



Leibniz Data Manager – A Research Data Management System

Anna Beer¹ , Mauricio Brunet¹ , Vibhav Srivastava¹,
and Maria-Esther Vidal^{1,2} 

¹ TIB - Leibniz Information Centre for Science and Technology, Hannover, Germany
{Anna.Beer,Mauricio.Brunet,Vibhav.Srivastava,Maria.Vidal}@tib.eu
² Leibniz University Hannover, Hannover, Germany

Abstract. FAIR principles aim to enhance machine-actionability of research data management, and enable data consumers and providers to scale up to incoming data avalanches. This demo paper describes Leibniz Data Manager (LDM), a research data management repository that resorts to Semantic Web technologies to empower FAIR principles. During the demonstration, the attendees will create various digital objects, and observe the crucial role of metadata in efficient and effective management and analysis of research data management. LDM is publicly available: <https://service.tib.eu/ldmservice/>.

Keywords: Research data management · RDF · FAIR principles

1 Introduction

FAIR data principles emphasize the crucial role of machine-processable metadata to find, access, interoperate, and reuse data with minimal human intervention [3]. Leibniz Data Manager is built on Semantic Web technologies to support researchers in documenting, analyzing, and sharing research datasets. LDM solves interoperability across repositories and integrates datasets published in other repositories. To present dataset metadata, it relies on existing vocabularies, e.g., DCAT¹ and DataCite². Also, data services implemented as Jupyter notebooks³ enable the execution of live code over LDM repositories. The definition of various access privileges facilitates the access and management of the LDM datasets and data services. Lastly, a wide variety of available data visualizations enables the preview of the main characteristics of a dataset without downloading it. This demo demonstrates the LDM features in the whole lifecycle of research data management [2]. First, attendees will collect and describe a dataset, and generate a Digital Object Identifier (DOI)⁴ that will persistently and globally

¹ <https://www.w3.org/TR/vocab-dcat-2/>.

² <https://schema.datacite.org/>.

³ <https://jupyter.org/>.

⁴ <https://www.doi.org/>.

identify their datasets. Next, they will explore metadata, describing the defined datasets, in various RDF serializations. Previews of the uploaded data will be visualized using a myriad of plots. Jupyter notebooks will be included as data services to demonstrate on-the-flight analyses. Lastly, datasets from other data repositories or data providers will be integrated; the attendees will be able to set up different synchronization schedules to keep datasets up to date.

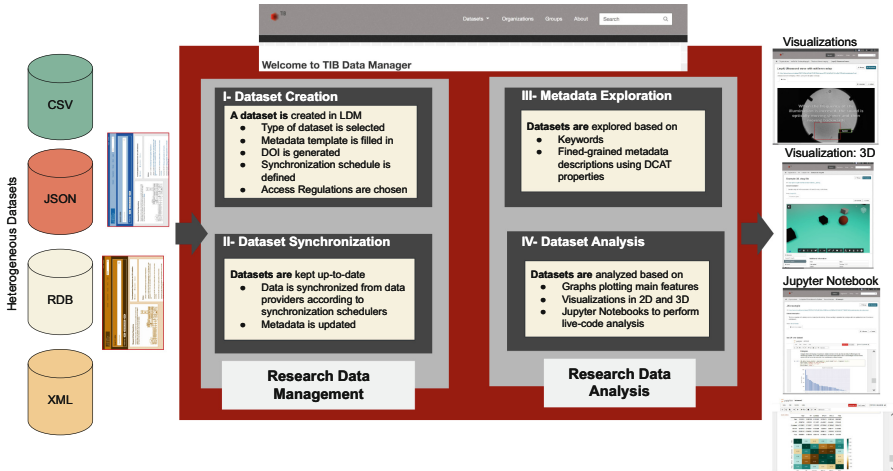


Fig. 1. The Leibniz Data Manager main components.

2 The Leibniz Data Manager Architecture

Leibniz Data Manager aims at supporting the lifecycle of research data management: a) Planning research; b) Collecting data; c) Processing and Analyzing data; d) Publishing and Sharing; e) Preserving data, and f) Re-using data. Figure 1 depicts the main components for research data management and analysis. Data is collected from datasets in heterogeneous formats; also, data catalogs can be integrated from existing repositories (e.g., the data repository of the Leibniz University Hannover⁵). Metadata describing a dataset is collected from the data provider; and existing vocabularies, e.g., DCAT and DataCite are utilized to describe the metadata following the Linked Data⁶ and FAIR principles. The newly created dataset is uniquely and persistently identified by generating a DOI. Moreover, the user can define a scheduler for synchronizing the dataset with the other dataset providers [1]. Lastly, the user can describe the dataset access regulations. Once a dataset is part of the LDM catalog, data and metadata are created and synchronized according to the schedule defined during the data

⁵ <https://data.uni-hannover.de/>.

⁶ <https://www.w3.org/wiki/LinkedData>.

creation step. At the analysis level, LDM enables users to explore the datasets based on keyword queries or searches defined on DCAT properties (e.g., object types, formats, licenses). Metadata is presented in various RDF serializations and described using DCAT or DataCite. Datasets can be explored using multiple plots or visualized in 2D or 3D. Lastly, data services allow for the analysis of datasets via the use of interactive programming via Jupyter notebooks. LDM is implemented as an open source and extends the open data repository system CKAN⁷ along with extra features developed on top of CKAN extensions, e.g., ckanext-dcat⁸. LDM is available as a Docker container to facilitate installing LDM distributions⁹. LDM is a publicly available resource maintained by the TIB – Leibniz Information Center for Science and Technology in Hannover¹⁰. TIB is one of the largest libraries for Science and Technology in the world¹¹, and actively promotes open access to digital research artifacts, e.g., research data, scientific literature, non-textual material, and software. Similar to other TIB services, LDM is regularly maintained and supported.

The screenshot shows the LDM dataset creation interface. It is divided into several sections:

- Dataset Metadata:** Includes fields for Title (e.g., "A descriptive title"), Description (e.g., "Some useful notes about the data"), Type (e.g., "economy, mental health, government"), Licenses (License not specified), Organization (TIB), Visibility (Private), Source (http://example.com/dataset.json), and Version (1.0).
- Unique Author Identifier:** Includes fields for Author (Joe Blogg) and ORCID (e.g., "0000-0002-1025-0987").
- Dataset Services:** Includes a "Data Services selection tool" with options to "Show by Organization" (TIB) and "Select Data Services".
- Annotations:** Includes a "Custom Fields" section with a "Custom Fields 1" field and a "Remove" button.

Fig. 2. Dataset creation step. Metadata is collected to describe datasets, licenses, authors, data services, and semantic annotations.

3 Demonstration of Use Cases

The demonstration aims at illustrating the LDM main functionalities and the support provided in each of the steps of the research data management lifecycle. During the demonstration, the attendees will be able to interact with LDM, and experiment the tasks of dataset creation and management, and dataset analysis.

⁷ <https://ckan.org/>.

⁸ <https://github.com/ckan/ckanext-dcat>.

⁹ https://github.com/SDM-TIB/LDM_Docker/.

¹⁰ <https://www.tib.eu/en/research-development/scientific-data-management/>.

¹¹ <https://www.tib.eu/en/tib/profile/>.

Social	@prefix spdx: <http://spdx.org/rdf/terms#> . @prefix time: <http://www.w3.org/2006/time#> . @prefix vcard: <http://www.w3.org/2006/vcard/ns#> . @prefix xml: <http://www.w3.org/XML/1998/namespace> . @prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
Twitter	
Facebook	
License	<https://service.tib.eu/ldmservice/dataset/f45c42a7-50d2-46e1-a247-e40db3255d50> a dcat:Service ; dct:description "A demo for the Falcon approach" ; dct:identifier "f45c42a7-50d2-46e1-a247-e40db3255d50" ; dct:issued "2022-01-31T15:37:16.684571"^^xsd:dateTime ; dct:modified "2022-02-24T09:11:39.255308"^^xsd:dateTime ; dct:publisher <https://service.tib.eu/ldmservice/organization/0c5362f5-b99e-41db-8256-3d0d754f> ; dct:title "Falcon Demo" ; dct:contactPoint [a vcard:Organization ; vcard:fn "Ahmad Sakor"] ; dcat:distribution <https://service.tib.eu/ldmservice/dataset/f45c42a7-50d2-46e1-a247-e40db3255d50/resource/> ; <https://service.tib.eu/ldmservice/dataset/f45c42a7-50d2-46e1-a247-e40db3255d50/resource/> <https://service.tib.eu/ldmservice/dataset/f45c42a7-50d2-46e1-a247-e40db3255d50/resource/> ; <https://service.tib.eu/ldmservice/dataset/f45c42a7-50d2-46e1-a247-e40db3255d50/resource/> <https://service.tib.eu/ldmservice/dataset/f45c42a7-50d2-46e1-a247-e40db3255d50/resource/> .
License not specified	<https://service.tib.eu/ldmservice/dataset/f45c42a7-50d2-46e1-a247-e40db3255d50/resource/1b232a7b- dct:description ""A dataset of 100 drugs from DrugBank\r the DrugBank Id and the label for each drug is provided"" ; dct:format "CSV" ; dct:title "Drugs Dataset" ; dcat:accessURL <https://service.tib.eu/ldmservice/dataset/f45c42a7-50d2-46e1-a247-e40db3255d50/> ; dct:byteSize 4863.0 ; dcat:mediaType "text/csv" .
Export	<https://service.tib.eu/ldmservice/dataset/f45c42a7-50d2-46e1-a247-e40db3255d50/resource/7dbdf1e- dct:format "py" ; dct:title "Falcon Functions" ; dcat:accessURL <https://service.tib.eu/ldmservice/dataset/f45c42a7-50d2-46e1-a247-e40db3255d50/> ; dct:byteSize 4863.0 ; dcat:mediaType "text/x-python" .
DCAT(rdf/xml) DCAT(xml) DCAT(N3) DCAT(ttl) DCAT(jsonld) DataCite CSL DublinCore BibTex	

Fig. 3. Semantically describing datasets.

3.1 Dataset Creation and Management

Attendees will go through dataset creation and specify metadata that characterizes the defined dataset; it includes title, description, tags, and license. Additionally, the dataset authors can be uniquely identified using their ORCID¹² identifiers. Similarly, attendees will define data services for the datasets, and use controlled vocabularies, to express the meaning of the published data. Figure 2 illustrates the part of the interface used to collect this metadata and create a dataset. Attendees will create two types of datasets, i.e., local and imported from other repositories. Different schedulers for data synchronization will be defined. They will explore metadata in various vocabularies, e.g., DataCite, DCAT, or DublinCore, to analyze machine-readable descriptions of the defined datasets (Fig. 3). Furthermore, attendees will be able to explore and search the datasets based on metadata represented using these vocabularies. Different schedulers for data synchronization will allow for LDM adaptability and synchronization.

3.2 Dataset Analysis

Three types of datasets are available: (i) Local datasets, including data resources presented in various formats (e.g., CSV, JSON, text, or MP4). (ii) Imported datasets collected from existing data repositories on the Web. (iii) Data services running on top of datasets and providing data processing results in the form of non-alterable Python code. The attendees will publish these three different types of data resources (Fig. 4) and analyze main properties using live code implemented as Jupyter notebook services¹³.

¹² <https://orcid.org/>.

¹³ <https://service.tib.eu/ldmservice/service/>.

Service Example

This is an example of a dummy service created just for testing. We are working to populate the prototype with new updated and real life services and datasets.

Data and Resources

- ▶ Text Explore
- ▶ car data Explore
- ▶ JN example Explore
- ▶ testing big file Explore

Cite this as

Brunel Mauricio (2021). Dataset: Service Example. <https://doi.org/10.57702/onhcz285>
DOI/retrieved: March 2, 2022

test JN over dataset

jupyter (autosaved)

View Cell Kernel Help

⌂ ↻ ↵ ⬆ ⬇ ▶ Run ■ C ⏪ ⏩ Markdown ▾

This is the reason in the above step while counting both Cylinders rows.

```
In [21]: df = df.dropna() # Dropping the missing values.
df.count()
```

Out[21]:	Make	10827
	Model	10827
	Year	10827
	HP	10827
	Cylinders	10827
	Transmission	10827
	Drive Mode	10827
	MPG-H	10827
	MPG-C	10827
	Price	10827
	dtype:	int64

Now we have removed all the rows which contain the Null or N/A v

```
In [22]: print(df.isnull().sum()) # After dropping the value
```




Fig. 4. Jupyter notebook integrated over dataset for live code analysis.

4 Conclusions

We demonstrate a data management system for supporting the lifecycle of research data management, i.e., data creation, documentation, analysis, preservation, and sharing. Datasets from other repositories can be imported and maintained up to date. The LDM demo puts in perspective the crucial role of Semantic Web technologies, and W3C recommended vocabularies, in the generation of machine-readable metadata respecting Linked Data and FAIR principles.

Acknowledgements. The project is funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in the LIS Funding Programme *e-Research Technologies* (grant no. 438302423).

References

1. Chamanara, J., Kraft, A., Auer, S., Koepler, O.: Towards semantic integration of federated research data. *Datenbank-Spektrum* **19**(2), 87–94 (2019). <https://doi.org/10.1007/s13222-019-00315-w>
2. Mosconi, G., et al.: Three gaps in opening science. *Comput. Support. Coop. Work* **28**(3–4), 749–789 (2019). <https://doi.org/10.1007/s10606-019-09354-z>
3. Wilkinson, M., et al.: The fair guiding principles for scientific data management and stewardship. *Sci. Data* **3**(1), 1–9 (2016). <https://doi.org/10.1038/sdata.2016.18>