

Deliverable 8.1.2.4

REPORT OF FINAL VERSION AND ACHIEVED CHANGES FOR DATABUS AND
MOSS
30.06.2025

Report of Final version and achieved Changes for databus and MOSS

Carsten Hoyer-Klick ¹, Jan Forberg ², and Sebastian Hellmann ²

¹German Aerospace Center, Institute of Networked Energy Systems, ROR, Curiestr. 4, 70563 Stuttgart

²InfAI, Institut für Angewandte Informatik (InfAI) e. V., Leipzig University ROR Goedelerring 9, 04109 Leipzig

Published by

German Aerospace Center, Institute of Networked Energy Systems, ROR, Curiestr. 4, 70563 Stuttgart

Acknowledgements

The authors of this article have used various preparatory works from the NFDI4Energy to create this portrait, and references have been made where possible. Thanks to all those who are not named. The authors would like to thank the German Federal Government, the German State Governments, and the Joint Science Conference (GWK) for their funding and support as part of the NFDI4Energy consortium. The work was funded by the German Research Foundation (DFG) – 501865131 within the German National Research Data Infrastructure (NFDI, www.nfdi.de).

License

This document is published under the Attribution 4.0 International (CC BY 4.0). This license allows users to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator. The license allows for commercial use.



Table of Contents

General Information	3
Summary	3
Deliverable within NFDI4Energy	3
Deliverable	4
Introduction	4
System Capabilities and Concept	4
Multi-Repository Catalog (Databus)	4
Metadata Annotation for Energy Systems Research (MOSS)	5
Search over Multi-Layer Graphs (MOSS)	6
Status at the Beginning of the Project	6
Workshops	7
Workshop at the NFDI4Energy Conference March 25th, 2025	7
Searching and Filtering	7
How do you like a user interface for searching?	11
Workshop at the Forschungsnetzwerk Systemanalyse May, 6th, 2025	13
Searching and Filtering	13
How do you like a user interface for searching?	13
Workshop at Helmholtz Metadata Collaboration Conference, May 12th, 2025	15
Searching and Filtering	15
How do you like a user interface for searching?	16
Implementation	16
Admin Interface	16
General Tab	17
Template Tab	17
SHACL Tab	17
Indexer Tab	17
JSON-LD Context Tab	17
Searching data	18
Conclusions and Outlook	20

General Information

Summary

Modeling future energy systems and technologies requires large amounts of heterogeneous input data and also generates heterogeneous output data in various scientific disciplines such as meteorology, engineering, materials science, geography, social sciences, and economics. To support the transformation process toward a more sustainable society, this data should be open and transparent. The FAIR (Findable, Accessible, Interoperable, and Reusable) principles have been established for this purpose. The focused research data infrastructure consists of the Energy Databus as a comprehensive data catalog, the Open Energy Ontology as a controlled vocabulary, the Open Energy Metadata for describing research data, and the Open Energy Platform as an open database. Other repositories such as Zenodo or those from research institutions can be integrated in to the catalog as federated data sources. This deliverable describes how Databus and MOSS (Metadata Overlay Search System) can be used to organize research data catalogs across multiple data repositories and data silos to federate data sources and ease data discovery based on semantic metadata. We describe the workshops we conducted within the NFDI4Energy consortium and the resulting adaptations to Databus and MOSS to improve the usability of the tools for research data management.

Deliverable within NFDI4Energy

This report is part of Task Area 8 whose aim is to support researchers by developing and providing core services and solutions for energy research data management. Databus and MOSS are two services which support the development of data catalogs across many different repositories and scientific domains. The Databus is able to provide a minimal, homogeneous model to accurately capture and integrate various models and create a consistent, interoperable metadata knowledge graph. The Databus provides an ID for each dataset as a root of the dataset specific metadata graph. MOSS enhances this metadata graph by data and domain specific subgraphs, which get attached to the basic root graph. The tools in general are agnostic of data repositories and scientific domains, the only requirement for a repository is that it can provide a URL for the data and that the metadata is described as linked data in RDF or JSON-LD. These two tools together form a service which makes data discoverable across repositories and scientific domains. The enhancements of Databus and MOSS are integrated as measure M8.1 "Search over Multi-Layer Metadata for NFDI4Energy Repositories."

Deliverable

Introduction

DBpedia Databus[1] and MOSS provide an advanced, flexible cataloging system designed to meet FAIR (Findable, Accessible, Interoperable, and Reusable) [2] 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 Concept

The system enables the deployment and population of an interdisciplinary, multi-domain catalog (using Databus)[3] across multiple dataset repositories relevant to a research field. This catalog, enriched with a metadata annotation system and a search interface called MOSS (Metadata Overlay Search System)[4], 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 Open Energy Platform[5]). 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 schema 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 an ID and therefore a root of a metadata graph. The MOSS is used to attach multiple metadata subgraphs to each of the IDs. 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.

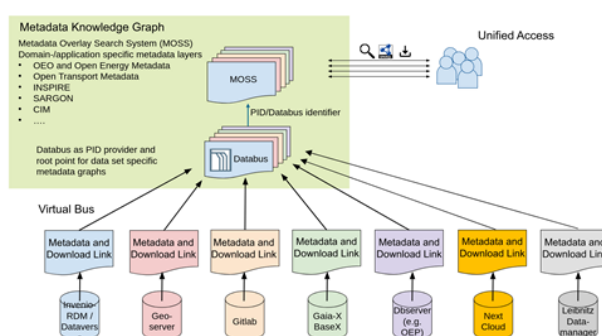


Figure 1: Architecture of Databus and MOSS

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 Digital Objects (DO)“. We have identified a number of typical repositories and repository software which are currently used in energy system research: Zenodo¹ as a publicly available platform; Repositories based on InvenioRDM², the open-source software

¹<https://zenodo.org>

²<https://inveniosoftware.org/products/rdm/>

used in Zenodo, that can be deployed at institutions (e.g. it is used by DLR) or other tools as Harvard Dataverse³; Git-based repositories such as the public GitHub platform by Microsoft or the self-deployed GitLab; Geoserver⁴, a server to view and edit geodata; Web based relational databases as the Open Energy Database (oedb)⁵ within the Open Energy Platform; Cloud-based repositories such as NextCloud⁶ and Google Drive⁷; CKAN-based repositories such as the Leibniz Data Manager (LDM)⁸.

Over the last 8 years, starting with the work on *DataId* [6], 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 IDs, data life-cycle, 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 URI to each version, while git-based solutions use commit hashes, but provide stable IDs pointing to the latest version. Other solutions like NextCloud work in overwrite mode, where the retrieved DOs for an ID change.
2. Incompatibilities are found on a lower level such as metadata describing the format with e.g. "csv.gz" as either "csv" or "gz" file, differences in measuring file size, inconsistencies in date formats and in checksums.

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 [7]). 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 and RDF. Thus, it is highly interoperable and complementary to the already developed services that rely on the same standard as namely LDM, ORKG, TIB Terminology and PID services.

Metadata Annotation for 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 [7], is an

³<https://dataverse.org/>

⁴<https://geoserver.org/>

⁵<https://openenergyplatform.org/database/>

⁶<https://nextcloud.com/>

⁷<https://drive.google.com>

⁸<https://projects.tib.eu/datamanager/>

addition to the Databus catalog system [1]. The Databus provides an identifier (ID) with basic metadata and this ID serves as a root to a metadata graph for each of the registered data sets. Different “modules” of metadata information can be configured in the MOSS. Each module 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 [8]. The modules 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 (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.

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 modules will focus on existing standards such as OEO and Open Energy Metadata 2.0, new modules can be added over time including an ORKG module 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.

Search over Multi-Layer 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 (cf. [9]) by enabling each user community to define and deploy its own metadata standards in modules and then configure via SPARQL queries 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 autocompletion 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.

Status at the Beginning of the Project

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 shows the parts of the Open Energy

Family that have been successfully integrated via Databus/MOSS in LOD GEOSS (cf. [7], [10]) namely the OEMetadata Standard, the Open Energy Ontology, the Energy Databus and the OEDatabase. The Databus itself is in productive use at the DBpedia community project to register (1) the monthly DBpedia Knowledge Graph releases (ca. 5000 files and 20 Billion facts, cf. [11]) 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 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.

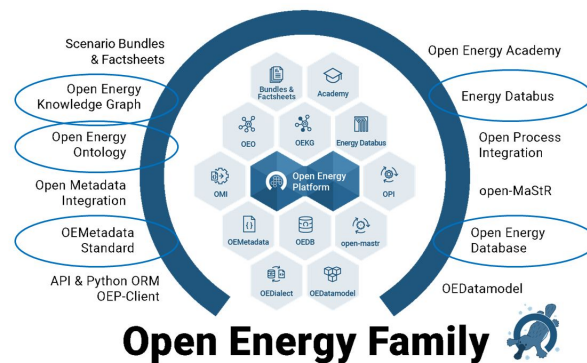


Figure 2: Databus and MOSS related elements of the Open Energy Family.

Workshops

In order to analyze ideas and demands of the users, we organized three workshops along three events in spring of 2025.

Workshop at the NFDI4Energy Conference March 25th, 2025

The first workshop took place as part of the NFDI4Energy Conference⁹ in Karlsruhe on March 25th, 2025. After a presentation of the current status of databus and MOSS, the workshop participants discussed two sets of questions:

- What are typical keywords you would use for searching? How would you like to filter larger numbers of search results?
- How do you like a user interface for searching? What feature are important.

The discussion on the first questions was more focused on the possible filtering techniques.

Searching and Filtering

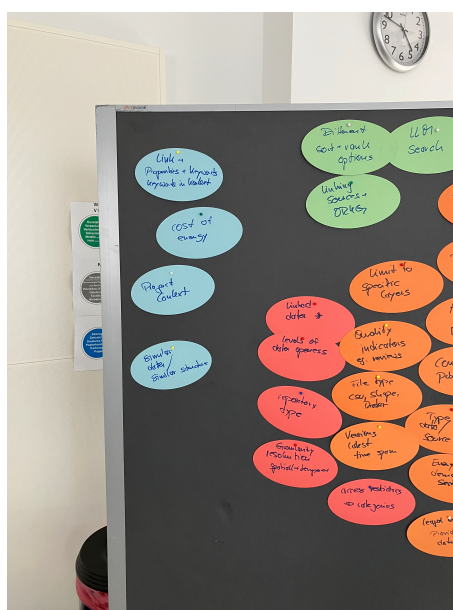
Ideas for searching:

- Link properties and keywords, enable to put search keywords in context, e.g. only search for certain keywords in `subject:` or in `is_about:` keys.
- Project context, e.g. searching for (more) data from the same project.
- Searching for similar data in terms of similar content or similar structure (e.g. for comparisons)

Ideas for features:

⁹<https://nfdi4energy.uol.de/sites/conference/2025/>

- Different options for sorting and ranking of the results
- Linking the sources of data to the ORKG
- Search based on Large Language models



Ideas for filters:

- Time ranges for temporal information in the metadata, either of the data or data creation.
- Spatial context, by selecting regions, locations or bounding boxes.
- Project context (limit results only to certain project contexts)
- Type of data, e.g. real measured data vs. modeled, vs. synthetic, source of data
- Limit to specific MOSS layers
- Complex search Queries with AND / OR / NOT ...
- Limit to voltage levels
- Quality indicators, e.g. OEMetadata badges, user ratings, levels of data openness, reviews
- Limit to certain contributors, publishers,
- Legal entities providing the data, e.g. only public agencies
- Type of energy
- File type, e.g. csv, shape, raster,
- Limit to e.g. latest version or versions from a certain time span
- Energy demand sectors
- Create and modification date
- Limit to source publications, only show data linked to certain publications
- License of the data
- Repository type, e.g. only Open Energy Platform or Zenodo.
- Granularity and Resolution
- Categories of access restriction, e.g. not able to overcome, can be overcome by a fee or signing an NDA.

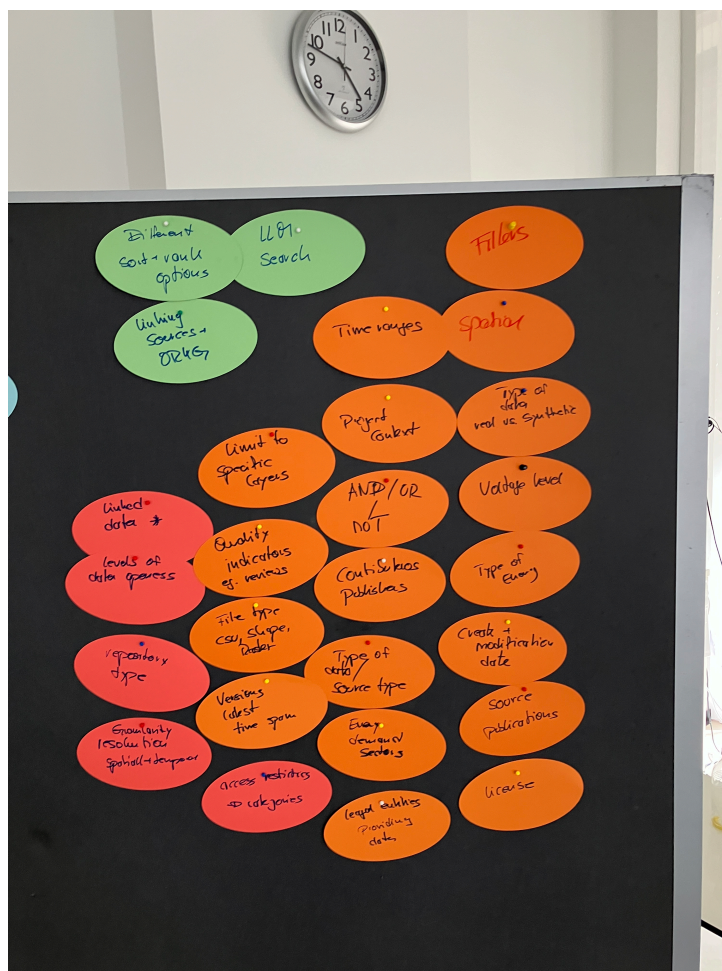


Figure 5: Collected cards for searching and filtering part 2

How do you like a user interface for searching?

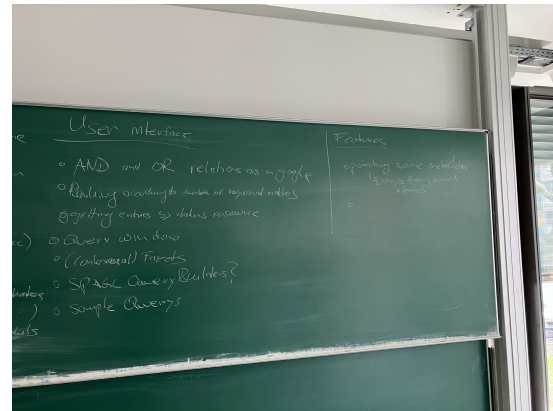
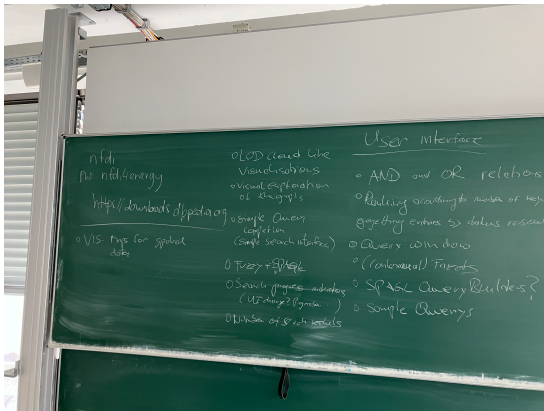


Figure 6: Collected results for the user interface (1) Figure 7: Collected results for the user interface (2)

Ideas for the user interface:

- AND and OR relations in keyword searches, as e.g. in Google
- Ranking according to number of keyword matches
- Listing entries by Databus resources (e.g. all by a certain user, for a certain group, or certain artifact).
- Query windows for complex queries
- Sample queries
- Contextual Facets
- SPARQL query builders
- LOD cloud like visualizations
- Visualization of the graphs
- Simple Query completion e.g. by Terminology service, in the simple search interface
- Fuzzy search
- Direct use of SPARQL
- Indication of search progress (e.g. through the change of the UI)
- Show number of search results
- Maps for spatial data
- Protection of some metadata, may be done by several MOSS instances inside and outside protected area (background: Queries to the graph database can not easily be limited)

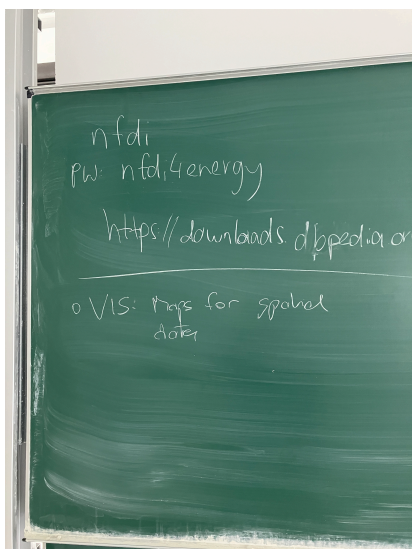


Figure 8: Collected results for the user interface (3)

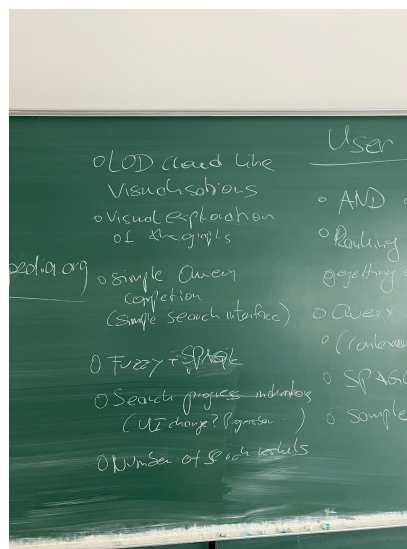


Figure 9: Collected results for the user interface (4)

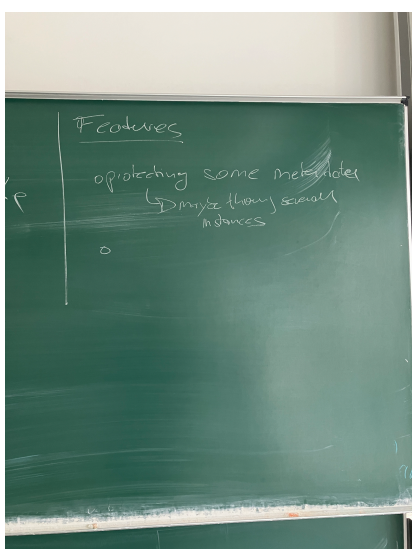


Figure 10: Collected results for the user interface (5)

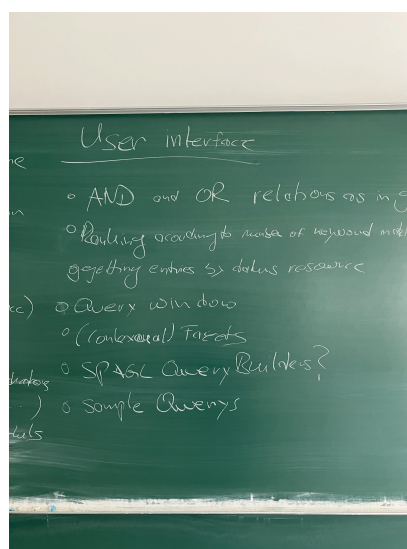


Figure 11: Collected results for the user interface (6)

Workshop at the Forschungsnetzwerk Systemanalyse May, 6th, 2025

The second workshop took place as part of the "Netzwerktreffen Forschungsnetzwerk Systemanalyse"¹⁰ in Berlin on May 6th, 2025. The audience there is more focused on the application of the tools for publishing and discovery of data. The participants discussed the same questions, as before. We only document new additions to the lists from above

- What are typical keywords you would use for searching? How would you like to filter larger numbers of search results?
- How do you like a user interface for searching? What feature are important.

The Discussion has been more integrated, but we try to separate the feedback into the categories of the questions. The results have been collected on a pin board:

Searching and Filtering

- Limit search results only to latest version. If there are several versions, hide older versions or indicate that there are more versions of this dataset.
- Search for specific projects (e.g. through the grant ID)
- Search for data generated by a specific software
- Search data from a specific person or institution.
- Selecting e.g. specific regions, e.g. federal states
- Intelligent geographical search: e.g. get all datasets which contain a region or all data sets where a region is contained. (e.g. if I look for data for Baden-Württemberg, should I also get German, European and Global data, because they contain Baden-Württemberg. Or the other way around, should I also find data for Stuttgart or Freiburg, because they are in Baden-Württemberg).

How do you like a user interface for searching?

Ideas for the user interface:

- Analysis of similarities of the search results, to generate ideas and hints for filtering.
- Modular search interfaces, change of search interface depending on domain or datatype
- Sample workflows and scripts how to download data and check if there are data updates on the server

¹⁰<https://www.forschungsnetzwerke-energie.de/events/de/jahrestreffen-2025-forschungsnetzwerk-energiesystemanalyse>

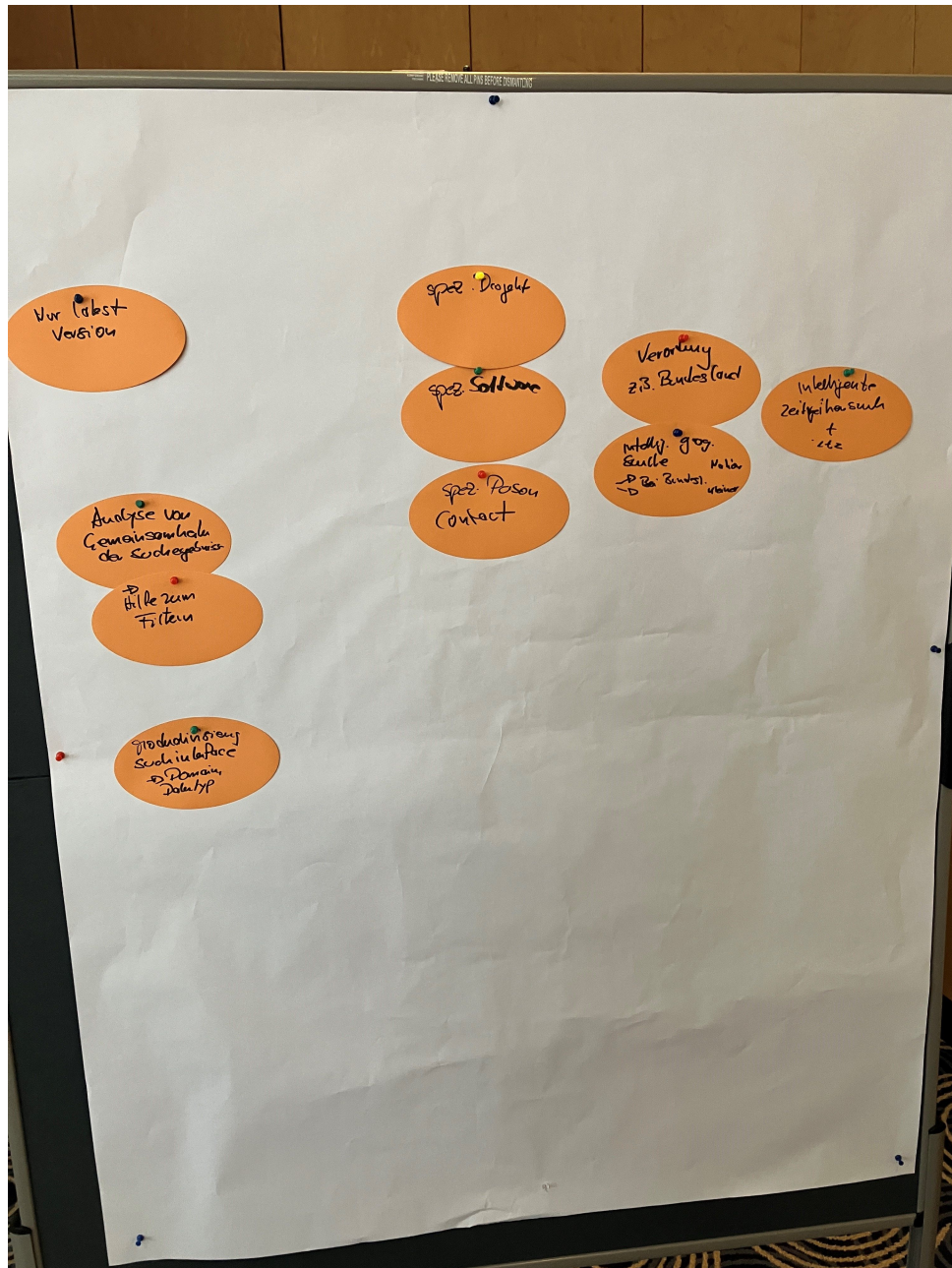


Figure 12: Feedback from the Workshop at "Forschungsnetzwerk Systemanalyse"

Workshop at Helmholtz Metadata Collaboration Conference, May 12th, 2025

The third workshop took place as part of the Helmholtz Metadata Collaboration Conference 2025¹¹ on May 12th, 2025 in Cologne. The audience there is more broad as it comes from a number of different disciplines which are part of Helmholtz, e.g. life sciences, matter, earth and environmental sciences, information sciences. The focus of the workshop has been on bridging the silos between different disciplines by using Databus and MOSS for metadata publication and discovery. We kept to the same questions:

- What are typical keywords you would use for searching? How would you like to filter larger numbers of search results?
- How do you like a user interface for searching? What feature are important.

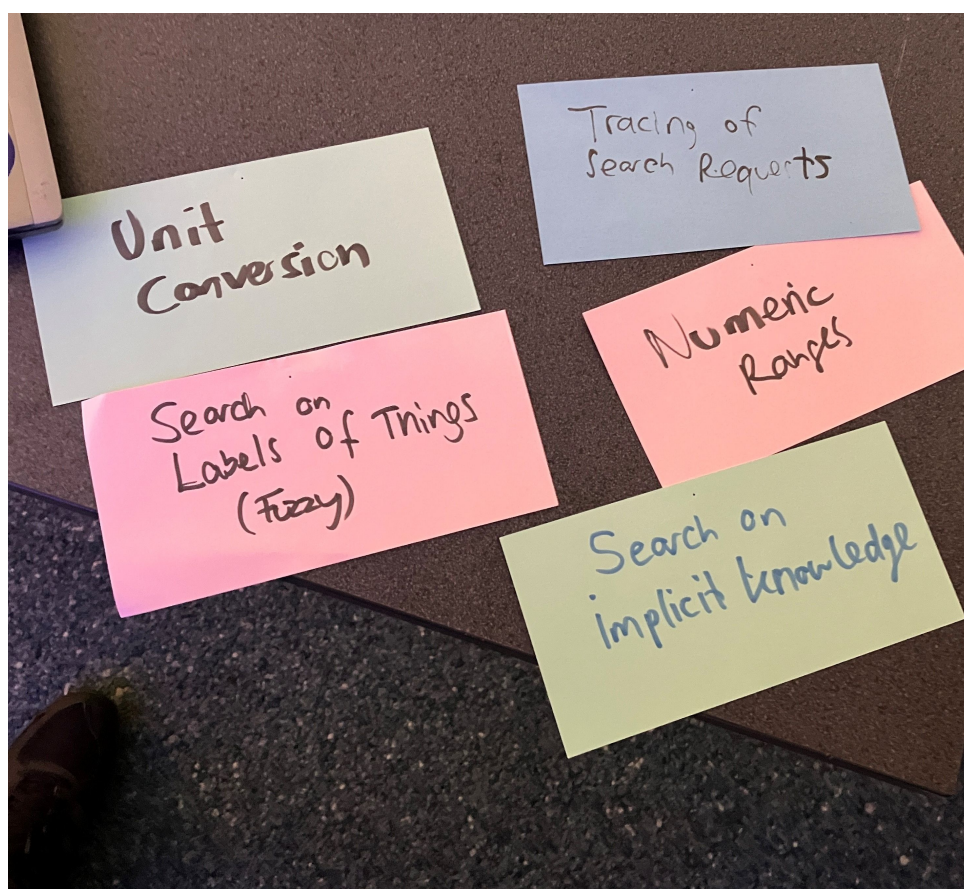


Figure 13: Feedback from the Workshop at Helmholtz Metadata Conference 2025

Searching and Filtering

- Unit Conversions, search for data of the same scale (e.g. anything measures in meters, or centimeter or micrometers.)
- Searching for research samples, e.g. from material or geological analysis, finding all the data which belongs to a certain sample
- searching for numerical ranges
- search on implicit knowledge, all terms which belong to a certain branch of an ontology

¹¹<https://helmholtz-metadaten.de/de/veranstaltungen/hmc-konferenz-2025>

How do you like a user interface for searching?

Ideas for the user interface:

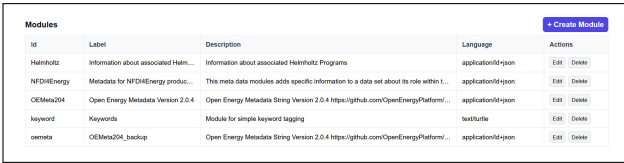
- Tracing of search request, e.g. go back to search from search results and reuse the previous query

Implementation

To address the user needs, a number of updates have been made to the Databus and MOSS software. The additions addressed two main areas: The creation of an admin interface to ease the curation of the metadata modules and a new search interface which uses the metadata information in the modules to improve the discovery of data. The implementation includes a front end and back end server which are linked in the README of the main MOSS repository on Github (<https://github.com/dbpedia/databus-moss>).

Admin Interface

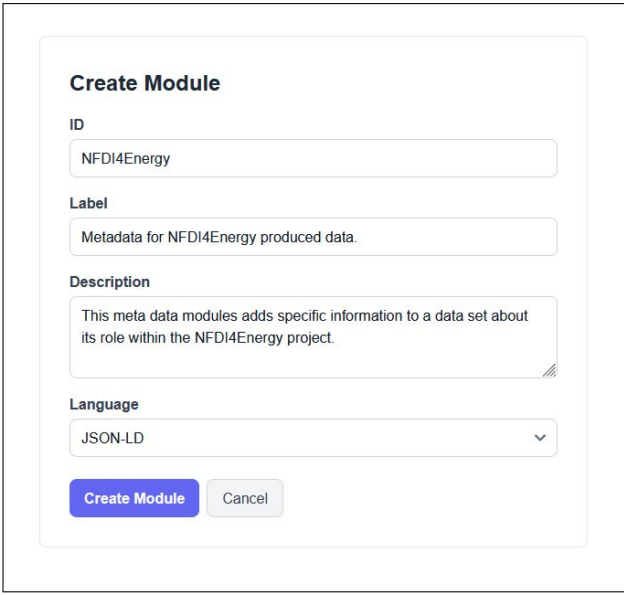
The admin interface is accessible to users with admin privileges and can be used to configure the MOSS instance. The most important configurations are the metadata extension schemata called modules, additional terminologies to load to the graph database and the search facets for the front page. All MOSS modules are listed at the top of the admin page. The list shows all existing modules. Here admins can edit modules or create new ones (see figure 14).



Id	Label	Description	Language	Actions
Helmholtz	Information about associated Helmholtz	Information about associated Helmholtz Programs	application/rdf+xml	Edit Delete
NFDI4Energy	Metadata for NFDI4Energy produced data	This meta data modules adds specific information to a data set about its role within L...	application/rdf+xml	Edit Delete
OEMeta204	Open Energy Metadata Version 2.0.4	Open Energy Metadata String Version 2.0.4 https://github.com/OpenEnergyPlatform/...	application/rdf+xml	Edit Delete
keyword	Keywords	Module for simple keyword tagging	text/turtle	Edit Delete
oemeta	OEMeta204_backup	Open Energy Metadata String Version 2.0.4 https://github.com/OpenEnergyPlatform/...	application/rdf+xml	Edit Delete

Figure 14: Listing of existing metadata modules and option to create new modules

A new module can be created by clicking on "+Create Module". A module requires an identifier string, a human readable label, a description and an RDF language. The language specifies the RDF language in which all related metadata is serialized. Figure 15 shows as an example the creation of an NFDI4Energy module.



Create Module

ID

Label

Description

Language

Create Module **Cancel**

Figure 15: Initialization of a sample NFDI4Energy module

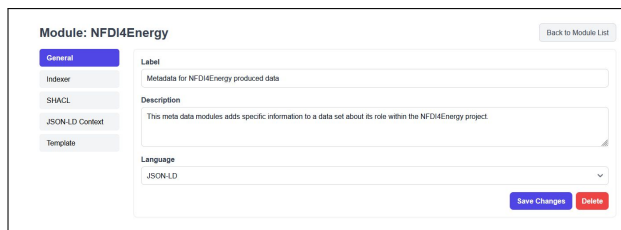


Figure 16: Edit page for a metadata module

Once the module has been created it can be edited by clicking on the module in the modules list to go to the module view. The module view features five tabs on the left hand side (for a JSON-LD module) to configure different aspects of the module (see figure16).

General Tab

The "General" tab lets admins alter the basic module information such as label and description.

Template Tab

The template tab allows the creation of a template that is used to give users a starting point when creating an entry for the module. This is recommended when manual entry creation is frequent and can help new users get accustomed with the metadata schema.

SHACL Tab

While the template helps to give a clearer understanding how a MOSS entry should look like, it is usually best to validate inputs to ensure that the information provided is complete. With the metadata being in an RDF format, our go-to validation language is SHACL. SHACL constraints (also called SHACL shapes) are a way of defining a more formal schema that data inputs can be validated against.

If you add SHACL shapes to your module, all entries created for this module will be validated against the shapes. Metadata can only be saved when conforming to the specified shapes.

Indexer Tab

This feature is paused and will likely be deprecated.

With an indexer you can specify how entries for this module will be written to a Lucene index. The Lucene indexer used to be the main driver of the search-strategy over MOSS instances and has been as such replaced by search facets paired with a SPARQL query generator.

You can however still use this feature to create a powerful search over the contents of the MOSS instance e.g. for use in other applications. The search indexer runs on the DBpedia Lookup project, which is documented here (<https://github.com/dbpedia/lookup>). Its core idea is to create key-value pairs for an inverse index by running specific SPARQL queries on RDF data.

The "Indexer" tab configures SPARQL queries which are used to fill the tables of the search indexer. In the "SHACL" tab, SHACL constraints for the validation of the entered data can be added.

JSON-LD Context Tab

The "JSON-LD context" tab can be used to host @context information to be used in the JSON-LD Metadata directly at the MOSS web page. MOSS will ensure that it is delivered as "application/LD+JSON" MIME type.

Facets					+ Create Facet
Id	Label	Predicate	Sort Order	Actions	
subject	Subject	<http://purl.org/dc/elements/1.1/subject> (<http://www.w3.org/ns/dcat#distribution> <http://purl.org/dc/terms/subje...	0	Edit	Delete
keyword	Keyword	<https://schema.org/keywords> <http://www.w3.org/ns/dcat#distribution> <http://www.w3.org/ns/dcat#keyword>	1	Edit	Delete
location	Location	<http://purl.org/dc/terms/spatial>	2	Edit	Delete
fundingAgency	Funding Agency	<https://dbpedia.org/ns/oem/fundingAgency>	3	Edit	Delete
grantno	Grant Number	<http://www.w3.org/ns/dcat#distribution> <http://schema.org/Grant>	4	Edit	Delete
license	License	<http://www.w3.org/ns/dcat#distribution> <http://purl.org/dc/terms/license> <http://purl.org/dc/terms/title>	5	Edit	Delete
format	Format	<http://www.w3.org/ns/dcat#distribution> <http://purl.org/dc/terms/format>	6	Edit	Delete
Publisher	Publisher	<http://www.w3.org/ns/dcat#distribution> <http://purl.org/dc/terms/publisher>	7	Edit	Delete
projecttitle	Project Title	<http://www.w3.org/ns/dcat#distribution> <https://openenergyplatform.org/OEMeta204#projectContext> <http://pu...	8	Edit	Delete

Figure 17: List of configured facets on a MOSS instance

Searching data

The search interface of the MOSS service consists of freely configurable facets. We first explored indexed search leveraging the modular structure of MOSS. Whenever an entry for any module is saved, a module specific indexer takes care of retrieving key-value-pairs from the metadata entry and updating a Lucene index that can then be used for API-based or front end search. This system works well for fuzzy search over arbitrary fields and even displaying explanations as to why an item shows up in the search results. However, when trying to realize incremental search, displaying counts of available items and doing more complex joins this system fell short. Several workarounds have been tried, such as indexing terminologies and using Lucene joins similar to the querying approach by Mitschick et al. [12]. However, configuring the module-specific indexers correctly and then mapping indexed fields to inputs and result listings of the application required a lot of manual work and even hard-coding some elements which went against the core idea of keeping MOSS as modular and easily configurable as possible.

With MOSS providing an RDF structure to retrieve and reason about metadata entries an RDF centric approach, namely a SPARQL-query generator, made both configuration and incremental search much easier. The only weakness of this system - the lack of proper fuzzy search over literals - was tackled by fetching data more broadly via SPARQL and then doing fuzzy matching of literals against search field inputs on the client side.

Instead of a complex indexing configuration, each facets merely specifies a predicate path connecting the described resource (Databus resource) and a target literal or terminology term. Using these predicate paths, MOSS will construct SPARQL-queries fetching distinct values of all entries along with counts and labels for the distinct values. Literal values will use themselves as label, terminology term labels are fetched from the respective terminology data. Search over labels is then done on the client side using the popular *fuzzy sort* JavaScript library.

This creates an easy-to-configure and flexible search setup with MOSS being able to target entries from multiple modules with a single search facet. This can even be achieved if modules use different predicates for similar things by using special SPARQL characters ("|" and "/") in the predicate path of the facet. A path like <: subject > |(<: distribution > / <: subject >) will allow search over values directly associated with a Databus resource via the <: subject > predicate as well as values associated with any object via the <: subject >" predicate, with the object in turn associated with the Databus resource via the <: distribution > predicate. Search over the facet will then allow MOSS to provide results from two different modules handling the declaration of subject metadata differently.

Already selected facets will be used to construct a final search query to retrieve results as well as update the queries of all facets to update the counts for incremental search. Figure 18 shows a sample configuration for a search interface. The first facet "Subject" is configured as described above, selecting metadata information from both Open Transport and Open Energy metadata. The information describing data-related subjects or topic is contained in both metadata formats. The subject terms used in the

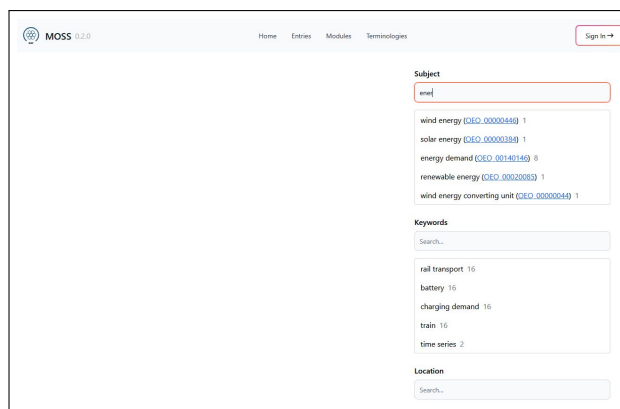


Figure 18: Sample configuration of the MOSS search interface

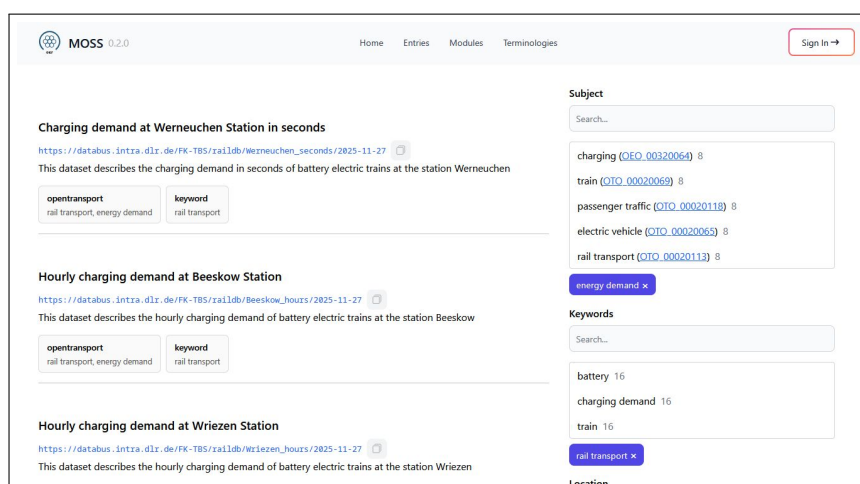


Figure 19: Sample Search results in MOSS. Multiple selections in the facets can be combined to narrow down search results

RDF are URIs from the Open Transport and Open Energy Ontology respectively. Both ontologies have been loaded into the MOSS instance as terminologies, providing human readable labels for energy and transport related concepts. The listing below the input field shows the 5 terminology terms with the most occurrences in metadata entries. The numbers at the end indicate how often this term has been used in the metadata. Providing any input in the facets input field will run a local fuzzy search on all terminology terms that have been declared a subject in any metadata entry.

The keywords facet uses the "keywords" field in the metadata modules. Further configured facets in this case are location and funding agencies. Any field in the metadata which has a valid linked data IRI can be used to configure a facet.

Figure 17 shows several facets that have been configured on a development MOSS instance. It lists configured facets with their id, human readable label, predicate chain and sorting order in the search interface.

Figure 19 shows a sample search result. Selections from multiple facets can be used to narrow down a search result. The titles and descriptions of the search result are directly pulled from the Databus metadata. By clicking on the Databus link, the Databus entry with the data download links can be accessed. The boxes below the result entry show information, why this entry has been found and by clicking on the boxes, the metadata entries on the MOSS can be viewed and edited.

Conclusions and Outlook

Databus and MOSS together form a flexible, lightweight metadata catalog designed to bridge data silos in research data management. Deployment is straightforward, relying on predefined Docker containers. Most configuration can be handled through an administrator interface, with a few basic settings initially requiring adjustments in the containerized environment - a process that will be fully integrated into the administrator interface in future releases. The platform's support for metadata module management enables the curation of community standards, fostering interoperability across diverse data sources. Currently, Databus and MOSS are being deployed at the Open Energy Family and within DLR to gather practical insights on implementing FAIR research data management across different domains and organizational structures.

Looking ahead, several avenues are being explored to enhance the platform's utility and scalability. This includes more fine-grained modularization, breaking down larger metadata formats such as Open Energy Metadata into smaller, reusable modules. Cross-domain reuse is another key prospect, exemplified by potential overlaps between Open Transport Metadata and Open Energy Metadata schemas. While the current implementation is expected to handle thousands of metadata entries with extensive terminologies, performance at the scale of millions of entries remains to be evaluated. Additionally, ongoing work aims to support plug-and-play installation of metadata modules, further simplifying deployment and adoption across research infrastructures.

References

- [1] DBpedia Association, *Databus gitbook*. [Online]. Available: <https://dbpedia.gitbook.io/databus> (visited on 05/27/2026).
- [2] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, “The FAIR Guiding Principles for scientific data management and stewardship,” en, *Scientific Data*, vol. 3, no. 1, p. 160 018, Mar. 2016, Publisher: Nature Publishing Group, ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18. [Online]. Available: <https://www.nature.com/articles/sdata201618> (visited on 06/01/2024).
- [3] DBpedia Association. “Databus.” (2025), [Online]. Available: <https://databus.dbpedia.org> (visited on 10/24/2025).
- [4] DBpedia Association. “Databus moss.” (2025), [Online]. Available: <https://github.com/dbpedia/databus-moss> (visited on 10/24/2025).
- [5] Open Energy Familiy. “Open energy platform.” (2025), [Online]. Available: <https://openenergyplatform.org> (visited on 10/24/2025).
- [6] M. Freudenberg, M. Brümmer, J. Rücknagel, *et al.*, “The metadata ecosystem of dataid,” in *Metadata and Semantics Research*, E. Garoufallou, I. Subirats Coll, A. Stellato, and J. Greenberg, Eds., Cham: Springer International Publishing, 2016, pp. 317–332, ISBN: 978-3-319-49157-8.
- [7] C. Hoyer-Klick, U. Frey, V. Sehn, *et al.*, “Demonstration and best practices,” LOD-GEOSS Project, funded by German Ministry for Economic Affairs and Climate Action, Grand 03EI1005A-G, Tech. Rep., 2023. [Online]. Available: <https://doi.org/10.5281/zenodo.8119096>.
- [8] M. Booshehri, L. Emele, S. Flügel, *et al.*, “Introducing the open energy ontology: Enhancing data interpretation and interfacing in energy systems analysis,” *Energy and AI*, vol. 5, p. 100 074, 2021, ISSN: 2666-5468. DOI: <https://doi.org/10.1016/j.egyai.2021.100074>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666546821000288>.
- [9] D. H. Wolpert and W. G. Macready, “No Free Lunch Theorems for Search,” Santa Fe Institute, Working Papers 95-02-010, Feb. 1995. [Online]. Available: <https://ideas.repec.org/p/wop/safiwp/95-02-010.html>.
- [10] J. Frey, C. Hoyer-Klick, C. Muschner, L. Hülk, D. Streitmatter, and S. Hellmann, “Distributed data infrastructure,” LOD-GEOSS Project, funded by German Ministry for Economic Affairs and Climate Action, Grand 03EI1005A-G, Tech. Rep., 2023. [Online]. Available: <https://doi.org/10.5281/zenodo.8119055>.
- [11] M. Hofer, S. Hellmann, M. Dojchinovski, and J. Frey, “The new dbpedia release cycle: Increasing agility and efficiency in knowledge extraction workflows,” in *Semantic Systems. In the Era of Knowledge Graphs*, E. Blomqvist, P. Groth, V. de Boer, *et al.*, Eds., Cham: Springer International Publishing, 2020, pp. 1–18, ISBN: 978-3-030-59833-4.
- [12] A. Mitschick, F. Nieschalk, M. Voigt, and R. Dachsel, “Iciclequery: A web search interface for fluid semantic query construction,” in *VOILA@ ISWC*, 2017, pp. 99–110.