Recommendation S0

Recommendation for implementing harmonized semantic concepts in data infrastructures and products

Description

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

Motivation for this Recommendation

The use of shared, community-endorsed vocabularies for metadata annotation is key to ensuring unambiguous and standardized descriptions of data. This not only supports the alignment and integration of heterogeneous datasets but also enhances data discovery and reuse. Crucially, such practices form the foundation for machine-readability of metadata, which is essential for achieving semantic interoperability.

Within the Helmholtz research field Earth and Environment, there is a growing need for consistent approaches to metadata annotation that ensure semantic interoperability. This recommendation aims to address that need by guiding the selection and prioritization of controlled vocabularies and by supporting the optimization of metadata annotation workflows.

Recommendation summary

Data infrastructures and data hosts—such as data repositories, sensor registries, electronic lab notebooks, or other platforms that store and make data available—should ensure the annotation of the vast majority of metadata using standardized terms from established and, where appropriate, FAIR-compliant controlled vocabularies (e.g., lists, thesauruses, taxonomies, standardized terminologies, or, ideally, ontologies) to promote semantic consistency, clarity, and interoperability. Data platforms such as data portals or knowledge graphs should incorporate these terms into their tools and reuse them for standardized representation and improved search.

Binding Convention

	mandatory	conditional	optional
Helmholtz FAIR Principle		Annotation is mandatory when appropriate controlled vocabularies, expert recommendations on their use, and the necessary domain expertise are available, and when the systems support annotation technically.	

Precondition for Implementation

The basis for a comprehensive metadata annotation is the is that data is provided with sufficient and structured metadata and that there is agreement about which metadata is considered essential in communities. Metadata annotation with semantic ressources is only effective if there is consensus within a research community about which controlled vocabularies or other semantic resources best meet the community's needs, and if these resources have clear governance, provenance, and documentation. Furthermore, they should be available and maintained over the long term (at least 5 years) and cover the vast majority of requirements.

Contributors

Content

1. Explanation of the Background and Benefits of the Recommendation

Consistent implementation of semantic concepts across a research community is essential for advancing the FAIR principles—particularly Findability, Interoperability, and Reusability. By relying on shared, community-endorsed vocabularies and standardized metadata structures, data can be described in a clear and unambiguous way, enabling automated systems to interpret, link, and integrate datasets across disciplinary and institutional boundaries. This supports not only transparent and reproducible research but also facilitates the meaningful reuse of data. Especially in complex domains such as Earth and Environment, semantic concepts serve as a means to standardize metadata, promote interdisciplinary understanding, and represent complex scientific phenomena in a machine-readable, structured format.

In the current German research landscape, where environmental data are fragmented, inconsistent in quality, and lack standard formats, consistent semantic annotation is key to improving interoperability. Community agreement on essential metadata elements ensures data are well-described, discoverable, and reusable. This recommendation supports the Helmholtz Earth and Environment community by guiding the selection of controlled vocabularies and the improvement of annotation workflows.

What is meant with „controlled vocabulary“ in these recommendations?

There are different types of structured terminologies used in semantic data annotation, each offering varying levels of complexity and expressiveness. According to Le Franc et al. (2019), these can be understood as part of a spectrum of controlled vocabularies, ranging from simple to highly formalized structures. In our recommendations, the term controlled vocabularies is used in this broad sense and thus also includes ontologies.

• A **glossary** is a simple, alphabetically ordered list of terms from a specific domain, each accompanied by a definition. It supports a shared understanding of terminology within a community.

• A **taxonomy** is a controlled vocabulary with a hierarchical structure. Terms are related via broader–narrower (parent–child) relationships, and the taxonomy helps classify concepts and organize knowledge systematically.

• A **thesaurus** builds upon a taxonomy by incorporating not only hierarchical but also associative (e.g., related terms) and equivalence relationships (e.g., synonyms or preferred terms). Thesauri are useful for enhancing semantic navigation and improving information retrieval.

• An **ontology** represents the most expressive and formal type of controlled vocabulary. While it may include terms and structures from glossaries, taxonomies, or thesauri, it adds formal semantics through logical relationships defined in machine-readable languages (e.g., OWL, Description Logic). Ontologies enable reasoning, inference, and advanced semantic interoperability across systems.

In line with Le Franc et al. (2019), we explicitly consider ontologies to be part of the family of controlled vocabularies addressed in this recommendation. Depending on the use case and required level of formality, different types of controlled vocabularies may be appropriate for metadata annotation and semantic integration.

2. Possible alternative solutions

-

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

Aspect	Advantages	Challenges / Limitations
————————	————————————————————	————————————————————————————
Quality of content	Enhances clarity, consistency, and semantic richness of metadata.	Requires ongoing curation; quality depends on community engagement and domain expertise.
Interoperability	Enables cross-disciplinary data integration and supports machine-readability.	Limited if vocabularies are poorly aligned, domain-specific, or not widely adopted.
Sustainability	Promotes long-term reuse through shared standards and semantic resources.	Many vocabularies/tools lack sustained funding and maintenance; often limited to project durations.
Technical availability	Supports automation, validation, and FAIR-aligned workflows.	Tools and vocabularies may become obsolete or unavailable without long-term infrastructure support.
Community fit	Encourages reuse of existing vocabularies and avoids duplication of effort.	Existing vocabularies may not fully meet community needs; new ones are difficult to build and maintain.
Funding and effort	Shared solutions improve efficiency and reduce redundant work.	High initial effort; long-term success requires stable funding, governance, and institutional backing.

> **Note:** While the benefits of standardized semantic practices are clear, their successful implementation depends on collective coordination, sustainable infrastructure, and adequate funding. Efforts should focus on reusing and adapting existing resources where possible, rather than creating isolated or redundant solutions.

4. The Recommendation

Data stewards, archivists, and tool developers—including those responsible for systems used at various stages of the data lifecycle, such as data acquisition, processing, documentation, and storage—should ensure that metadata is captured in a structured and standardized manner, using harmonized metadata schemas aligned with community standards. This includes platforms such as electronic lab notebooks, archiving tools, sensor registries, and other software environments that support data generation, transformation, or submission. Metadata must be consistently annotated with well-governed controlled vocabularies to guarantee semantic clarity, interoperability, and long-term reusability across diverse data infrastructures. Providing clear documentation of the vocabularies and semantic resources in use, alongside transparent, user-friendly annotation

workflows, supports consistent metadata quality and facilitates semantic integration.

Developers of data portals, knowledge graphs, and discovery tools should incorporate these controlled vocabularies and ontologies into their software environments. This enhances machine-readability, promotes semantic consistency across systems, and enables users to efficiently search, filter, and combine data from multiple sources.

To enable seamless semantic annotation from the start, data producers need to be supported through targeted training and awareness initiatives that emphasize the use of community-endorsed vocabularies, structured metadata practices, and annotation best practices. Transparent user guidance and easily accessible documentation of recommended semantic resources are essential to ensure metadata quality and simplify the semantic linkage of data throughout its lifecycle.

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

- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). [The FAIR Guiding Principles for scientific data management and stewardship](https://doi.org/10.1038/sdata.2016.18). *Scientific Data*, 3, 160018. https://doi.org/10.1038/sdata.2016.18

- Le Franc, Y., Hettne, K., & Ó Carragáin, E. (2019). *D2.5 FAIR Semantics Recommendations Second Iteration*. Zenodo. https://doi.org/10.5281/zenodo.4314321

Relevant Community Recommendations

9. History of this document