HELMHOLTZ Metadata Collaboration

Helmholtz Metadata Collaboration

List of Standards for Matter

July 2025

HELMHOLTZ

Research for grand challenges.

Contents

List of Standards for Matter

Özlem Özkan, Oonagh Brendike-Mannix, Gerrit Günther, Markus Kubin & Heike Görzig

Helmholtz-Zentrum Berlin für Materialien und Energie & Helmholtz Metadata Collaboration Hub Matter

Preamble

This list provides an overview of data standards relevant to the Helmholtz research field Matter. It was compiled by integrating several sources: results from the HMC 2021 Community Survey (Helmholtz Metadata Collaboration, 2022) (in which participants identified data formats and standards used or generated in recent research projects), standards associated with specific use cases, manual mapping of standards found in repositories relevant to the field (Kubin et al., 2024), and additional manual searches.

To support easier navigation and application, the list is organized into three sections:

- Standards in Matter – standards that are specifically used within the research field Matter;
- Generic Standards – broadly applicable standards that are not domain-specific but relevant;
- Generic (Meta)Data Formats or Models – data formats or models which often lack semantic enrichment but support interoperability and data management.

Please note that this list is not exhaustive. It is intended as a starting point to help community members identify and explore relevant data standards.

Matter-specific standards

ANZLIC Metadata Profile

Link: https://www.anzlic.gov.au/

The ANZLIC Metadata Profile is a geospatial metadata standard based on ISO 19115 and the AGLS extension of Dublin Core. It is used for describing and discovering spatial and geographic datasets and services, particularly in Australia and New Zealand. Although primarily relevant to Earth science and environmental research, it has potential relevance in geospatially aware astronomy datasets (Karlsruhe Institute of Technology, n.d.).

Astronomy Visualization Metadata (AVM)

Link: https://www.virtualastronomy.org/avm_metadata.php

AVM is a metadata schema developed to support cross-searching of astronomy images, especially those generated for public and educational use. It is compatible with the Adobe XMP specification and can be embedded in formats such as JPEG, PNG, and TIFF (Karlsruhe Institute of Technology, n.d.).

Core Scientific Metadata Model (CSMD)

Link: http://icatproject-contrib.github.io/CSMD/

The Core Scientific Metadata Model (CSMD) is a data-oriented framework developed at STFC, designed to capture comprehensive information about scientific studies and the data they generate. It serves as the foundation for the ICAT schema, although ICAT incorporates additional implementation-specific features. CSMD supports the organization and management of metadata related to scientific research activities. Please note that CSMD was last updated in 2013. Inclusion here is for completeness and potential users should instead consider the ICAT schema (Karlsruhe Institute of Technology, n.d.).

Crystallographic Information Framework (CIF)

Link: https://www.iucr.org/resources/cif

CIF, which stands for Crystallographic Information File, serves as a standard format for data exchange in crystallography. It also refers to the Crystallographic Information Framework, which encompasses a set of exchange protocols based on data dictionaries and relational rules that can be represented in various machine-readable formats, including XML (International Union of Crystallography, n.d.).

Electron Microscopy (EM) Glossary

Link: https://emglossary.helmholtz-metadaten.de/

The EM Glossary is an initiative to standardize terminology in electron microscopy by bringing together experts in the field with metadata and information engineering professionals. It aims to create a harmonized vocabulary to enhance communication and interoperability within the electron microscopy community (HMC, n.d.).

FITS (Flexible Image Transport System)

Link: https://fits.gsfc.nasa.gov/

FITS is the standard data format used in astronomy for storing, transmitting, and processing scientific and image data. It supports metadata in ASCII headers and enables complex data structures like images, tables, and multidimensional arrays (Karlsruhe Institute of Technology, n.d.).

FITS World Coordinate System (WCS)

Link: https://fits.gsfc.nasa.gov/fits_wcs.html

WCS is an extension to the FITS format that provides a standard for mapping image pixels to physical (sky) coordinates. It was proposed in 2002 and integrated into the 3.0 release of the FITS standard to enable celestial referencing in image data (Karlsruhe Institute of Technology, n.d.).

ICAT Schema

Link: https://repo.icatproject.org/site/icat/server/5.0.0/schema.html

The ICAT Schema extends the Core Scientific Metadata Model (CSMD) with practical elements for managing experimental data at large-scale facilities. It is typically used for internal organization of the ICAT metadata catalogue, and should be understood as such (ICAT Project, n.d.).

IMPEx Data Model

Link: https://impex-fp7.oeaw.ac.at/

The IMPEx (Integrated Medium for Planetary Exploration) Data Model extends the SPASE model to support planetary and heliospheric simulations, providing interoperability across simulation tools and data archives (Karlsruhe Institute of Technology, n.d.).

IUPAC Gold Book

Link: https://goldbook.iupac.org/

The IUPAC Compendium of Chemical Terminology, known as the Gold Book, provides authoritative definitions of terms in chemistry. It serves as a reference for standardized chemical nomenclature and terminology (IUPAC, n.d.).

International Virtual Observatory Alliance Technical Specifications

Link: https://www.ivoa.net/documents/

The International Virtual Observatory Alliance (IVOA) develops and maintains a suite of metadata standards and models to enable interoperability of astronomical archives worldwide. These include models for photometry, spectra, time coordinates, and simulation data (Karlsruhe Institute of Technology, n.d.).

Metadata Schemas for Materials Science (KIT DM)

Link: https://github.com/kit-data-manager/Metadata-Schemas-for-Materials-Science

The Metadata Schemas for Materials Science initiative, developed by the Karlsruhe Institute of Technology (KIT), provides a modular and extensible framework for describing experimental and simulation data in materials science (KIT DM, n.d.).

NeXus and the NeXus Definition Language (NXDL)

Link: https://www.nexusformat.org/

NeXus is an international standard created by scientists and programmers to enhance collaboration in the analysis and visualization of data from neutron, x-ray, and muon experiments. It aims to provide a standardized data format based on HDF5 (see below) and the NeXus Definition Language (NXDL) terminology to facilitate data sharing and interoperability among scientific communities (NeXus Data Format, n.d.).

OME-xml (newer version OME-tiff)

Link: http://www.openmicroscopy.org/Schemas/

OME-XML is a standardized metadata schema for biological light microscopy data, developed by the Open Microscopy Environment Consortium. It is vendor-neutral and supports metadata for light microscopy experiments. OME-TIFF is an extension of OME-XML, allowing metadata to be embedded within TIFF or BigTIFF files, ensuring consistent data and metadata integration for biological imaging (Open Microscopy Environment, n.d.).

Open Databases Integration for Materials Design (OPTIMADE)

Link: https://www.optimade.org/

OPTIMADE is a consortium-driven initiative aimed at enhancing interoperability between materials science databases. It establishes a common REST API specification, enabling consistent and standardized data access across different materials databases. The goal is to promote data sharing and integration within the materials design community (OPTIMADE Consortium, n.d.).

Open Particle-Mesh Data (OpenPMD)

Link: https://github.com/openPMD/openPMD-standard

OpenPMD, or Open Particle-Mesh Data, is a standard for metadata and naming conventions used to exchange particle and mesh-based data from scientific simulations and experiments. It is not a file format but provides standardized naming and attribute guidelines to promote interoperability in scientific data sharing (OpenPMD Collaboration, n.d.).

Open Reflectometry Standards Organisation (ORSO)

Link: https://www.reflectometry.org/

The Open Reflectometry Standards Organisation (ORSO) is a global community of scientists dedicated to advancing neutron and X-ray reflectometry through international collaboration and standardization. ORSO focuses on improving reproducibility, interoperability, and data sharing within the reflectometry community (Open Reflectometry Standards Organisation ORSO, n.d.).

Protein Data Bank Exchange Dictionary and the Macromolecular Crystallographic Information Framework (PDBx/mmCIF)

Link: https://mmcif.wwpdb.org/

The Protein Data Bank (PDB) is the central global archive for 3D structures of proteins, nucleic acids, and complex biological assemblies, managed by the Worldwide PDB (wwPDB). The PDB Exchange Dictionary (PDBx) is utilized to define the metadata for the deposition, annotation, and storage of PDB entries. It incorporates the Macromolecular Crystallographic Information Framework (mmCIF), which was developed under the guidance of the International Union of Crystallography (IUCr). The PDBx framework has been extended to include metadata for other structural determination methods, such as Nuclear Magnetic Resonance Spectroscopy and 3D Electron Microscopy (Worldwide Protein Data Bank, n.d.).

Quantities, Units, Dimensions and Types (QUDT)

Link: http://qudt.org/

QUDT (Quantities, Units, Dimensions, and Types) provides ontologies that define the fundamental classes, properties, and limitations necessary for representing physical quantities, measurement units, and dimensions across various systems. It establishes a cohesive framework for measurable quantities, units, numerical values in different units, and the data structures used to store and manipulate them within software systems (QUDT.org, n.d.).

Resource Metadata for the Virtual Observatory

Link: https://www.ivoa.net/documents/ResourceMetadata/

This standard defines the metadata necessary for discovering and using astronomical data services. It extends the Dublin Core model with astronomy-specific elements and is used in the Virtual Observatory registry systems (Karlsruhe Institute of Technology, n.d.).

Sample Environment Communication Protocol (SECoP)

Link: https://content.iospress.com/articles/journal-of-neutron-research/jnr190143

SECoP is a global communication standard that facilitates interaction between sample environment equipment and experimental control software, particularly in neutron and photon research facilities. It ensures seamless integration and communication within scientific research infrastructures (Schulz, M., et al., 2019).

Standard for Documentation of Astronomical Catalogues (SDAC)

Link: https://cds.unistra.fr/doc/catstd.htx

Developed by the Centre de Données astronomiques de Strasbourg (CDS), SDAC defines conventions for archiving astronomical data in human-readable formats. It specifies naming, structure, and metadata requirements for tabular data and associated documentation (Karlsruhe Institute of Technology, n.d.).

SPASE Data Model

Link: https://spase-group.org/data/model/

The Space Physics Archive Search and Extract (SPASE) Data Model is designed for describing heliophysics data and its scientific context, including data sources, content, provenance, and access. It supports a federated data system, and the preferred serialization format is XML. It is maintained by the SPASE Consortium (Karlsruhe Institute of Technology, n.d.).

XAS Data Interchange Format (XDI)

Link: https://github.com/XraySpectroscopy/XAS-Data-Interchange/blob/master/specification/spec.md

XDI is a standard file format for storing X-ray Absorption Spectroscopy (XAS) data. It provides a consistent and well-defined structure for tabular data and metadata associated with XAS experiments. Designed for interoperability and archival purposes, XDI ensures reproducibility and easier sharing of XAS datasets across the scientific community (XraySpectroscopy GitHub, n.d.).

Generic Standards

Common European Research Information Format (CERIF):

Link: https://eurocris.org/Uploads/Web%20pages/CERIF-1.6/documentation/MInfo.html

CERIF is a metadata standard recommended by the European Union for documenting research activities. It supports the structured recording of metadata related to research projects, publications, datasets, and other research outputs. CERIF is maintained by EuroCRIS and promotes interoperability within the European research ecosystem (EuroCRIS, n.d.).

CSV on the Web (CSVW):

Link: https://csvw.org/standards.html

CSVW extends the basic CSV format by including machine-readable metadata in JSON format. It defines table structures, column data types, and relationships, making tabular data more interoperable and easier to integrate across systems. It improves interoperability of CSV files by providing machine-readable descriptions of the data structure, such as column names, data types, and relationships (W3C, n.d.-a).

Datacite Metadata Schema

Link: https://schema.datacite.org/

The DataCite Metadata Schema is designed to enable accurate and consistent identification of research outputs for citation and retrieval. It defines core metadata elements applicable to various resource types, particularly datasets. This schema supports the documentation and dissemination of research objects, contributing to DataCite's mission of advancing data sharing and accessibility (DataCite, n.d.).

DCAT

Link: http://www.w3.org/TR/vocab-dcat/

DCAT is an RDF-based vocabulary that facilitates interoperability between data catalogs on the web. It provides a standardized schema to describe datasets and data services, enabling data aggregation from multiple catalogs and enhancing dataset discoverability through federated search mechanisms (W3C, n.d. -b).

Dublin Core 1.1

Link: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/#source

Dublin Core is a set of fifteen metadata elements used for resource description and cataloging. It is part of the Dublin Core Metadata Initiative (DCMI) and provides a standardized vocabulary to describe resources consistently across digital libraries, repositories, and data catalogs, enhancing discoverability and interoperability (Dublin Core Metadata Initiative, n.d.).

ISO-Standards

Link: https://www.iso.org/standards.html

ISO Standards are globally recognized guidelines developed by the International Organization for Standardization. They establish consistent requirements and best practices across different industries and sectors, ensuring quality, safety, and efficiency. ISO standards are widely adopted for interoperability, compliance, and international trade (International Organization for Standardization, n.d.).

PROV-O

Link: https://www.w3.org/TR/prov-o/

PROV-O, or Provenance Ontology, is a framework designed using the OWL2 Web Ontology Language (OWL2) to represent provenance information. It provides a set of classes, properties, and constraints to model the origin and history of data across various systems and contexts. PROV-O is flexible and can be customized to create domain-specific provenance models, supporting data traceability and accountability (W3C, n.d.-c).

Schema.org

Link: https://schema.org/

Schema.org is a shared vocabulary that supports structured data markup for websites, enabling improved data visibility and search engine optimization. It uses various encodings, including RDFa, Microdata, and JSON-LD, to describe entities, their attributes, and relationships. Schema.org is widely used across the web to enhance discoverability and semantic interoperability (Schema.org, n.d.).

Generic (Meta)data Formats or Models

Citation File Format

Link: https://citation-file-format.github.io/

Citation File Format (CFF) provides a standardized way for researchers and software developers to share citation information. CITATION.cff files are plain text files that contain metadata about the software or datasets, including author names, title, version, license, and how to cite the resource. These files are human-readable and machine-readable, enabling consistent and accurate citations. Developers can include them in their software repositories to guide users on the proper citation format (Citation File Format Initiative, (n.d.).

JSON

Link: https://ecma-international.org/publications-and-standards/standards/ecma-404/

JSON is a lightweight, text-based syntax for data interchange that is platform-independent and easy to read for both humans and machines. Originally derived from JavaScript, JSON is widely used for structuring data in web applications and APIs. It provides a standardized way to represent complex data structures using key-value pairs (Ecma International, 2017).

JSON-LD

Link: https://www.w3.org/TR/json-ld11/

JSON-LD is an extension of JSON that is specifically designed for representing Linked Data. It enables data to be connected and Linked across different systems using unique identifiers (URIs), promoting semantic interoperability. JSON-LD is commonly used in web-based programming environments to integrate Linked Data with existing JSON-based systems (W3C, 2020).

HDF5

Link: https://www.hdfgroup.org/solutions/hdf5/

HDF5 (Hierarchical Data Format version 5) is a data model, library, and file format designed for managing large and complex datasets. It provides a hierarchical structure that supports metadata, compression, and parallel I/O, making it highly efficient for storing and retrieving scientific data. HDF5 is widely used in various fields, including physics, engineering, and geosciences, for handling structured and unstructured data in high-performance computing environments (The HDF Group, n.d.).

odML

Link: http://g-node.github.io/python-odml/

odML is a structured file format for storing metadata in a human- and machine-readable way. It is designed to organize metadata related to scientific experiments, ensuring reproducibility and traceability. odML supports hierarchical data organization and is commonly used in neuroscience and biomedical research (German Neuroinformatics Node, n.d.).

RDF

Link: https://www.w3.org/RDF/

RDF is a framework designed for data exchange on the web. It enables the integration of data from different sources even if they use distinct schemas. RDF is particularly advantageous for maintaining data compatibility over time, as it supports the evolution of schemas without requiring changes from all data users. It provides a flexible and consistent way to share and merge information across diverse systems (W3C, n.d. -d).

RO-Crate

Link: https://w3id.org/ro/crate

RO-Crate is a standard for organizing and packaging research data as a "Research Object" (RO). It consists of a set of data files (referred to as a crate) along with a ro-crate-metadata.json file that describes the contents. This metadata file provides detailed information about each item, including authorship, licensing, identifiers, and provenance, making the data understandable to both humans and machines.

The collection can include various types of research outputs such as papers, datasets, software, or references, either as a folder of files or as a collection of Linked references. Any directory can be transformed into an RO-Crate by simply adding a ro-crate-metadata.json file.

RO-Crate ensures that research data is distributed along with its metadata, enhancing traceability, archiving, and proper attribution. It utilizes the JSON-LD format for metadata representation, which is an extension of JSON. Metadata can be written manually or generated using available tools. Additionally, some workflow management systems, like Galaxy, support exporting data as RO-Crates (RO-Crate Community, n.d.).

ROOT

Link: https://root.cern/

ROOT is an object-oriented framework developed at CERN, designed for high-energy physics data processing and analysis. It provides a powerful file format, data storage, and statistical analysis tools tailored for large-scale scientific computations. ROOT files can store complex data structures efficiently and integrate with advanced visualization techniques, making them a standard format in particle physics and related research areas (CERN, n.d.).

SKOS

Link: https://www.w3.org/TR/skos-reference/

SKOS is a lightweight data model for organizing and Linking knowledge systems on the web. It is built on RDF (Resource Description Framework) and OWL (Web Ontology Language), allowing for the sharing and integration of knowledge organization systems on the Semantic Web. SKOS is commonly used for creating controlled vocabularies, taxonomies, and thesauri. Its data can be represented in multiple RDF-compatible formats, such as RDF/XML or Turtle, to enhance interoperability across systems (W3C, n.d.-e).

XML

Link: https://www.w3.org/XML/

XML is a flexible, text-based format for representing structured data. It is widely used for data interchange, electronic publishing, and configuration files. XML allows for custom tags and hierarchical structuring, making it suitable for a wide range of applications, from document storage to web services (W3C, n.d.-f).

References

CERN. (n.d.). ROOT Data Analysis Framework. https://root.cern/. Retrieved March 4, 2025.

Citation File Format Initiative. (n.d.). Citation File Format. https://citation-file-format.github.io/. Retrieved March 4, 2025.

DataCite. (n.d.). DataCite Metadata Schema. https://schema.datacite.org/. Retrieved March 4, 2025.

Dublin Core Metadata Initiative. (n.d.). Dublin Core Metadata Terms. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/#source. Retrieved March 4, 2025.

Ecma International. (2017). Standard ECMA-404: The JSON Data Interchange Syntax. https://ecma-international.org/publications-and-standards/standards/ecma-404/. Retrieved March 4, 2025.

EuroCRIS. (n.d.). CERIF: Common European Research Information Format. https://eurocris.org/Uploads/Web%20pages/CERIF-1.6/documentation/MInfo.html. Retrieved March 4, 2025.

German Neuroinformatics Node. (n.d.). odML: Open Metadata Markup Language. http://g-node.github.io/python-odml/. Retrieved March 4, 2025.

Helmholtz Metadata Collaboration. (2022). A survey on research data management practices among researchers in the Helmholtz Association. https://doi.org/10.3289/HMC_publ_05

Helmholtz Metadata Collaboration. (n.d.). Electron Microscopy Glossary. https://emglossary.helmholtz-metadaten.de/. Retrieved March 4, 2025.

ICAT Project. (n.d.). ICAT Schema. https://repo.icatproject.org/site/icat/server/5.0.0/schema.html. Retrieved March 4, 2025.

IUPAC. (n.d.). Gold Book: Compendium of Chemical Terminology. https://goldbook.iupac.org/. Retrieved March 4, 2025.

International Organization for Standardization. (n.d.). ISO Standards Catalogue. https://www.iso.org/standards.html. Retrieved March 4, 2025.

International Union of Crystallography. (n.d.). Crystallographic Information Framework (CIF). https://www.iucr.org/resources/cif. Retrieved March 4, 2025.

Karlsruhe Institute of Technology. (n.d.). Metadata Standards Catalog Matter. Helmholtz Metadata Collaboration. https://msc.datamanager.kit.edu/matter. Retrieved July 22, 2025.

Kubin, M., Sedeqi, M. R., Schmidt, A., Gilein, A., Glodowski, T., Serve, V., Günther, G., Weisweiler, N. L., Preuß, G., & Mannix, O. (2024). *A data-driven approach to monitor and improve open and FAIR research data in a federated research ecosystem.* Data Science Journal, 23(1), 41, 1–10. https://doi.org/10.5334/dsj-2024-041

Metadata Schemas for Materials Science (KIT DM). (n.d.). https://github.com/kit-data-manager/Metadata-Schemas-for-Materials-Science. Retrieved May 7, 2025.

NeXus Data Format. (n.d.). NeXus and NXDL. https://www.nexusformat.org/. Retrieved March 4, 2025.

OPTIMADE Consortium. (n.d.). Open Databases Integration for Materials Design. https://www.optimade.org/. Retrieved March 4, 2025.

Open Microscopy Environment. (n.d.). OME-XML and OME-TIFF. http://www.openmicroscopy.org/Schemas/. Retrieved March 4, 2025.

Open Reflectometry Standards Organisation (ORSO). (n.d.). https://www.reflectometry.org/. Retrieved May 7, 2025.

OpenPMD Collaboration. (n.d.). Open Particle-Mesh Data (OpenPMD). https://github.com/openPMD/openPMD-standard. Retrieved March 4, 2025.

QUDT.org. (n.d.). Quantities, Units, Dimensions and Types (QUDT). http://qudt.org/. Retrieved March 4, 2025.

RO-Crate Community. (n.d.). RO-Crate Metadata Specification. https://w3id.org/ro/crate. Retrieved March 4, 2025.

Schulz, M., et al. (2019). Sample Environment Communication Protocol (SECop). Journal of Neutron Research. https://content.iospress.com/articles/journal-of-neutron-research/jnr190143. Retrieved March 4, 2025.

W3C. (2020). JSON-LD 1.1. https://www.w3.org/TR/json-ld11/. Retrieved July 22, 2025.

W3C. (n.d.-a). CSV on the Web (CSVW). https://csvw.org/standards.html. Retrieved July 22, 2025.

W3C. (n.d.-b). DCAT Vocabulary. http://www.w3.org/TR/vocab-dcat/. Retrieved July 22, 2025.

W3C. (n.d.-c). PROV-O: The PROV Ontology. https://www.w3.org/TR/prov-o/. Retrieved July 22, 2025.

W3C. (n.d.-d). RDF 1.1 Concepts and Abstract Syntax. https://www.w3.org/RDF/. Retrieved July 22, 2025.

W3C. (n.d.-e). SKOS: Simple Knowledge Organization System. https://www.w3.org/TR/skos-reference/. Retrieved July 22, 2025.

W3C. (n.d.-f). Extensible Markup Language (XML). https://www.w3.org/XML/. Retrieved July 22, 2025.

Worldwide Protein Data Bank. (n.d.). PDBx/mmCIF Dictionary. https://mmcif.wwpdb.org/. Retrieved March 4, 2025.

XraySpectroscopy GitHub. (n.d.). XAS Data Interchange Format (XDI). https://github.com/XraySpectroscopy/XAS-Data-Interchange/blob/master/specification/spec.md. Retrieved March 4, 2025.