

# Databus and MOSS – Search over a flexible, multi-domain, multi-repository metadata catalog

Sebastian Hellmann<sup>1</sup>, Carsten Hoyer-Klick<sup>2</sup>

<sup>1</sup>Institute for Applied Informatics (InfAI) at Leipzig University,  
hellmann@informatik.uni-leipzig.de

<sup>2</sup>German Aerospace Center (DLR), Institute of networked Energy Systems,  
carsten.hoyer-klick@dlr.de

## Abstract

Databus and MOSS provide an advanced, flexible cataloging system designed to meet FAIR (Findable, Accessible, Interoperable, and Reusable) [FAIR] data management standards across domains. By creating a unified metadata registry, Databus and MOSS enable researchers to add, refine, and search rich metadata records, ensuring data accessibility and interoperability across diverse repositories and research fields.

## System Capabilities and Concepts

The system enables the deployment and population of an **interdisciplinary, multi-domain catalog (using Databus)** across multiple dataset repositories relevant to a research field. This catalog, enriched with a **metadata annotation system** and a **search user interface called MOSS (Metadata Overlay Search System)**, allows researchers to

- (1) flexibly add comprehensive metadata records and
- (2) search and discover data relevant to their field and their current research.

Both Databus and MOSS are conceptually and technically mature tools that have been in active use for several years (in particular in energy system research and the OpenEnergyPlatform<sup>1</sup>). In the following, we describe the primary architecture and the four key concepts that form the foundation of this system.

Figure 1 shows the main architecture (using example repositories and metadata schemas from energy system research). The data itself stays in various types of data repositories (Virtual Bus). They each follow their own standards. Their data sets are registered on the Databus to create a PID and therefore a

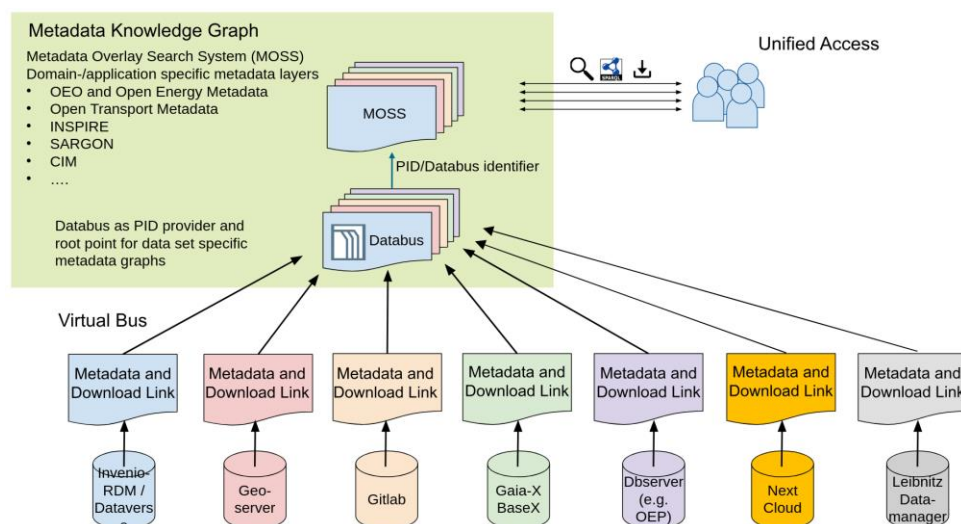


Figure 1: Architecture of databus and MOSS.

<sup>1</sup><https://openenergyplatform.org/>

root of a metadata graph. The MOSS is used to attach multiple metadata subgraphs to each of the PIDs. Essential metadata is available as SPARQL and additionally each subgraph is indexed in a search index within the MOSS. The user mainly interacts with the MOSS. When searching for data, the user will get the available metadata and a download link to the data in the respective repository of the data.

### *Multi-Repository Catalog (Databus)*

Currently, research data is already published across a wide range of online portals, repositories, platforms and websites. One of the main goals of research data management is to adequately catalog this existing, distributed data in a „registry for metadata for DOs“<sup>2</sup>. We have identified a number of typical repositories and repository software which are currently used in energy system research:

- (1) Zenodo as a publicly available platform,
- (2) Repositories based on InvenioRDM, the open-source software used in Zenodo, that can be deployed at institutions (e.g. it is used by DLR) or other tools as Harvard Dataverse,
- (3) Git-based repositories such as the public GitHub platform by Microsoft or the self-deployed GitLab,
- (4) Geoserver, a custom landing page for energy-related geo data,
- (5) Open Energy Platform and its databases,
- (6) Cloud-based repositories such as NextCloud and Google Drive,
- (7) CKAN-based repositories such as the Leibniz Data Manager (LDM).

Over the last 8 years, starting with the work on DataID [DataID], we investigated the capabilities of these repositories and many others in order to create a unified approach to catalog the contained data under FAIR aspects. Development on the Databus catalog system started around 2018. Although all the above repositories share the same goal of providing online access and metadata, heterogeneity is extremely high in the details. Each of the systems implements PIDs, data lifecycle, metadata, API access and low-level data handling in a different way. In previous projects, we have successfully shown, that the Databus Ontology is able to provide a minimal, homogeneous model to accurately capture and integrate the various models and create a consistent, interoperable Metadata Knowledge Graph. Two examples of such heterogeneity are:

- (1) versioning is done in Zenodo by assigning new PIDs to each version, while git-based solutions use commit hashes, but provide stable PIDs pointing to the latest version. Other solutions like NextCloud work in overwrite mode, where the retrieved DOs for a PID change.
- (2) other incompatibilities are found on a lower level such as metadata describing the format with e.g. “csv.gz” could be described as “csv” or “gz” file, or even differences in measuring filesize, in date formats and in hashsums.

To clarify the benefits of Databus for research data management:

- (1) Databus has an established track record as a metadata registry, supporting platforms like Zenodo/InvenioRDM, HuggingFace, GitHub/GitLab, NextCloud/GoogleDrive, FTP file servers, and notably, the Open Energy Platform (see [LOD-GEOSS D6]). As the engineering for these integrations is implementation-specific, the work would have to be repeated (mostly from scratch) within research data management.
- (2) The Databus (as well as MOSS) is committed to the same open standards as the tooling of many research data management approaches: Knowledge Graphs, SPARQL, Linked Data, DCAT & DCAT-AP, SHACL, JSON-LD, RDF. Thus, it is highly interoperable and complementary to the already

---

<sup>2</sup>See Key Services / Registry <https://nfdi4energy.uol.de/sites/services/#best-practices>

developed services that rely on the same standard as namely LDM, ORKG, TIB Terminology and PID services.

### *Metadata Annotation in Energy Systems Research (MOSS)*

Energy system research is rooted in many different scientific domains and the data that is used in energy systems modeling comes from very different domains such as engineering, remote sensing, meteorology, economics, social sciences, geography among others. Documenting data is difficult as each of the above-mentioned domains use different standards and different vocabularies. Building a metadata schema which covers all the different aspects may prove to be difficult and may not always be compatible with the different domain standards.

We propose an easier way to handle multi-domain metadata by attaching multiple metadata schemata of different domains to a single data set. Our approach is consistent with the construction of specific, minimal metadata “modules” for specific purposes. This eases the discovery of the data while searching for data with different domain perspectives.

The MOSS (Metadata Overlay Search System), which has been developed during the LOD-GEOSS project [LOD-GEOSS D6], is an addition to the Databus catalog system [Databus]. The Databus provides a persistent identifier (PID) with basic metadata and this PID serves as a root to a metadata graph for each of the registered data sets. Different “layers” of metadata information can be configured in the MOSS. Each layer contains a domain-specific subgraph that can be added to describe the data resource from a multi-domain perspective. The MOSS supports the use of ontologies for a unified data vocabulary within the research domains, e.g. the Open Energy Ontology [OEO]. The layers are curated by the administrator of the MOSS, therefore the adherence to domain standards can be enforced in opposition to user extensions on the catalog which may vary according to the users using it.

MOSS is a central entry point for users managing their data residing in different repositories. Rich metadata such as the Open Energy Metadata Schema 2.0<sup>3</sup> (JSON-LD) can be used to register data, as it contains all information for a registration in the MOSS and Databus. The MOSS and Databus therefore serve as a flexible, multi-domain and multi-repository metadata catalog and an entry point for data users where they can start to catalog and describe their data. In this manner, Databus and MOSS can bridge data repositories on the data and metadata level, and enable the development of uniform data catalogs. This can consolidate all the metadata from the different data silos and data domains by making them searchable on a single platform. Databus and the current MOSS are under active development and planned to be deployed to the Open Energy Platform, at the DLR research data management system and within the Helmholtz Federated IT Services (HIFIS).

To clarify the benefits of MOSS for research data management:

- (1) MOSS allows users to annotate data sets from many repositories in a central place that can later be part of a larger platform.
- (2) Initial metadata layers will focus on existing standards such as OEO and Open Energy Metadata 2.0, new layers can be added over time including an ORKG layer or others.
- (3) Subgraphs are versioned in Git allowing traceability and versioning of metadata edits following a Wiki workflow.
- (4) Adherence to RDF and SPARQL standard makes the collected metadata in MOSS interchangeable with LDM, federated search and other Knowledge Graph approaches.

---

<sup>3</sup> Already compatible with MOSS

## Search over Multi-Layer Metadata Graphs (MOSS)

A good search is always highly contextual and geared towards the needs of the targeted audience. With MOSS, we provide a practical workaround to the No Free Lunch in Search Theorems<sup>4</sup> (cf. [Wolpert]) by enabling each user community to define and deploy its own metadata standards in layers and then configure via SPARQL queries<sup>5</sup> which parts from these metadata subgraphs are indexed in the search and exploited in the user interface via facets, plaintext/keyword search, ontology-based search with autocomplete or hierarchical browsing of ontologies and other search paradigms. This modular approach enables MOSS to avoid the inefficiencies of a one-size-fits-all solution by aligning the search index directly with the unique structure and semantic needs of each community's metadata and thus deliver optimized search quality and efficiency within their context — effectively addressing the challenge posed by the No Free Lunch theorem.

To clarify the benefits of the search in MOSS for research data management:

- (1) The successful development of a user-accepted search over multi-domain, multi-repository datasets can increase visibility and acceptance of research data management within the research community, instead of perceiving it as a burden.
- (2) The employed methods and tools aim to have an impact on the whole research data management as a search solution that can be adopted and re-used across different domains.

## Current Status

The development builds heavily on previous work from the last 6 years that originated on the one side from the LOD-GEOSS project, the Open Energy Family (OEF) and DLR internal projects and the knowledge graph community around Dbpedia on the other side.

Figure 2 on the right shows the parts of the Open Energy Family that have been successfully integrated via Databus/MOSS in LOD GEOSS (cf. [LOD GEOSS D3, LOD GEOSS D6]) namely the OEMetadata Standard, the OEOntology, the Energy Databus<sup>6</sup> and the OEDatabase.

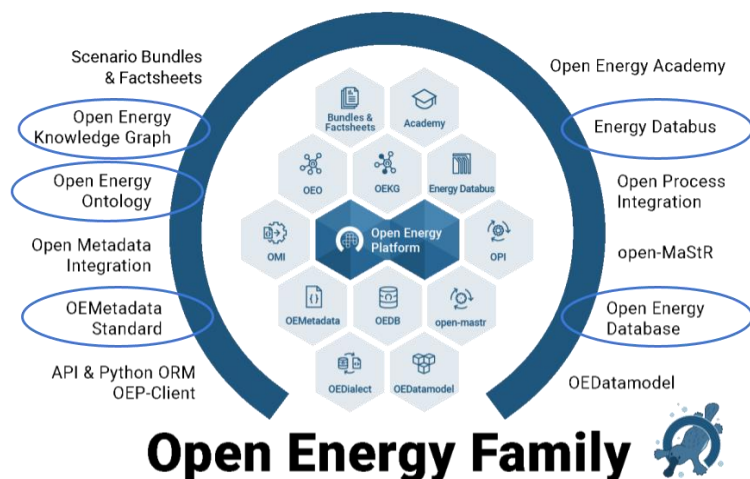


Figure 2: Open Energy Family

The Databus itself is in productive use at the Dbpedia community project<sup>7</sup> to register (1) the monthly Dbpedia Knowledge Graph releases (ca. 5000 files and 20 Billion facts, cf. [Hofer]) as well as (2) the community generated datasets. Approximately 40,000 files are downloaded via the Dbpedia Databus each day (1.2 Million downloads per month) that are distributed over 25 distinct servers and repositories (Zenodo, GitHub, Huggingface).

The Databus software<sup>8</sup> is in a mature state (version 2.1.0) and ready to deploy. The development of the MOSS backend (editing and managing subgraph layers, search index and configuration) recently completed.

<sup>4</sup> [https://en.wikipedia.org/wiki/No\\_free\\_lunch\\_in\\_search\\_and\\_optimization](https://en.wikipedia.org/wiki/No_free_lunch_in_search_and_optimization)



<sup>5</sup> The community can decide how to configure the search, which is then added to the configuration by an admin

<sup>6</sup> Open Energy Databus <https://databus.openenergyplatform.org/>

<sup>7</sup> <https://dbpedia.org> and <https://databus.dbpedia.org>

<sup>8</sup> <https://github.com/dbpedia/databus>

## References

- [FAIR] Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>
- [OEO] Boosheri, Meisam und Emele, Lukas und Flügel, Simon und Förster, Hannah und Frey, Johannes und Frey, Ulrich J und Glauer, Martin und Hastings, Janna und Hofmann, Christian und Hoyer-Klick, Carsten und Hülk, Ludwig und Kleinau, Anna und Knosala, Kevin und Kotzur, Leander und Kuckertz, Patrick und Mossakowski, Till und Muschner, Christoph und Neuhaus, Fabian und Pehl, Michaja und Robinius, Martin und Sehn, Vera und Stappel, Mirjam (2021) *Introducing the Open Energy Ontology: Enhancing data interpretation and interfacing in energy systems analysis*. Energy and AI, 5. Elsevier. doi: [10.1016/j.egyai.2021.100074](https://doi.org/10.1016/j.egyai.2021.100074) 
- [Databus] Documentation: <https://dbpedia.gitbook.io/databus>
- [DataID] Freudenberg, M., Brümmer, M., Rücknagel, J., Ulrich, R., Eckart, T., Kontokostas, D., & Hellmann, S. (2016) *The metadata ecosystem of DataID*, Metadata and Semantics Research (MTSR) [https://doi.org/10.1007/978-3-319-49157-8\\_28](https://doi.org/10.1007/978-3-319-49157-8_28) preprint [https://svn.aksw.org/papers/2016/MSOR\\_DataID2/public.pdf](https://svn.aksw.org/papers/2016/MSOR_DataID2/public.pdf)
- [LOD-GEOSS D3] Frey, J., Hoyer-Klick, C., Muschner, C., Hülk, L., Streitmatter, D., & Hellmann, S. (2023). Distributed Data Infrastructure. Project Report. doi: 10.5281/zenodo.8119055. Zenodo. <https://zenodo.org/records/8119055>
- [LOD-GEOSS D6] Hoyer-Klick, C., Frey, U., Sehn, V., Launer, J., Hülk, L., Kronshage, S., & Kuckertz, P. (2023). Demonstration and Best Practices (1.0). Zenodo. <https://doi.org/10.5281/zenodo.8119096> 
- [Hofer] Marvin Hofer, Sebastian Hellmann, Milan Dojchinovski, Johannes Frey: The New DBpedia Release Cycle: Increasing Agility and Efficiency in Knowledge Extraction Workflows. SEMANTiCS 2020: 1-18
- [Wolpert], D. H.; Macready, W. G. (1995). "No Free Lunch Theorems for Search". Technical Report SFI-TR-95-02-010. Santa Fe Institute. S2CID 12890367.