

Recommendation S1.0

Recommendation to enrich data with rich metadata

Description

Status: Under development, Date: 2025/07/08 10:18, Version: 001

Motivation for this Recommendation:

As data can only be embedded in semantic frameworks when it is described with rich metadata, the first step toward a standardized approach for implementing semantic resources is for data producers to enrich data with metadata, and as part of this for data producers or given communities to agree on which metadata should be provided by data producers and required by data infrastructures. Only the standardized use of metadata enable annotation with identifiable terms from recognized controlled vocabularies, which allows machines to interpret and connect data across disciplinary and institutional boundaries. Since the needs for these metadata can vary greatly between research communities, data infrastructures, and use cases, we recommend using existing metadata schemas commonly used in the respective communities, data infrastructures, and use cases as a guide. Below, we will list just a few of the many possible ones. In addition, there are generally applicable schemas that can be recommended.

Recommendation summary

All data producers in Helmholtz Earth & Environment should enrich their datasets with rich, standardized metadata at the time of dataset creation, submission to repositories or publication. This should be done following established schemas such as DataCite for bibliographic metadata, or other (among others, discipline-specific) schemas or workflows. By ensuring compliance with these standards and repository-specific workflows, datasets become preserved, interoperable, and reusable across Helmholtz centers and the national and international scientific community.

Binding Convention:

	mandatory	conditional	optional
Helmholtz FAIR Principle	x		

Precondition for Implementation:

Related Recommendations

Parent: S0

Dependent: S3.0

Other: none

Contributors

Content

1. Explanation of the Background and Benefits of the Recommendation

About

A metadata schema defines the structure, content, and semantics of metadata elements used to describe a dataset. It specifies what metadata should be captured, how it should be named, and in which format it should be stored. Schemas often include controlled vocabularies and formal structures, allowing metadata to be understood both by humans and machines. Widely used schemas include ISO 19115 for geospatial data, DataCite Metadata Schema for citation metadata, and Dublin Core for general resource description.

History

Metadata standards emerged in the late 20th century to support the growing need for systematic data management and interoperability. For Earth and Environmental sciences, geospatial standards such as ISO 19115 and OGC SensorML were developed under the International Organization for Standardization and the Open Geospatial Consortium. Bibliographic metadata practices evolved from library science toward DataCite and Dublin Core. Increasingly, domain-specific descriptive metadata categories like observed variables/parameters or methodological metadata have been integrated to describe the actual measurements and observational procedures. These extensions ensure that both the content and the provenance of the data are captured for reuse, interpretation, and interoperability.

Current Use of Metadata Standards in Earth & Environmental Sciences

Across Helmholtz Earth and Environmental sciences, metadata standards are applied in diverse repositories and infrastructures, to name just a few of them:

- PANGAEA (AWI/MARUM) applies metadata workflows combining descriptive (dataset titles, abstracts, parameters, methods), structural (campaign and event hierarchies), and administrative (file formats, DOIs, licenses) elements. PANGAEA uses ISO 19115, DIF, Dublin Core, and its own metadata schema that emphasizes parameters, methods, and contextual information, ensuring interoperability across geosciences and marine research (<https://wiki.pangaea.de/wiki/Metadata>).
- GFZ Data Services (GFZ Potsdam) supports metadata based on ISO 19115, NASA GCMD DIF,

and DataCite, with its own metadata entry system providing templates and controlled vocabularies for FAIR compliance.

- Helmholtz Coastal Data Center (HCDC) relies on ISO 19115 and NetCDF CF Conventions, and in some cases OGC SensorML, to capture observational and sensor-based data.
- DataCite Metadata Schema provides the global backbone for dataset citation and retrieval (DataCite Schema)

Motivation

By aligning metadata practices with recognized standards, Helmholtz Earth & Environment can ensure that data are not only preserved but also discoverable, interpretable, and reusable across institutional and disciplinary boundaries.

2. Possible alternative solutions

3. Consideration of the advantages and disadvantages of implementing the recommendation

(quality of content, limitations, interoperability, sustainability: expected future dissemination / technical availability / funding)

4. The Recommendation

Bibliographic Metadata

It is recommended that for the accurate and consistent identification of a resource for citation and retrieval purposes, each published dataset should be provided with the core metadata elements defined in the most up-to-date DataCite Metadata Schema (see <https://schema.datacite.org/>).

Types of Metadata

We recommend distinguishing between the main categories of metadata as defined in ISO standards and adopted by infrastructures such as DataCite:

Descriptive Metadata: describe the intellectual content of the data.

Example: titles, abstracts, keywords, parameters, methods, temporal and geographic coverage.

Structural Metadata: describe the internal structure and organization of data.

Example: data tables, file formats (CSV, NetCDF), relations between files.

Administrative Metadata: describe management, rights, and provenance.

Example: license information, DOI assignment, versioning, funding project identifiers.

Provenance/Technical Metadata (sometimes included under administrative or treated separately in ISO frameworks): document how the data were created, transformed, and curated.

Example: instruments, laboratory methods, data processing workflows.

Community and Repository Alignment

When selecting metadata schemas, data producers should always consider the intended purpose of the metadata:

1. If datasets are to be published in repositories such as PANGAEA or GFZ Data Services, metadata must follow repository-specific workflows and schemas.
2. If datasets need to be interoperable within a scientific community (e.g., oceanography, climate science), community standards like CF Conventions, NetCDF Climate & Forecast metadata, or GCMD keywords should be adopted.
3. If datasets must be integrated into international portals and infrastructures, metadata must be aligned with globally recognized schemas such as ISO 19115 or DataCite.

5. Naming of communities that have already implemented the recommendation

6. Documentation of the test to validate correct implementation

7. Examples of Instances

8. Further Information

References

Relevant Community Recommendations

9. History of this document

From:
<https://earth-and-environment.helmholtz-metadaten.de/wiki/> - HMC
**Earth and Environment
Community Wiki**

Permanent link:
<https://earth-and-environment.helmholtz-metadaten.de/wiki/doku.php?id=wiki:s1.0&rev=1756297016>

Last update: 2025/08/27 12:16

