Leibniz Data Manager – Data Management Across Various Research Data Repositories

Angelina Kraft [1], Anna Beer [2], Mauricio Brunet [1], Ahmad Sakor [1], Maria-Esther Vidal [1,3]

[1]TIB – Leibniz Information Centre for Science and Technology;
[2]University of Hildesheim, University Library;
[3]Leibniz University Hannover & L3S Research Center

The Leibniz Data Manager (LDM) is a research data management system that resorts to Semantic Web technologies to empower FAIR principles. LDM supports searching and exploring research data across various digital repositories by providing a (meta-)data management layer for digital sources based on the web-based data catalog software CKAN (Comprehensive Knowledge Archive Network). The LDM allows users to preview research data, e.g., tables, audio-visual material like AutoCAD files or 2D and 3D data, or live programming code via Jupyter Notebook(s) so that their potential for re-use can be easily evaluated. LDM is available as a Docker container, enabling the installation of local LDM distributions to assist research data management in different phases of the research data life cycle. LDM is publicly accessible at `https://service.tib.eu/ldmservice`.

1 Introduction

The FAIR principles (Findable, Accessible, Interoperable, Reusable; Wilkinson et al. 2016; Hodson et al. 2018) and reproducible guidelines aim at advising the publication of scientific digital objects. Additionally, research data repositories and code repositories provide the basis for supporting researchers during publication and validation processes. In this process, research data repositories (i.e., listed via re3data.org) offer the possibility to explore published digital objects.

However, their scope does not allow for holistic management of scientific objects and research data so that computational transparency is endured over time. The data ecosystem of research data repositories consists of various available categories and types: Discipline-specific repositories, interdisciplinary repositories, institutional repositories, and mixtures thereof. With this heterogeneity comes large variations in data and metadata standards, APIs, file formats, license information, archival and publication guidelines, terms of re-use, and others. This is also the reason why a search across multiple repositories is considered

Publiziert in: Vincent Heuveline, Nina Bisheh und Philipp Kling (Hg.): E-Science-Tage 2023. Empower Your Research – Preserve Your Data. Heidelberg: heiBOOKS, 2023. DOI: `https://doi.org/10.11588/heibooks.1288.c18084`

Table 1: Key LDM features in the Research Data Life Cycle.

Steps in the Data Life Cycle	LDM features
Data collection	Data selection from repositories, metadata is mapped
Data curation	Definition of various access privileges
Data analysis	E.g. Jupyter Notebook(s) enable the execution of live code
Data re-use	Data visualizations enable preview without download

a time-consuming task to be carried out by researchers who want to re-use data but are still determining where to look for it. Geisler et al. (2021) discuss a list of requirements to be met in a data ecosystem. Knowledge-driven data ecosystems are positioned as frameworks for enhancing transparency in data exchange.

The Leibniz Data Manager (LDM) provides a tool that can aid in the transition from a publication- or article-based to an information-based (linked-data) research workflow (Table 1). This happens by further developing a CKAN-based software distribution that allows indexing metadata and data across digital repositories, using the existing vocabularies like the Data Catalog Vocabulary (DCAT) to map metadata standards. Currently, LDM connects three pilot repositories as a proof-of-concept. With this technology in place, an intuitive user interface allows performing a data search for relevant and related data sets across the connected repositories, screening for relevant data and ultimately taking another step towards the reproducibility of science.

2 Architecture

The Leibniz Data Manager (Figure 1) solves interoperability across repositories and integrates data sets published in other repositories. The LDM is available as an open source software distribution since April 2021.[1] An up-to-date user documentation and maintenance documentation is available and linked at the homepage, and requests for functionality updates of the LDM are managed via GitHub.

LDM offers a simple, small-scale, and open software distribution which can connect digital repositories in a way such that data sets and other digital scientific objects will stay in their respective repositories, while LDM provides an integrated view of the metadata. Figure 1 depicts the main components for research data management and analysis (modified after Beer et al. 2022b). Data collections are derived from data sets in heterogeneous formats; also, data catalogs can be integrated from existing repositories (e.g., the research data repository of the Leibniz University Hanover). Metadata describing the data set is collected from the data provider; and existing vocabularies, e.g., DCAT (Albertoni et al. 2023) and DataCite (DataCite Metadata Working Group 2021) are utilized to describe the metadata. A newly created data collection is uniquely and persistently

1 `https://service.tib.eu/ldmservice`

identified by generating a Digital Object Identifier (DOI). The user can define a scheduler for synchronizing the data collection with the other data set providers (Chamanara et al. 2019). Lastly, a user can describe the access regulations. Once a data set is part of the LDM catalog, data and metadata are created and synchronized according to the schedule defined during the data creation step.

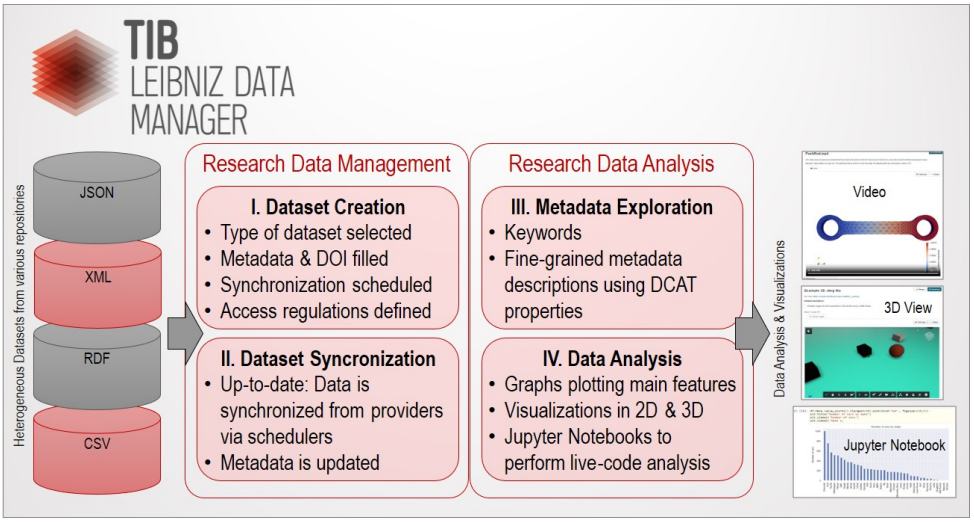

Figure 1: The Leibniz Data Manager: An adaptive RDM system which supports data search, exploration, analysis and visualization across various research data repositories.

With the LDM other information infrastructure providers may offer data "showcases" by semantically connecting existing data catalogs and repositories (Beer et al. 2022a). This enables researchers to find and analyze published digital research objects that may be of relevance for their own research. LDM offers Web APIs for traversing the LDM catalog and uploading new digital objects. Moreover, it enables the publication of data services and live code over LDM data sets. Following Linked Data principles, the LDM catalog is modeled as an RDF knowledge graph. A SPARQL endpoint allows for querying the RDF factual statements of all LDM digital objects.

Another main use of LDM is its application as a data management training tool for researchers. As a small-scale software distribution, LDM is easily installed at a local computer for training purposes. Demonstrating data sets are also available for this purpose. In 2023, LDM-training sessions were created and implemented within Master and PhD courses provided by the Leibniz University Hanover. An example are the courses "Responsible Research Data Management" in the doctoral program BIOMEDAS (BIOMEDAS Management 2023). BIOMEDAS was developed as a cross-university PhD program within the Academy of the Translational Alliance of Lower Saxony (TRAIN).

3 Support

TIB operates the hosting and long-term availability of LDM, which includes the data collection and analysis service and the availability of code and documentation on GitHub.

This also includes server infrastructure, storage, and personnel for server and application administration.

4 Conclusion and Outlook

The Leibniz Data Manager is a research data management system that supports searching and exploring research data across various repositories. LDM follows the FAIR data principles and resorts to standard vocabularies to represent metadata about research digital objects. Furthermore, LDM features a (meta-)data management layer for digital objects, which enables researchers to preview research data, e.g., tables, audio-visual material like AutoCAD files or 2D and 3D data, or live programming code via Jupyter Notebook(s) so that their potential for reuse can be easily evaluated. In a next step, an LDM extension with the most promising FAIR developments and recommendations – e.g., indexing of FAIR Digital Objects (Smedt, Koureas, and Wittenburg 2020) and FAIR Signposting (FAIRsharing Team 2023) – is planned. This will contribute to a better machine-actionability and semantic interoperability of the LDM from a data provider and data consumer point of view towards a knowledge-driven ecosystem of research data.

Acknowledgements

The project is funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in the LIS Funding Programme *e-Research Technologies* (grant no. 438302-423). This article is based on the accepted poster for the E-Science-Tage 2023 (DOI: `https://doi.org/10.11588/heidok.00033144`).

ORCID:

- Angelina Kraft https://orcid.org/0000-0002-6454-335X
- Anna Beer https://orcid.org/0000-0002-3447-0575
- Mauricio Brunet https://orcid.org/0000-0001-9576-8845
- Ahmad Sakor https://orcid.org/0000-0001-8007-7021
- Maria-Esther Vidal https://orcid.org/0000-0003-1160-8727

References

Albertoni, Riccardo, David Browning, Simon Cox, Alejandra Gonzalez Beltran, Andrea Perego, and Peter Winstanley. 2023. "Data Catalog Vocabulary (DCAT) – Version 2". `https://www.w3.org/TR/vocab-dcat-2/`.

Beer, Anna, Mauricio Brunet, Vibhav Srivastava, and Maria-Esther Vidal. 2022a. *Dataset: LDM Demo.* Technical report. Leibniz Information Centre For Science and Technology University Library (TIB). DOI: https://doi.org/10.57702/wb5bok01.

———. 2022b. "Leibniz Data Manager – A Research Data Management System". In *The Semantic Web: ESWC 2022 Satellite Events,* 73–77. Springer International Publishing. DOI: https://doi.org/10.1007/978-3-031-11609-4_14.

BIOMEDAS Management. 2023. "Translation Alliance Lower Saxony (TRAIN)". Visited on April 27, 2022. https://translationsallianz.de/train-academy/biomedas.

Chamanara, Javad, Angelina Kraft, Sören Auer, and Oliver Koepler. 2019. "Towards Semantic Integration of Federated Research Data". *Datenbank-Spektrum* 19 (2): 87–94. DOI: https://doi.org/10.1007/s13222-019-00315-w.

DataCite Metadata Working Group. 2021. "DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs v4.4". DOI: https://doi.org/10.14454/3W3Z-SA82. https://schema.datacite.org/meta/kernel-4.4/.

FAIRsharing Team. 2023. *FAIR Signposting.* DOI: https://doi.org/10.25504/FAIRsharing.d7622d.

Geisler, Sandra, Maria-Esther Vidal, Cinzia Cappiello, Bernadette Farias Lóscio, Avigdor Gal, Matthias Jarke, Maurizio Lenzerini, et al. 2021. "Knowledge-Driven Data Ecosystems Toward Data Transparency". *Journal of Data and Information Quality* 14 (1): 1–12. DOI: https://doi.org/10.1145/3467022.

Hodson, Simon, Sarah Jones, Sandra Collins, Françoise Genova, Natalie Harrower, Leif Laaksonen, Daniel Mietchen, Rūta Petrauskaité, and Peter Wittenburg. 2018. *Turning FAIR data into reality: interim report from the European Commission Expert Group on FAIR data.* Technical report. European Commission Expert Group on FAIR data. DOI: https://doi.org/10.5281/zenodo.1285272.

Smedt, Koenraad de, Dimitris Koureas, and Peter Wittenburg. 2020. "FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units". PII: publications8020021, *Publications* 8 (2): 21. DOI: https://doi.org/10.3390/publications8020021.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersbergand, Gabrielle Appleton, Myles Axtonand, Arie Baakand, Niklas Blombergand, et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific data* 3 (1): 1–9. DOI: https://doi.org/10.1038/sdata.2016.18.