The Helmholtz Digitization Ontology: Representing Digital Assets in the Helmholtz Digital Ecosystem

Said Fathalla[1,*], Gerrit Günther[2], Leon Steinmeier[3], Christine Lemster[4], Dorothee Kottmeier[4,6], Lakxmi Sivapatham[5], Pier Luigi Buttigieg[4], Volker Hofmann[1,*] and Stefan Sandfeld[1]

[1] *Forschungszentrum Jülich GmbH, Institute for Advanced Simulation – Materials Data Science and Informatics (IAS-9), Jülich, Germany*

[2] *Helmoltz-Zentrum Berlin für Materialien und Energie, Berlin, Germany*

[3] *Helmholtz Center Dresden Rossendorf, Dresden, Germany*

[4] *GEOMAR Helmholtz-Zentrum für Ozeanforschung, Kiel, Germany*

[5] *Deutsches Zentrum für Luft und Raumfahrt, Cologne, Germany*

[6] *Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung, Bremerhaven, Germany*

Abstract

The Helmholtz Association is actively digitizing research outcomes to drive progress and innovation. The vast volumes of digital data are diverse in terms of their formats and the semantic descriptions used in their interchange, and storage. Therefore, a semantic frame of reference is required to facilitate interoperability throughout the Helmholtz digital ecosystem. This paper presents the Helmholtz Digitization Ontology (HDO), which is intended to serve that purpose. HDO is a mid-level ontology that contains concepts representing digital assets relevant to the Helmholtz digital ecosystem, data creation, management, and exchange. It is developed within the framework of the Helmholtz Metadata Collaboration (HMC) with contributors from various scientific backgrounds. HDO serves as a harmonized semantic framework and machine-actionable reference across all Helmholtz research fields.

Keywords

Metadata, Metadata Management, Data Management, FAIR, Harmonization, OWL, Bilingual ontology

1. Introduction

The Helmholtz Association comprises 18 research centers operating across Germany, focusing on a wide range of scientific topics and methods within six research areas fields[1]. The Helmholtz Metadata Collaboration (HMC)[2] is an association-wide operating platform that supports (meta)data harmonization and information engineering across all research fields intending to make data within Helmholtz adhere to the FAIR principles [1] and establish an interoperable FAIR data space.

Motivation and requirements. The heterogeneity of scientific contexts within Helmholtz leads to ambiguity and conflicts regarding metadata semantics, e.g. in developed metadata

SEMANTiCS 2024: 20th International Conference on Semantic Systems, Sep 17–19, 2024, Amsterdam, The Netherlands

*Corresponding author.

s.fathalla@fz-juelich.de (S. Fathalla); v.hofmann@fz-juelich.de (V. Hofmann)

[1] https://www.helmholtz.de/en/research/research-fields/

[2] https://www.helmholtz-metadaten.de/en

schemas, tools or general communication between collaborators, when data is exchanged in interdisciplinary collaboration. Due to their ability to establish clear context and relationships between concepts [2], ontologies are widely used towards facilitating efficient and interoperable data management and data exploitation. As such, ontologies are important towards making data FAIR, specifically towards achieving interoperability and machine actionability. Thus, HMC realized that it is required to provide a semantic frame of reference for all stakeholders involved in Helmholtz's digitization efforts. Contributors from all Helmholtz research fields have been involved in this process.

Objectives. The main objectives of the Helmholtz Digitization Ontology (HDO) are: 1) Creating a standardized semantic framework with terminology that can reduce semantic uncertainty and ambiguity and thereby increase semantic interoperability between various Helmholtz systems. 2) Facilitating data integration in different Helmholtz systems, e.g., the institutional Helmholtz Knowledge Graph [3] or domain-specific use cases. 3) Providing a basis for reasoning based on existing data to allow inferring new knowledge, e.g., towards predictive analyses. 4) Supporting harmonized knowledge management within Helmholtz to facilitate decision-making and the preservation of research findings.

2. Concepts Definitions

In HDO, we provide well-defined concepts with rigorous semantic and unambiguous definitions. Class definitions aim to: 1) outline the intrinsic characteristics of the term being defined, 2) avoid circularity, 3) be neither excessively broad (to avoid ambiguity), nor too narrow (to allow implementation of further sub-classes where necessary), and 4) be easily understood using common, unambiguous terminology, which is important, especially concerning HDO's purpose of serving as a mid-level ontology that provides common understanding and reduces miscommunication across different research fields.

To create class definitions, we follow Aristotelian logic, specifically, definitions adhere to the genus-differentia form[3]. Genus-differentia definitions follow the form: "`A (the class label) is a B (the genus or superclass) which is C (the differentia)`". For instance, consider the definition of "`JSON file`", which is "`A file which conforms to JSON format`". Here, the genus part is "file", which is the superclass of `JSON file`, from which all differentia and properties are inherited. The remaining part is the differentia, which comprises the features distinguishing the currently defined term from its genus and siblings.

3. Development Strategy

The key aspects of HDO development are illustrated in Figure 1. The development is carried out in three phases: *initialization*, *implementation*, and *adoption and adaption.*

1. Initialization phase: In the early stages of the development, an internal GitLab repository was created to gather a set of core terms and their definitions in per-term YAML files. We followed a template with keys such as `definition`, `synonyms`, `comments`, and `seeAlso`.

[3] https://en.wikiversity.org/wiki/Dominant_group/Genus_differentia_definition

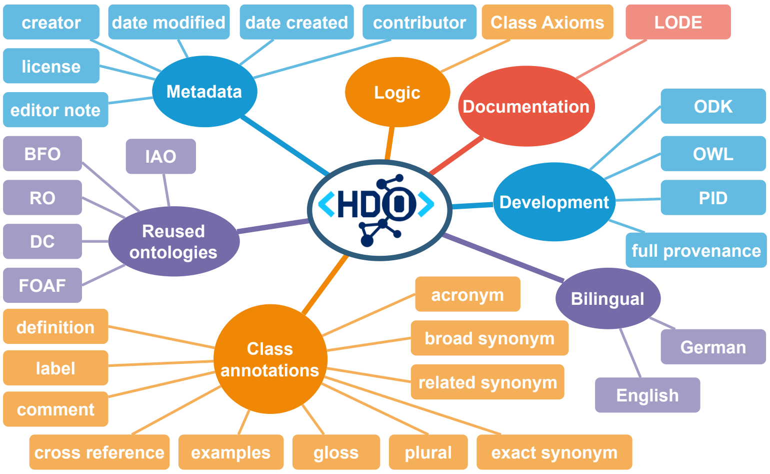

Figure 1: Key development aspects used during HDO development.

2. Implementation phase: After collecting terms, YAML files were converted and merged into one OWL file, where keys of the template were mapped onto existing and imported annotation properties (e.g. `definition` was mapped to `iao:definition`). Further development was carried out in the OWL file in a public repository[4], and managed using the Ontology Development Kit [4]. HDO is made accessible via a persistent identifier (https://purls.helmholtz-metadaten.de/hob/hdo.owl) and terms are derefrenced via their IRIs (e.g., HDO_00004001). PIDA is used to dereference a single ontology term IRI, and display the HTML documentation[5] on a web browser for human users or provide the OWL file for machines. The class hierarchy is extended according to the following workflow: 1) Contributors create a GitLab issue and propose a term definition and properties, 2) Collaborative discussion was carried out within the issue thread. 3) Upon general agreement, the class was implemented in the development file (i.e., hdo-edit.owl) within separate, sequentially generated branches, and 4) A merge request is created, and upon approval by at least three contributors, these branches are merged into the main.

3. Adoption and Adaption: Upon publication of HDO, it will be used in use cases across the different Helmholtz research fields. This will test the ontology against use case-specific requirements and allow further adaption based on iterative exchange.

3.1. Reuse of existing terms

We focused on reusing classes from existing, well-known semantic artifacts, wherever possible, to ensure semantic interoperability. HDO is top-level aligned with the Basic Formal Ontology (BFO)[6] to ensure interoperability with other mid- and domain-level ontologies. Furthermore, classes and properties from well-known Open Biological and Biomedical Ontologies (OBO)

[4] https://codebase.helmholtz.cloud/hmc/hmc-public/hob/hdo
[5] https://purls.helmholtz-metadaten.de/hob/HDO_00000000
[6] https://obofoundry.org/ontology/bfo

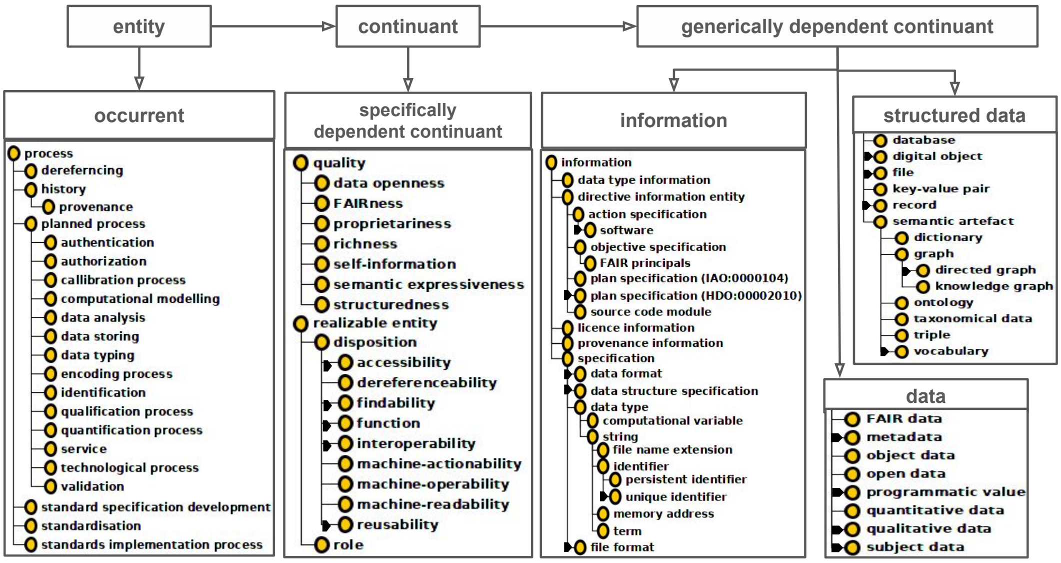

Figure 2: An overview of HDO class hierarchy. Open arrowheads denote `subClassOf` properties between the classes. For visualization purpose, certain classes have been excluded from the representation.

were re-used, including: Information Artifact Ontology (IAO)[7] (`iao:action specification`, `iao:information content entity`, `iao:plan specification`) and the Relation Ontology (RO)[8] (`ro:input of`, `ro:has characteristic` and `ro:has input`).

3.2. Concepts overview

An overview of the HDO core classes is shown in Figure 2. HDO establishes semantics for the core concepts of digital infrastructure, digital information management, and processes in data management and exchange. For example, HDO offers a rigorous semantic context for the aspects of the FAIR principles. These concepts are modelled as `bfo:disposition` that inhere in `hdo:data` (HDO_00000009) either according to a practical understanding of `hdo:findability`, as well as specified according to the FAIR principles as `hdo:findability according to the FAIR principles`. Further, we created classes for different aspects of metadata standardization under `hdo:information` and aligned this with IAO classes (e.g., `iao:information content entity`).

3.3. Logic

We asserted pairwise disjointness between mutually disjoint classes, e.g., `hdo:digital infrastructure` is disjoint with `hdo:hardware`. We used `bfo:role` and `bfo:realizes` to create a pattern that allows populating certain classes by inference. For example, the class `hdo:agent` is *"An entity which realises an agent role"*. This is inferred through `hdo:agent`

[7] https://github.com/information-artifact-ontology/IAO/
[8] https://obofoundry.org/ontology/ro.html

`role` and `bfo:realizes`. A similar example is the class `hdo:tool` which is defined as *"A continuant which realises a tool role"*. For this, as well as to allow reasoning about relevant concepts, several OWL axioms have been asserted in HDO. The `EquivalentClasses` axiom allows to state that several class expressions are equivalent to each other. For instance, the class `hdo:data` has the following axiom assertion:

```
'generically dependent continuant'  and
      ('output of' some ('encoding process'  and ('has input' some signifier)))
```

The `SubClassOf` axiom allows to state that each instance that fits a class expression is also an instance of that class. For instance, the class `hdo:structured vocabulary` has the following `SubClassOf` axiom assertion:

```
vocabulary and 'has quality' some structuredness
```

4. Conclusions and Future Work

The Helmholtz Digitization Ontology was developed with the objective of providing harmonized semantics as a reference framework for the Helmholtz digital ecosystem. The development of HDO was open and transparent, and its full provenance is recorded. Further, we follow an established ontology development framework (e.g., ODK) and align HDO with well-established semantic frameworks in order to increase acceptance towards domain-level re-use and application. One of the further use cases will be the semantic representation of FAIR digital objects (FDOs) that will allow data integration between FDOs and the HMC Helmholtz KG [3]. Such implementations will extend HDO core semantics and facilitate the representation, interoperability, and analysis of scientific metadata within the Helmholtz digital ecosystem.

References

[1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The FAIR guiding principles for scientific data management and stewardship, Scientific data 3 (2016).

[2] S. Fathalla, S. Vahdati, C. Lange, S. Auer, Seo: A scientific events data model, in: The Semantic Web – ISWC 2019, Springer International Publishing, Cham, 2019, pp. 79–95.

[3] J. Bröder, G. Preuß, F. D'Mello, S. Fathalla, V. Hofmann, S. Sandfeld, The Helmholtz knowledge graph: driving the transition towards a FAIR data ecosystem in the Helmholtz Association, in: European Semantic Web Conference, Springer, 2024.

[4] N. Matentzoglu, D. Goutte-Gattat, S. Z. K. Tan, J. P. Balhoff, S. Carbon, A. R. Caron, W. D. Duncan, J. E. Flack, M. Haendel, N. L. Harris, W. R. Hogan, C. T. Hoyt, R. C. Jackson, H. Kim, H. Kir, M. Larralde, J. A. McMurry, J. A. Overton, B. Peters, C. Pilgrim, R. Stefancsik, S. M. Robb, S. Toro, N. A. Vasilevsky, R. Walls, C. J. Mungall, D. Osumi-Sutherland, Ontology Development Kit: a toolkit for building, maintaining and standardizing biomedical ontologies, Database 2022 (2022).