3rd NFDI4Energy Conference: Defining Energy Research Data Management Together
Proceedings to the 3rd NFDI4Energy Conference - Full Papers
https://doi.org/10.52825/ocp.v9i.3327

Published: 23 Mar. 2026

Mapping the Metadata Landscape of Energy Data Repositories

A Multi-Repository Assessment of Conformance to OEMetadata

Philipp D. Rohde[1,2,*], Enrique Iglesias[2,3], and Maria-Esther Vidal[1,2,3]

[1]TIB Leibniz Information Centre for Science and Technology, Hannover, Germany
[2]Leibniz University of Hannover, Hannover, Germany
[3]L3S Research Center, Hannover, Germany
*Correspondence: Philipp D. Rohde, philipp.rohde@tib.eu

Abstract. This paper empirically assesses how well energy-relevant repositories supply the dataset- and column-level metadata required by OEMetadata Schema. Records from ten catalogs are harvested, crosswalked to OEMetadata v2, normalized, and analyzed to compute dataset- and resource-level coverage. Coverage is weak and uneven: provenance and contributor data appear mainly in OEP/OPSD; licenses are absent in half the portals; temporal/spatial metadata are sparse; access URLs are often missing; and column-level data dictionaries are largely absent. We release crosswalks and a reproducible harvesting workflow, establishing a baseline for measuring and comparing interoperability across repositories.

Keywords: Metadata, Metadata Quality, Interoperability

1. Introduction

The accelerating digitalization of the energy sector has made high-quality, interoperable metadata a prerequisite for reproducible research, transparent policy analysis, and robust operational decision-making. Energy datasets span diverse modalities—tabular time series, geospatial layers, model inputs and outputs, and aggregated statistics—and are disseminated through a heterogeneous landscape of repositories and portals. While generic repository schemas facilitate discovery at a broad level, domain workflows increasingly depend on richer, machine-actionable descriptions that capture temporal and spatial coverage and, crucially, column-level semantics for tabular resources. Without such detail, integration across sources, automation of pipelines, and meaningful reuse remain costly and error-prone.

To address these needs, the Open Energy Platform community has specified the OEMetadata Schema v2 [1], a domain-specific profile designed to standardize both dataset-level and field-level metadata for energy research and modeling. OEMetadata v2 introduces explicit keys for identification and citation, provenance, licensing, temporal and spatial coverage, and a structured *data dictionary* that describes each column of

a table or CSV, including names, descriptions, data types, units, code lists, and other constraints. By aligning repository records to OEMetadata v2, data providers can enable downstream consumers to discover, interpret, and validate energy datasets with far less manual effort.

In practice, however, energy-relevant repositories expose metadata using a variety of native models—such as DCAT[1]/DCAT-AP[2], Dublin Core[3], DataCite[4], CKAN defaults, ISO 19115[5]/19139[6], schema.org/Dataset[7], or SDMX[8]—and often lack systematic field-level descriptions for tabular resources. This heterogeneity complicates cross- repository discovery and hampers the machine-actionability needed for automated validation and integration. It also raises a concrete question for the community: to what extent do current repositories already provide the information required by OEMetadata v2, and where are the gaps that most impede interoperability?

This paper provides an empirical answer by assessing the coverage of OEMetadata v2 for energy-related datasets drawn from a representative set of ten repositories and portals: Energy Data eXchange, EU Open Data Portal, Eurostat, FfE Open Data Portal, Inspire-HEP, the Open Energy Data Initiative, the Open Energy Platform, OpenAIRE, Open Power System Data, and the Open Access Power-Grid Frequency Database. We focus our analysis on the coverage, i.e., presence of OEMetadata v2 keys. We do not evaluate the correctness of the underlying data, assess broader quality dimensions such as accuracy or timeliness, or attempt full semantic validation beyond what the schema specifies.

Our contributions are:

1. Crosswalks from each repository's native metadata to OEMetadata;
2. An automated harvesting workflow computing reproducible coverage metrics; and
3. Descriptive results and recommendations for improving OEMetadata alignment.

By establishing a clear baseline of coverage, the paper highlights where repositories already provide OEMetadata v2 elements and where gaps remain. This baseline supports targeted improvements and enables future tracking of progress toward more interoperable, machine-actionable energy data.

2. Multi-Repository Assessment

This section reports a multi-repository assessment of metadata coverage against the OEMetadata Schema v2. We first outline the scope and intent of OEMetadata v2, then describe the set of repositories examined and how their metadata were harvested, and finally present coverage results for the schema's dataset-level and resource-level elements.

[1] https://www.w3.org/TR/2024/REC-vocab-dcat-3-20240822/
[2] https://semiceu.github.io/DCAT-AP/releases/3.0.0/
[3] http://dublincore.org/specifications/dublin-core/dcmi-terms/2020-01-20/
[4] https://schema.datacite.org/meta/kernel-4.6/
[5] https://www.iso.org/standard/53798.html
[6] https://www.iso.org/standard/67253.html
[7] https://schema.org/Dataset
[8] https://sdmx.org/standards-2/

2.1 OEMetadata Schema

OEMetadata Schema v2 is a community-defined, domain-specific metadata profile developed within the Open Energy Platform (OEP) and the broader Open Energy Family to make energy datasets easier to find, interpret, and reuse. Unlike generic repository schemas that stop at high-level descriptions, OEMetadata v2 is designed to capture the details energy workflows actually depend on, including machine-readable temporal and spatial coverage and column-level "data dictionaries" for tabular resources. The specification is published openly and documents each key, its intent, expected value type, and examples. At a high level, OEMetadata v2 organizes information into two complementary layers:

- *Dataset-level metadata* that supports identification, citation, licensing, provenance, temporal and spatial coverage, and basic access information; and
- *Resource- and table-level metadata* that describes individual distributions (for example, CSV files) and provides a structured data dictionary for each table, including column names, human-readable descriptions, datatypes, units, and (where applicable) allowed values or code lists.

Core elements covered by the schema include:

- *Identification and citation:* titles, descriptions, version information, persistent identifiers for datasets where available, and links to related publications or projects to support attribution and provenance.
- *Actors and provenance:* creators, contributors, and contacts (ideally with person or organization identifiers) and references to sources or methods to establish lineage.
- *Licensing and access:* explicit license declarations and basic access information to enable lawful reuse.
- *Temporal and spatial coverage:* standardized representations of time intervals and spatial footprints to make datasets discoverable and comparable across studies.
- *Resources and distributions:* per-file metadata for each provided resource (format, size, access URL) to support automation.
- *Data dictionaries for tabular data:* column-level metadata specifying names, descriptions, datatypes, and units, enabling unambiguous interpretation and safer downstream processing.

The schema is intentionally interoperable with broader standards: its concepts can be crosswalked to widely used models such as DCAT/DCAT-AP, DataCite, ISO 19115, and schema.org/Dataset. In practice, OEMetadata v2 fills a gap these general schemas leave by requiring energy-relevant detail at the column level and by encouraging consistent, machine-readable temporal and spatial descriptions. It also promotes use of persistent identifiers (for example, DOIs for datasets, ORCID/ROR for actors) and common conventions (for example, ISO 8601 for times), which improves reproducibility, traceability, and automated integration. In this study, we use OEMetadata v2 as the target profile to measure coverage: whether dataset records and tabular resources provide the keys it specifies. This establishes a domain-relevant baseline for how well current repositories support the metadata needed for interoperable, machine-actionable energy research.

2.2 Metadata Retrieval & Analysis

Additionally to the metadata fields defined in the OEMetadata Schema, we also evaluate the coverage of author information. The quality of the metadata describing energy-related

datasets from the following repositories is assessed in terms of their conformance to OEMetadata Schema v2:

1. **Energy Data eXchange (EDX)** – U.S. DOE/NETL platform for sharing energy and environmental datasets and tools. https://edx.netl.doe.gov/
2. **EU Open Data Portal (EU Open)** – The EU's central open data aggregator (data.europa.eu) for datasets from EU institutions. https://data.europa.eu/data/datasets
3. **Eurostat** – Statistical Office of the European Union providing official statistics and SDMX APIs (includes energy). https://ec.europa.eu/eurostat
4. **FfE Open Data Portal (FfE)** – Forschungsstelle für Energiewirtschaft's portal for energy research datasets. https://opendata.ffe.de/
5. **Inspire-HEP (HEP)** – High-energy physics discovery platform (INSPIRE collaboration; not energy-sector specific). https://inspirehep.net
6. **Open Energy Data Initiative (OEDI)** – NREL/US DOE catalog of open energy datasets across technologies and domains. https://data.openei.org/
7. **Open Energy Platform (OEP)** – Community platform for energy system modelling data and metadata (Open Energy Family). https://openenergyplatform.org/database/
8. **OpenAIRE** – European research aggregator for publications, datasets, and other outputs. https://explore.openaire.eu/
9. **Open Power System Data (OPSD)** – Curated, versioned European power-system datasets (time series, market data, weather links). https://open-power-system-data.org/
10. **Open Access Power-Grid Frequency Database (PGF)** – OSF-hosted repository of open time-series measurements of power grid frequency. https://osf.io/m43tg/

For each repository, the available programmatic interfaces are first identified and exercised to obtain a complete, machine-readable snapshot of dataset records and their distributions. Records are then examined to ensure that all distributions associated with a dataset are captured, that versioned entries are not double-counted, and that landing-page metadata is distinguished from per-file metadata. A schema crosswalk from each native metadata model to OEMetadata v2 is developed by manual analysis of representative records. The crosswalk specifies target keys and value transformations, for example normalization of author or contributor names. Particular attention is paid to nested structures in which parent keys appear populated while all required subfields are empty; in such cases, empty children are deleted and parent keys without surviving children are dropped to avoid inflating coverage. After cleaning, the crosswalks are executed to produce OEMetadata-conformant records for all ten repositories. The transformed metadata are analyzed to compute coverage of OEMetadata attributes. Coverage at the dataset level is calculated as the percentage of dataset records with a non-empty, post-normalization value that satisfies the target key's expected type; coverage at the resource level is calculated analogously over individual distributions. For composite attributes with subfields, such as temporal coverage, spatial coverage, sources, contributors, and schema elements, coverage is reported at the parent level only when at least one semantically meaningful subfield survives normalization. This procedure yields conservative, reproducible estimates of effective OEMetadata presence that reflect not only whether keys exist in principle, but whether they contain usable, machine-actionable content.

Table 1. Coverage of the OEMetadata Schema v2 in Percent. *While dataset title and description are well covered since they are standard metadata fields, OEMetadata specific fields like schema information, delimiter, and decimal separator are badly covered.*
– denotes the absence of the respective metadata information

Attribute	Repository									
	EDX	EU Open	Eurostat	FfE	HEP	OEDI	OEP	OpenAIRE	OPSD	PGF
author	–	–	–	–	99.65	–	–	37.57	–	100
dataset title	100	99.96	100	100	100	100	0.49	100	100	100
dataset description	100	99.84	100	98.89	96.04	–	0.0	39.69	–	100
resource title	99.96	63.69	100	–	100	–	92.36	–	23.98	100
access URL	100	–	–	100	100	–	100[a]	–	100	100
resource description	7.8	–	–	–	–	99.98	88.46	–	11.7	–
language	–	99.9	100	100	–	–	99.03	6.73	–	–
keywords	–	72.86	75.14	–	98.86	–	62.24	21.25	–	–
publication date	–	–	–	–	100	–	49.11	98.95	–	–
spatial	6.07	55.44	70.0	–	–	78.14	70.66	–	23.26	100
temporal	–	36.83	100	–	–	–	61.10	–	30.23	–
source	–	–	–	–	–	–	87.84	–	76.74	–
license	–	51.56	89.46	–	–	100	99.19	–	41.86	–
contributor	–	–	–	–	–	–	96.92	–	81.40	–
resource format	100	60.25	100	100	–	92.43	66.67	3.63	100	–
resource encoding	–	–	–	100	–	–	37.89	–	73.68	–
schema fields	–	–	–	–	–	–	100	–	67.84	–
primary key	–	–	–	–	–	–	100	–	20.47	–
foreign keys	–	–	–	–	–	–	7.97	–	–	–
delimiter	–	–	–	–	–	–	1.63	–	43.86	–
decimal separator	–	–	–	–	–	–	29.76	–	9.36	–

[a]but only 63.41% are valid URLs

2.3 Observed Results

2.3.1 OEMetadata Coverage

Table 1 reports the coverage of OEMetadata fields in percent per repository. Since the metadata schema includes many subfields, e.g., for spatial and temporal information, the coverage of the following attributes is grouped and reported only for the parent key, indicating that at least one of the subfields is filled: spatial, temportal, source, license, contributor, and schema fields.

Surprisingly only three of the repositories, namely Inspire-HEP, OpenAIRE, and the Open Access Power-Grid Frequency Database, include author information in the metadata. OEDI has author information but it is not present in the metadata provided. But even in OpenAIRE only slightly more than a third of the datasets have author information. Without author information, datasets lack clear provenance and accountability, making it harder to assess methods, trust data quality, and reproduce results; users also lose a contact point to resolve ambiguities, errors, or versioning issues. Its absence undermines citation and credit workflows—disambiguation via identifiers such as ORCID enables proper attribution and linking to publications and grants—reducing discoverability, reuse, and the ability to track impact.

Basic information like the dataset title and description as well as resource title are given for almost all of the datasets in the repositories. The OEP currently has many datasets without a title due to the change to the new metadata version that included the dataset level. FfE, OEDI, and OpenAIRE, however, do not provide resource titles. In OPSD about a quarter or the resources have a title. Access URLs for the resources are not present in the metadata retrieved from EU Open, Eurostat, OEDI,

and OpenAIRE. Without resource access URLs, users cannot retrieve the exact files underpinning reported results, making reproduction, verification, and citation of specific versions difficult or impossible. Their absence also breaks machine-actionable workflows—harvesters, validators, and analysis pipelines—undermining FAIR [2] accessibility and forcing manual, error-prone navigation via landing pages.

Having spatial and temporal metadata in the datasets is crucial as it allows for the analysis of energy consumption patterns at specific locations and times, enabling a more accurate and detailed understanding of urban energy usage. This information helps identify areas of high energy demand, peak usage periods, and potential opportunities for energy efficiency improvements. With this metadata, correlations between energy consumption and other urban factors, such as population density and climate, can be explored. However, some repositories do not capture this information, and even those that do typically provide it for only a third to two thirds of their datasets.

Provenance information—particularly explicit identification of sources and contributors—underpins trust, interpretability, and reproducibility in secondary analysis and data integration. Source disclosure provides a methodological pedigree: how data were collected, with which instruments or models, under which institutional standards, and with what preprocessing, calibration, quality control, and gap-filling procedures. In energy research, semantic and methodological heterogeneity is pervasive—installed versus net capacity, gross versus net generation, AC versus DC ratings for photovoltaics, aggregation boundaries for demand, and time zone handling in time series are recurrent points of divergence. Without clear provenance, datasets that are not commensurate may be combined, signals derived from common upstream measurements may be double-counted, and uncertainty may be mischaracterized. Contributor information adds accountability and practical utility beyond credit. Identified maintainers and curators provide a route for clarification, correction, and updates; declared roles and affiliations make implicit assumptions and potential biases more transparent; and contactable stewardship supports ongoing curation. Provenance also enables precise citation and versioning, both of which are essential for scholarly integrity and for freezing analyses to specific dataset states when revisions occur. However, only OEP and OPSD provide this information with OPSD recording the sources for 76.74% versus 87.84% in OEP, and 81.40% and 96.92% for contributors, respectively.

Licensing information is equally critical because it defines the legal permissions and obligations that govern reuse, modification, combination, and redistribution. License terms such as attribution, share-alike, non-commercial, or no-derivatives directly shape research workflows, constrain integration strategies, and determine dissemination pathways for derived products. In composite datasets assembled from multiple sources, license compatibility is a binding constraint: incompatible terms can render an otherwise valuable aggregation non-distributable or inconsistent with open-science mandates. Clear, explicit, and preferably machine-readable license metadata reduces legal uncertainty, minimizes transaction costs associated with seeking permissions ex post, and facilitates collaboration across institutions. Transparent licensing also supports compliance with funder, institutional, and journal policies, clarifies obligations such as attribution placement and notice propagation, and secures the long-term viability of both the original data and derivatives by making downstream rights unambiguous for archiving, mirroring, and replication studies. Nevertheless, only five out of ten repositories even capture licensing information. Apart from EU Open Data and OPSD, the percentage of datasets including licensing information is about 90% or higher, while EU Open Data provides this information for 51.56% of the datasets and OPSD for 41.86%.

Table 2. Repository Statistics. *Due to limitations in the provided APIs, the metadata of a limited number of datasets can be retrieved. The total number of datasets refers to energy-related datasets, repositories like EU Open and OpenAIRE host many more datasets of other domains.*

Repository	Datasets (downloaded)	Datasets (total)	Resources
EDX	17,221	17,221	32,416
EU Open	10,000	24,743	55,281
Eurostat	370	370+[a]	1,577
FfE	90	90+[b]	90
HEP	10,000	10,691	133,035
OEDI	2,703	2,724	14,221
OEP	617	617[c]	615
OpenAIRE	10,000	37,746	10,000
OPSD	43	43	171
PGF	1	1	1

[a] exact number unknown
[b] exact number unknown
[c] datasets that are not in draft mode

Explicit schema metadata—column names, data types, descriptions, units, and value domains—ensures unambiguous semantic interpretation and type-safe processing, preventing dimensional and parsing errors and enabling validation and automated integration across heterogeneous sources. It is essential for reproducibility and comparability, as consistent schema definitions allow correct alignment of variables, facilitate unit conversions and aggregation, and support quality control and provenance-aware transformations. However, this information is only captured in OEP and OPSD with varying coverage, e.g., the fields are described for all the resources in the OEP but only for about two thirds in OPSD. On the other hand, field delimiters are more ofter recorded in OPSD, 43.86% versus 1.63% in the OEP.

2.3.2 API Limitations

Cross-repository harvesting is impeded by API limitations and inconsistencies that reduce coverage and compromise provenance; as reported in Table 2 which shows the number of datasets and resources for which the metadata is downloaded as well as the total number of energy-related repositories. In OEDI, author information exists but is not exposed in the API response, obscuring attribution. Several catalogs impose a hard cap of 10,000 records despite larger holdings—OpenAIRE returns 10,000 of 37,746 datasets, Inspire 10,000 of 10,691, and the EU Open Data Portal 10,000 of 24,743—resulting in truncated, potentially biased samples. Some APIs are not up to date (e.g., OEDI exposes 2,703 of 2,724 datasets), further limiting completeness. FfE offers no API; metadata must be retrieved per dataset ID, and not all datasets have an ID, hindering systematic harvesting.

3. Conclusions

This multi-repository assessment shows that alignment with OEMetadata Schema v2 is weak not only in breadth but also in depth. Coverage of critical elements is systematically incomplete—provenance (sources, contributors) appears almost exclusively in OEP and OPSD; licensing is missing in half of the catalogs and sporadic where present; temporal and spatial descriptors are provided for only a minority of records; and

column-level data dictionaries are largely absent outside OEP and OPSD. Basic resource-level elements are frequently missing as well: many catalogs do not expose per-file access URLs, and several provide no resource titles. Authorship information is largely absent, with only a few repositories supplying it at all. Taken together, these gaps undermine discovery, reproducibility, and machine-actionable reuse. Equally important, the way metadata are entered often fails to meet the intent of OEMetadata v2. Manual crosswalking and cleanup revealed records where parent keys are populated while required subfields are empty, inflating apparent coverage without conveying usable detail. Across repositories, metadata tend to be exposed at the landing-page level rather than per resource, impeding automation; roles and provenance are conflated or omitted; and resource-level schema elements (for example, field delimiters and data dictionaries) are inconsistently supplied. These issues indicate not only missing fields but also structural and granularity mismatches between native models and OEMetadata v2, so that even "present" keys frequently lack the machine-readable content needed for validation and integration. Addressing these deficiencies requires both adoption of OEMetadata v2 and improvements in capture and exposure practices: repositories should provide per-resource access URLs, populate provenance, licensing, and actor fields with persistent identifiers, and supply complete, machine-readable temporal, spatial, and column-level dictionaries. Clearer guidance and validators, robust crosswalks from DCAT/DataCite/ISO profiles, and repository workflows that prevent empty parent keys and enforce subfield completeness would markedly raise effective interoperability. The baseline established here indicates that progress will depend as much on how fields are populated as on whether they exist, and it motivates targeted remediation strategies that couple schema alignment with quality controls on content.

Data availability statement

The data and scripts used for this report are published in the NFDI4Energy instance of the Leibniz Data Manager under DOI 10.71694/4tut8jmy [3].

Author contributions

Philipp D. Rohde: Software, Validation, Investigation, Resources, Writing - Original Draft, Writing - Review & Editing, Supervision **Enrique Iglesias:** Software, Validation, Investigation, Resources, Writing - Review & Editing **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

This work is funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in the Nationale Forschungsdateninfrastruktur (National Research Data Infrastructure) project *NFDI4Energy* (grant no. 501865131), and Leibniz Association, program "Leibniz Best Minds: Programme for Women Professors", project TrustKG-Transforming Data in Trustable Insights; Grant P99/2020.

References

[1] L. Hülk, J. Huber, C. Hofmann, and C. Muschner, *Open Energy Family - Open Energy Metadata (OEMetadata)*, version 2.0.4, Jan. 2025. [Online]. Available: https://github.com/OpenEnergyPlatform/oemetadata.

[2] M. D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship", *Sci Data*, vol. 3, no. 160018, 2016. DOI: 10.1038/sdata.2016.18.

[3] P. D. Rohde and E. Iglesias, *Multi-Repository Metadata Assessment*, Leibniz Data Manager Instance for NFDI4Energy, Hannover, DEU, Dec. 2025. DOI: 10.71694/4tut8jmy.