital Object Identifier (DOI)[4] for each uploaded ROD, making them citeable, and giving the original authors credit for publishing their data if others reuse it. LDM is implemented as an open source and extends the open data repository system CKAN[5] along with extra features developed on top of CKAN.

In addition to publishing RODs, LDM offers researchers the possibility of importing RODs already published in other repositories, such as the Leibniz University Hannover[6], PANGAEA[7], or RADAR[8]. Additionally, datasets and software published in GitHub repositories can be imported into LDM. These RODs are logically integrated into LDM, i.e., RODs are kept in the original repository, but their metadata descriptions are generated by LDM. A scheduler for synchronization allows LDM to maintain up-to-date descriptions. Metadata is expressed using existing vocabularies, e.g., DCAT[9], DataCite[10], and DublinCore[11]. SKOS[12] and the Provenance Ontology (Prov-O[13]) describe RDOs' features (e.g., definitions and labels) and their provenance, respectively.

LDM creates a knowledge graph (KG) with the metadata of the RODs published by LDM users or imported from other repositories. RODs in the LDM KG are described in terms of the properties of a DCAT resource (e.g., creator, format, size, and distribution); also, the licenses that regulate the distribution and use of RODs are described in the KG. In case an ROD is related to a scientific publication, the description of the ROD in the LDM KG is linked to the description of this publication, e.g., its fine-grained representation in the Open Research Knowledge Graph (ORKG[14]. RODs are also linked to resources in Wikidata that provide a more detailed description of an ROD. The LDM KG is accessible via a SPARQL endpoint[15], and DeTrusty [6], a federated query engine, allows for the execution of queries over the LDM, ORKG, and Wikidata KGs[16].

LDM is publicly available[17] at the TIB - Leibniz Information Center for Science and Technology in Hannover[18]. Additionally, LDM is also published as a Docker container to facilitate installing LDM distributions[19]. The main features of LDM are demonstrated in an open accessible demo[20].

[4] https://www.doi.org/
[5] https://ckan.org/
[6] https://data.uni-hannover.de/
[7] https://www.pangaea.de/
[8] https://www.radar-service.eu/radar/en/home
[9] https://www.w3.org/TR/vocab-dcat-2/
[10] https://schema.datacite.org/
[11] https://www.dublincore.org/
[12] https://www.w3.org/TR/2008/WD-skos-reference-20080829/skos.html
[13] https://www.w3.org/TR/prov-o/
[14] https://orkg.org/
[15] https://labs.tib.eu/sdm/ldm_kg/sparql
[16] https://labs.tib.eu/sdm/ldm_federated/sparql
[17] https://service.tib.eu/ldmservice/
[18] https://www.tib.eu/en/research-development/scientific-data-management/
[19] https://github.com/SDM-TIB/LDM_Docker/
[20] https://service.tib.eu/ldmservice/demo

3 Conclusions and Future Directions

LDM emerges as a timely framework, contributing to ongoing initiatives focused on improving transparency, reproducibility, and collaboration in the dynamic field of Research Data Management. By integrating FAIR principles and Semantic Web technologies, such as vocabularies, KGs, SPARQL endpoints, and query engines, LDM stands out as a key player in addressing current challenges and fostering a more efficient research data ecosystem. These specific features of LDM underscore the critical role that FAIR principles and Semantic Web technologies play in addressing the complexities of research data management. Looking ahead, our research plan includes the definition of hybrid AI systems that enhance the connections between LDM, ORKG, and Wikidata KGs. This will provide fine-grained and detailed metadata of scholarly resources to better support the research data lifecycle.

Author contributions

Philipp D. Rohde: Software, Validation, Investigation, Resources, Writing - Review & Editing, Supervision
Ahmad Sakor: Methodology, Software, Validation, Investigation, Resources, Writing - Review & Editing, Supervision
Mauricio Brunet: Methodology, Software, Validation, Investigation, Resources, Writing - Review & Editing
Enrique Iglesias: Software, Validation, Investigation, Resources, Writing - Review & Editing
Mazen Bechara: Software, Validation, Investigation, Writing - Review & Editing
Susanne Arndt: Investigation, Writing - Review & Editing
Mathias Begoin: Conceptualization, Investigation, Writing - Review & Editing, Project administration, Funding acquisition
Angelina Kraft: Conceptualization, Investigation, Writing - Review & Editing, Supervision, Project administration, Funding acquisition
Maria-Esther Vidal: Conceptualization, Methodology, Investigation, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration, Funding acquisition

Competing interests

The authors declare that they have no competing interests.

Funding

The project is funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in the LIS Funding Programme *e-Research Technologies* (grant no. 438302423) and the Nationale Forschungsdateninfrastruktur (National Research Data Infrastructure) project *NFDI4Energy* (grant no. 501865131). Federal Ministry for Economic Affairs and Energy of Germany (BMWK) in the project CoyPu (project number 01MK21007[A-L]). Leibniz Association, program ”Leibniz Best Minds: Programme for Women Professors”, project TrustKG-Transforming Data in Trustable Insights; Grant P99/2020.

References

[1] S. Kanza and N. J. Knight, "Behind every great research project is great data management," *BMC Res Notes*, vol. 15, no. 1, Jan. 2022. DOI: 10.1186/s13104-022-05908-5.

[2] C. L. Borgman and P. E. Bourne, "Why it takes a village to manage and share data," *Harvard Data Science Review*, 2022. DOI: 10.1162/99608f92.42eec111.

[3] S. Geisler, M. Vidal, C. Cappiello, *et al.*, "Knowledge-driven data ecosystems toward data transparency," *ACM J. Data Inf. Qual.*, vol. 14, no. 1, 3:1–3:12, 2022. DOI: 10.1145/3467022.

[4] C. M. Klingner, M. Denker, S. Grün, *et al.*, "Research data management and data sharing for reproducible research—results of a community survey of the german national research data infrastructure initiative neuroscience," *eneuro*, vol. 10, no. 2, ENEURO.0215–22.2023, 2023. DOI: 10.1523/eneuro.0215-22.2023.

[5] A. Beer, M. Brunet, V. Srivastava, and M.-E. Vidal, "Leibniz Data Manager – A Research Data Management System," in *The Semantic Web: ESWC 2022 Satellite Events*, Cham: Springer, 2022, pp. 73–77. DOI: 10.1007/978-3-031-11609-4_14.

[6] P. D. Rohde, "SHACL Constraint Validation during SPARQL Query Processing," in *Proceedings of the VLDB 2021 PhD Workshop, co-located with the 47th International Conference on Very Large Databases (VLDB 2021)*, Aachen, Germany: CEUR-WS.org, 2021. [Online]. Available: http://ceur-ws.org/Vol-2971/paper05.pdf.