HELMHOLTZ
Metadata Collaboration

HMC Conference

12.-14.05.2025, Cologne

"...because context matters"

Book of Abstracts

Overview

Intro

Welcome to HMC Conference 2025!

This comprehensive compilation of abstracts showcases the cutting-edge research and innovative ideas that will be presented at the Helmholtz Metadata Collaboration (HMC) Conference 2025.

The HMC Conference 2025 brings together experts from academia, industry, and research institutions to share their latest findings and advancements in the field of metadata management. As a key initiative of the Helmholtz Association, the HMC aims to address the challenges and opportunities of metadata management in scientific research and beyond.

The abstracts included in this book represent the diverse range of topics and themes that will be explored during the conference. From data interoperability and metadata standards to data-driven research and innovative applications, these abstracts demonstrate the breadth and depth of research in the field of metadata management.

We hope that this Book of Abstracts will serve as a valuable resource for the conference attendees, providing a concise overview of the exciting contributions including keynotes, talks, posters, demos and workshops that will be presented at HMC Conference 2025. Hopefully the various contributions will inspire further collaboration and knowledge-sharing among the metadata community, and contribute to the advancement of metadata management and collaboration in the years to come.

Conference Programme

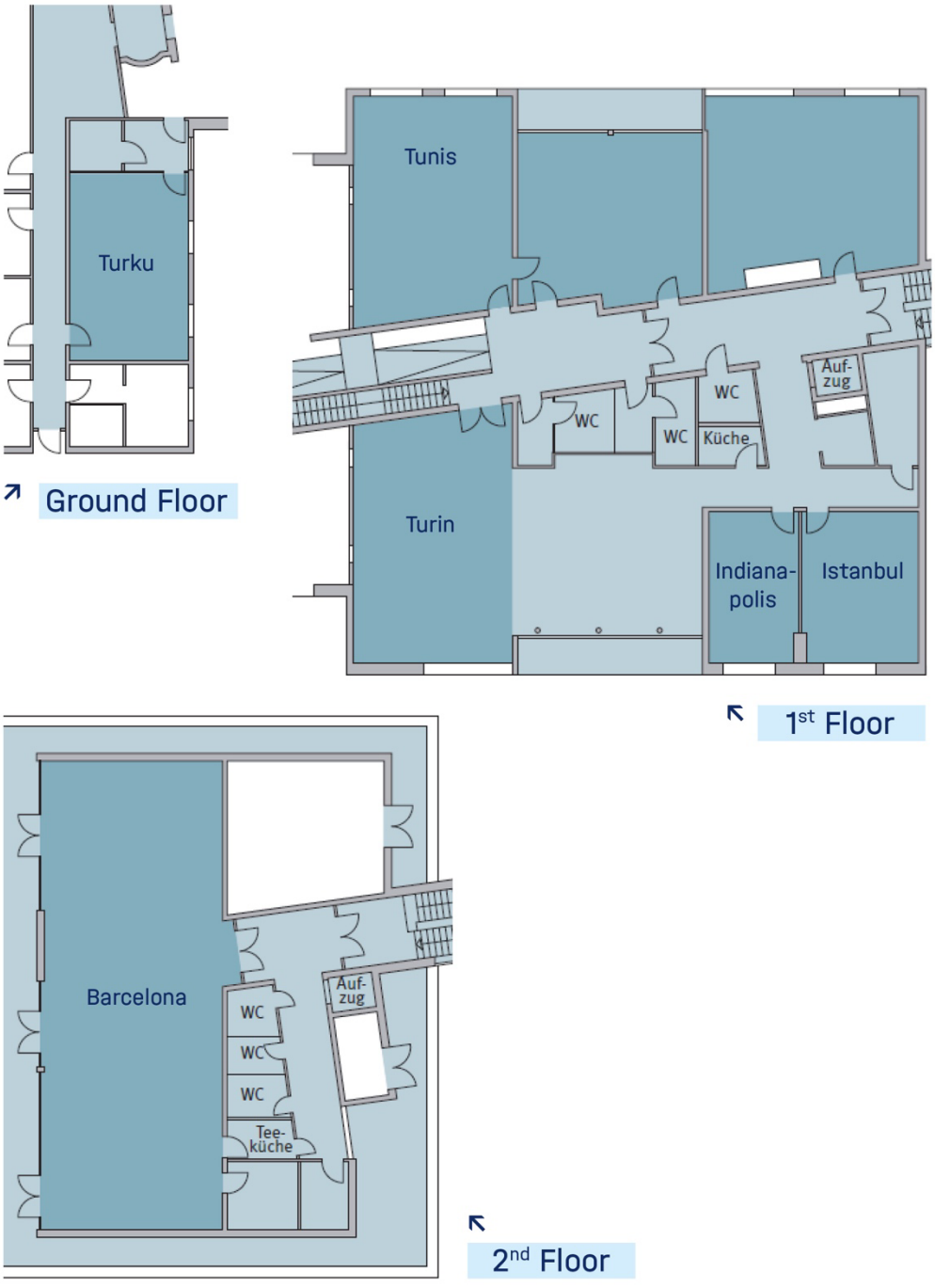
Overview of the HMC Conference Programme 2025

Here you can find our conference programme:
https://hmc2025.welcome-manager.de/front/content.php?id_article=581

Conference Location

Room Plan of the HMC Conference Location

Below you can find a map with all rooms of the HMC Conference 2025:



Monday, 12.05.2025

Room Barcelona, 14:30–15:30

Keynote

What is FAIR for? Some thoughts on progress, directions and priorities

Simon Hodson

[1] The Committee on Data of the International Science Council (CODATA), France

The FAIR principles have had a significant impact, and progress has been made towards their realisation. Nevertheless, it is important to reflect on the purpose of the FAIR principles: they are after all, a means to an end, rather than an end in themselves. This presentation will highlight the most important objectives of the FAIR principles and discuss how the scientific and data stewardship communities might prioritise activities to address the I and the R of FAIR.

Room Barcelona, 16:00–17:30

Session: Metadata Annotation and Management I

ID TO1

Standardizing Metadata for Electronic Laboratory Notebook objects: A Call for Community Collaboration

Author: Rory Macneil[1]
Co-author: Tilo Mathes[1]

[1] Research Space, Edingburgh, UK

Electronic Laboratory Notebooks (ELNs) have become essential and often institutionally required tools for modern scientific research, generating large amounts of digital objects including documents, notebooks, folders, projects, etc. These objects are crucial for recording the research process and provide unique context for research results and downstream research data. However, the lack of standardized metadata schemas for these digital artifacts limits their discoverability, accessibility, interoperability, and reusability. While physical samples benefit from established schemas like IGSN, ELN objects lack equivalent standardization. This presentation introduces our ongoing work with DataCite to use and potentially extend their metadata schema to effectively describe ELN digital objects, exemplified by RSpace. We will demonstrate how our current approach maps ELN-specific attributes to the DataCite metadata schema, addressing unique challenges such as versioning, permissions, and relationships between digital lab objects. We will present preliminary mappings between RSpace ELN objects and DataCite schema elements, highlighting gaps where extensions or modifications may be necessary. Critically, we recognize that establishing widely applicable standards requires perspectives from the wider research community. Therefore, we invite collaboration from ELN users, developers, and metadata specialists to contribute to refining these mappings.

Corresponding Author: Rory Macneil, rmacneil@researchspace.com

ID T02

Metadata Management of scientific instruments with inst.dlr

Author: Sac Medina[1]
Co-author: Federico Diaz Capriles[1]

[1] German Aerospace Centre (DLR)

Persistent Identifiers for Instruments (PIDINST) play a pivotal role in promoting FAIR (Findable, Accessible, Interoperable, and Reusable) principles by enabling the identification and tracking of research instruments across the data lifecycle. This work presents a comprehensive analysis of the PIDINST schema applicability, focusing on its capacity to capture and manage critical metadata elements for scientific instruments. We explore the applicability of the PIDINST metadata schema to three experimental facilities from different research areas of the German Aerospace Center (DLR). Our goal is to evaluate the degree of applicability of the schema to diverse use cases, highlighting its strengths and identifying areas for improvement. The methodology consisted of conducting a survey in which each facility was instructed to describe their instruments using the PIDINST schema. The schema properties were subsequently analysed, and their usage was quantified. Results indicate that the PIDINST schema is approximately 70% applicable to our three use cases. The remaining 30% may revealed limitations in our methodology or in the schema, which needs further analysis. To support the implementation of the PIDINST schema, we developed a software (Python code) tailored to implement the PIDINST schema efficiently. The software aims to streamline metadata annotation, ensure compliance with existing standards, and enhance interoperability between diverse data management systems. A key feature includes user guidance to ensure consistent metadata integration. Our presentation will highlight the methodological approach adopted for schema analysis, the technical architecture of the software, and key use cases demonstrating its utility in real-world research environments. Due to the connection between theoretical frameworks and practical application, our work contributes to the ongoing discourse on advancing metadata annotation and management in the research community

Corresponding Author: Sac Medina, sac.medina@dlr.de

ID T03

HELPMI and NAPMIX: Two Metadata-Initiatives in the field of matter

Author: Hans-Peter Schlenvoigt[1]
Co-authors: Andrew Mistry[1], Oliver Knodel[2]

[1] Helmholtz Centre for Heavy Ion Research (GSI), [2] Helmholtz-Zentrum Dresden-Rossendorf (HZDR)

Metadata is crucial for ensuring the FAIR principles (Findability, Accessibility, Interoperability, and Reusability) in scientific research, particularly in matter sciences, where diverse experiments and models generate vast heterogeneous data. To address these challenges, multiple metadata initiatives have emerged, enhancing data management and collaboration. This talk presents two key initiatives: HELPMI (Helmholtz Metadata Collaboration Initiative for Plasma and Laser-Plasma Physics) and NAPMIX (Nuclear, Astro and Particle Metadata Integration Experiments). HELPMI started the development of a domain-specific metadata framework for laser-plasma physics, aligning with the Helmholtz Metadata Collaboration (HMC) and FAIR principles, and building upon existing standards like NeXuS and openPMD. It focuses on easy-to-use metadata schemas for experiments at user facilities, and interoperability to facilitate data reuse, sharing, and validation. NAPMIX, part of the OSCARS project, facilitates metadata integration across nuclear, astroparticle, and particle physics, enhancing standardization and interoperability. It aims to support cross-domain collaboration through a common metadata framework while also providing a platform for end-user metadata generation. The development is carried out in collaboration with international European infrastructures. This presentation highlights the synergies between HELPMI and NAPMIX, comparing their metadata structuring, governance, and implementation approaches. By addressing shared challenges in plasma and nuclear/particle physics, these initiatives contribute to a broader, interoperable metadata ecosystem. We will also explore future directions, such as cross-disciplinary metadata integration and connections with emerging research infrastructures. Overall, this talk underscores the importance of metadata standardization in fostering open, reproducible science.

Corresponding Author: Hans-Peter Schlenvoigtr, h.schlenvoigt@hzdr.de

ID T04

Results of the 1st Interdisciplinary NFDI Metadata Workshop, January 2025: Can we agree on cross-disciplinary metadata standards to improve the reuse of research data?

Author: Oliver Koepler[1]

[1] TIB – Leibniz Information Centre for Science and Technology and University Library

The German National Research Data Infrastructure (NFDI) unites activities across all scientific disciplines to realise an overarching research data management. The interdisciplinary discussion and agreement on common standards for the description of research data, be it through metadata or terminology, is a key element for the successful implementation of FAIR data. The Task Force Metadata of the NFDI section "(Meta)data, Terminologies, Provenance" aims to moderate and coordinate the process of identifying and evaluating existing cross-disciplinary terminologies and metadata schemas as suitable candidates for NFDIagreed recommendations. The Task Force takes its motivation from the NFDI goals and the strategy of 2025/26 to develop cross-disciplinary metadata standards for the comprehensive (re-)usability of research data. With this vision in mind, the Task Force organised a two-day interactive NFDI Metadata Workshop at the Open Science Lab of TU Dresden to identify and evaluate suitable existing generic metadata schemas and recommendations on how to apply these in the disciplines represented by the NFDI. Altogether 46 participants from 26 NFDI consortia discussed common metadata schemas, disciplinary needs, and standardization approaches. Key sessions included expert talks from NFDI, re3data and EOSC, an overview of institutional processes for metadata standardization. In a preliminary landscape analysis the Task Force had identified DCAT, DataCite, and schema.org as three established and commonly applied metadata schemas to be discussed as potential candidates for NFDI-wide recommendations from the perspectives of the 26 NFDI communities. A World Café session facilitated discussions across disciplines about the utilization of the aforementioned metadata schema, , and the adoption of re3data for repository registration. The workshop also addressed questions about how repositories related to the NFDI can be represented in re3data in the future, including recommendation for registration and updating metadata of such repositories. In this presentation we will summarise the results of the workshop and give an outlook on how this dialogue can be continued towards an NFDI recommendation on a portfolio of cross-disciplinary metadata standards for research data.

Corresponding Author: Oliver Koepler, oliver.koepler@tib.eu

Tuesday, 13.05.2025

Room Barcelona, 09:00–10:00

Keynote

Data Management Makes Machine Learning Easier

Speaker: Till Korten[1]
Co-authors: Helene Hoffmann[1], Peter Steinbach[1], Özlem Özkan[2]

[1] Helmholtz-Zentrum Dresden-Rossendorf (HZDR), [2] Helmholtz-Zentrum Berlin (HZB)

Data management benefits the collaboration between Helmholtz AI consultants and researchers. We give practical examples how each letter in FAIR (data) makes it easier for machine learning experts to work with the data and therefore improves the quality of the outcome. This can serve as a general motivation for researchers to improve their data management.

Room Barcelona, 10:00-11:00

Session: Mapping & Community Initiatives

ID T05

Datathons--promoting equitability in sequence data reuse

Author: Stephanie Jurburg[1]
Co-author: Clara Arboleda[2]

[1] Helmholtz Centre for Environmental Research (UFZ), [2] German Centre for Integrative Biodiversity Research (iDiv)

Approaches to rapidly collecting global biodiversity data are increasingly important, but biodiversity blind spots persist. Here, I present the Datathon project, a community-driven initiative to stimulate the archiving and reuse of sequence data in biodiversity blindspots. Now starting its fourth yearly iteration, the Datathon project has consolidated over 6000 microbiome data points from biodiversity blindspots, and led to a network of >500 microbial ecologists who continue to learn and collaborate on sequence data reuse.

Corresponding Author: Stephanie Jurburg, stephanie.jurburg@ufz.de

ID T06

Development of Metadata Standards for 3D Seismic and Active Source Ocean Bottom Seismometer Data

Author: Janine Berndt[1]
Co-authors: Daniel Damaske[2], Mehrdad Soleimani Monfared[1], Estella Weigelt[3], Mechita Schmidt-Aursch[3], Hela Mehrtens[1], Christian Berndt[1], Janine Felden[3]

[1] GEOMAR Helmholtz Centre for Ocean Research Kiel, [2] MARUM – Center for Marine Environmental Sciences at the University of Bremen, [3] Alfred-Wegener Institute (AWI)

The reuse of marine seismic data requires standardised metadata. MetaSeis is the follow-up of a successfully completed similar project for two-dimensional multichannel seismic reflection raw data, conducted in collaboration with the DAM (German Marine Research Alliance) Underway Research Data initiative within the NFDI4Earth framework. The goal of MetaSeis is to develop and test metadata standards for three-dimensional seismic data and ocean bottom seismometer recordings of controlled seismic events. These events typically involve shots using airguns, as opposed to the passive recordings of earthquake seismicity that were addressed in the former HMC project eFAIRs. The first phase of the MetaSeis project involves two main tasks. First, we have adapted the two-dimensional seismic metadata standard for use with three-dimensional data, incorporating the added complexities of more extensive navigation data. This revised standard was initially tested during the R/V Maria S. Merian voyage MSM132, refined, and then reapplied during R/V Sonne voyage SO310. Further modifications are still required to accommodate the more complex data acquisition processes used by the Bundesanstalt für Geowissenschaften und Rohstoffe. Additionally, we are examining the latest industry standards to investigate if they are suitable for the scientific community. Second, MetaSeis assessed the status of legacy ocean bottom seismometer data at Alfred Wegener Institute and developed a workflow for archiving these data using a future metadata standard. In parallel, a separate standard is being developed for newly acquired ocean bottom seismometer data. The project aims to streamline the archiving of new three-dimensional reflection seismic and active-source ocean bottom seismometer data and to make these datasets more accessible for big data applications.

Corresponding Author: Janine Berndt, berndt@geomar.de

ID T07

OSCARS Composability Work: Beamline Finder at DESY based on the PaNET ontology

Author: Melanie Nentwich[1]
Co-authors: A. Paul Millar[1], Patrick Fuhrmann[1]

[1] Deutsches Elektronen-Synchrotron (DESY)

New users at a synchrotron facility like PETRA III are often overwhelmed by the multitude of possible experimental techniques (ETs) and their many variations with respect to energy ranges as well as sample and sample environment characteristics (e.g. crystalline and liquid samples; different in-situ setups). In order to help inexperienced users find an optimal beamline for their scientific questions, we are working on composing the capabilities of the PaNET ontology (Photon and Neutron ETs) with the web framework of the overview platform Way For Light (WFL). While WFL already offers a wealth of information for the more experienced users to explore the light sources in Europe, it currently does not cover the sample environment information and also only has a very rough categorization of the ETs. In contrast, the PaNET ontology lists more than 300 techniques that are (partially) interrelated and also reveal the purpose of the technique (e.g. creation of a 3D spatial map). By creating a web page that asks the right, intuitive questions, we aim to guide new users to find the perfect beamline for their experiments. Currently, we are creating a proof-of-concept that uses a functional programming language with a reactive paradigm (Elm) to query an underlying SPARQL endpoint. This endpoint includes the PaNET ontology as well as a customized database on the PETRA III beamlines. Using the customized database instead of WFL for now allows us to also include basic information on the available sample environments and to explore different approaches in handling dependencies between various experimental setups. The created web page provides users with an interactive way to navigate through the different options. This serves as a prototype for future work, a demonstration of the feasibility of the approach, and an opportunity to receive feedback from experts and users.

Corresponding Author: Melanie Nentwich, melanie.nentwich@desy.de

ID T08

Enhancing interoperability and integration: facilitating metadata transfer between research platforms and data repositories

Author: Marleen Marynissen[1]
Co-author: Dieuwertje Bloemen[1]

[1] KU Leuven, Belgium

To support researchers in sharing their data effectively, KU Leuven's institutional research data repository (RDR), based on Dataverse and launched in 2022, has focused on improving usability and ease of data and metadata entry. A key aspect of this is ensuring seamless integration with existing research tools and facilitating metadata transfer to enhance interoperability and integration. This enhances both metadata quality and the overall FAIRness of the data. Achieving this required close collaboration between (meta)data experts and technical specialists to design effective and scalable solutions for data and metadata transfer. In this presentation, we will discuss our collaborative approach to facilitating metadata transfer between active data management systems and data repositories, the challenges encountered in ensuring interoperability across diverse platforms, the impact on research workflows, and our future plans. In May 2023, we introduced an open-source Dataverse integration dashboard that enables researchers to transfer data easily from active data management systems such as GitHub, GitLab, and OSF, to the repository. This dashboard allows to efficient move files while maintaining the original data structure. Moreover, it ensures efficient file updates, prevents unnecessary duplication, and allows versioning at the file level by replacing only modified files using checksums. Building on this foundation of file transfers, our latest development extends the dashboard's functionality to include metadata import from these platforms as well. By streamlining metadata extraction and mapping, we eliminate manual entry, significantly reducing errors and inconsistencies. Despite challenges in mapping metadata, due to the slight differences in semantics behind similar metadata elements, and system-specific metadata quirks, the work proved invaluable as this functionality ensures that crucial contextual information is retained during data transfers. As a result, it improves data reliability, traceability, and compliance with the FAIR principles, while optimizing research workflows. Researchers benefit from a more efficient deposit process that minimizes administrative overhead while maintaining essential contextual information. By sharing our experiences, challenges, and lessons learned, we aim to provide valuable insights for other institutions looking to optimize their FAIR-aligned data infrastructures and enhance their metadata management practices.

Corresponding Author: Marleen Marynissen, marleen.marynissen@kuleuven.be

Room Barcelona, 11:30-12:30

Session: Metadata Annotation and Management II

ID T09

MeSyTo: Advancing Data Interoperability in Toxicology and Pharmacology Through Standardized Metadata and Ontologies

Author: Marina Pozhidaeva[1]
Co-authors: Kristin Schubert[1], Sebastian Canzler[1], Stephan Schreiber[1], Wibke Busch[1], Jörg Hackermüller[1]

[1] Helmholtz Centre for Environmental Research (UFZ)

Tackling chemical pollution as part of the Triple Planetary Crisis requires data science-based solutions that enable comprehensive environmental monitoring, risk assessment, and policy development. However, the lack of standardized and interoperable metadata in toxicology and pharmacology hinders the effective integration and reuse of critical data. The MeSyTo project aims to improve toxicological and pharmacological data interoperability and machine readability by developing standardised metadata frameworks. The project addresses inconsistencies in metadata annotations that hinder the integration and reuse of data in toxicology and pharmacology. By adopting the FAIR (Findable, Accessible, Interoperable, Reusable) principles, MeSyTo will create structured metadata to document exposure conditions, experimental settings and workflows for omics data from collection to storage in the data repositories. Our key accomplishments include the development of a comprehensive metadata catalog through the integration and harmonization of metadata elements from the OECD Omics Reporting Framework, various omics data repositories, and wet lab protocols. Furthermore, we have established a domain-specific ontology by defining core classes and aligning them with external ontologies to ensure semantic consistency and interoperability across datasets. Current efforts focus on validating the domain-specific ontology, developing constraints using SHACL for case-specific metadata annotation, and extending its application to multi-omics data integration. The overarching goal is to enhance reproducibility, data sharing and reuse, and to foster collaboration within Helmholtz and beyond, ultimately empowering researchers to tackle the grand challenges of chemical pollution and its impact on human and ecosystem health.

Corresponding Author: Marina Pozhidaeva, marina.pozhidaeva@ufz.de

ID T10

Kadi4Mat: Enhancing Metadata Management in Surface Science

Author: Johannes Steinhülb[1]
Co-author: Sven Berger[2], Darya Snihirova[2], Daniel Höche[2], Michael Selzer[1]

[1] Karlsruhe Institute of Technology (KIT), [2] Helmholtz-Zentrum Hereon (HEREON)

Surface science encompasses diverse applications from corrosion analysis to battery technology and advanced materials design that generate diverse datasets through both experimental measurements and simulations [1,2,3]. Ensuring data are well-described, traceable, and reproducible is increasingly critical, given the diversity of file formats and metadata standards in this field. To address these challenges, we introduce Kadi, an advanced research data management (RDM) platform tailored for surface science. Kadi allows to integrate flexible electronic lab notebooks (ELN) [4] with KadiStudio [5] an intuitive workflow management tool to streamline the design, execution, and documentation of complex processes. The platform facilitates the automatic capture of metadata from lab devices and standardizes data storage for both experimental and simulation outputs. Its graphical interface allows researchers to build workflows that incorporate subworkflows, loop conditions, and execution directly from the web frontend. This integrated approach enhances data interoperability and reduces administrative overhead in environments with strict IT regulations. Our presentation will detail examples demonstrating how Kadi improves data traceability and reproducibility. To illustrate Kadi's transformative impact, we present a case study on aluminum's interfacial behavior in a corrosive NaCl environment—capturing both its electrochemical response and detailed morphological evolution. This example underscores how systematic data recording and streamlined workflow management drive efficient, transparent research practices that ultimately accelerate scientific discovery in surface science.

References:

- [1] Neher, N. et al. (2025). TrixiParticles.jl: Particle-based multiphysics simulation in Julia. Journal of Open Source Software.
- [2] Deng, M. et al. (2018). Mg-Ca binary alloys as anodes for primary Mg-air batteries. Journal of Power Sources.
- [3] Gazenbiller, E. et al. (2024). Mechanistic insights into chemical corrosion of AA1050 in ethanol-blended fuels with water contamination via phase field modeling. Materials and Corrosion.
- [4] Schlabach, S. et al. (2024). Using ELN functionality of Kadi4Mat (KadiWeb) in a materials science case study of a user facility. Data Science Journal, 23(1).
- [5] Al-Salman, R. et al. (2023). KadiStudio use-case workflow: Automation of data-processing for in situ micropillar compression tests. Data Science Journal, 22(1)

Corresponding Author: Johannes Steinhülb, johannes.steinhuelb@kit.edu

ID T11

Metadata for Ionospheric and Space Weather Observations (MISO)

Author: Xingzhi Lyu[1]
Co-authors: Yuri Shprits[1], Dedong Wang[1], Miriam Sinnhuber[2], Jens Berdermann[3]

[1] German Research Centre for Geosciences (GFZ), [2] Karlsruhe Institute of Technology (KIT),
[3] German Aerospace Centre (DLR)

The MISO project aims to overcome existing barriers to fast and efficient data exchange among Helmholtz and international research groups working in space physics and space weather. Its primary goal is to enable seamless integration and interoperability of data across diverse scientific domains. To this end, we have analyzed the existing domain-specific data products and formats, identified key metadata requirements, and proposed discipline-specific variables to enhance metadata content in support of a unified metadata standard. Special attention is given to intersections of data usage among GFZ, DLR, and KIT. Although all three institutions share a focus on space weather, they operate in different regions of near-Earth space, such as the magnetosphere, atmosphere, and ionosphere, which involve distinct parameters, coordinate systems, and units. These differences present significant challenges for metadata harmonization. Establishing a flexible and standardized metadata framework is therefore essential to ensure data findability, accessibility, and interoperability within the Helmholtz Association and beyond.

Corresponding Author: Xingzhi Lyu, xlyu@gfz.de

ID T12

Project MEMAS: a framework for FAIR data storage in composite engineering

Author: Pradnil Kamble[1]
Co-authors: Mathieu Vinot[1], Nicolas Unger[1], Roland Glück[1], Nathalie Toso[1]

[1] German Aerospace Centre (DLR)

The management of data abroad various disciplines can be a very challenging task due to the variety of expert language and the heterogeneity of data and data formats. In this contribution, we present outcomes and technical solutions developed in the project MEMAS for the efficient and sustainable storage of data and metadata in robotics, manufacturing, testing and simulation of composite parts. Our work follows the FAIR principles by developing a multi-domain ontology that bridges the abovementioned fields of engineering. This integration presents unique challenges, as robots are not conventionally used in this manner, demanding new conceptual approaches in ontology design. Data interoperability and reusability is ensured by the use of the research data management system (RDMS) shepard to store heterogenous research data. The development of json schemas allowed for the automatic generation of user interfaces, which are used to enrich data sets and store reusable instances of ontology classes, for instance for testing instruments, machines or test standards. In a second phase, we focused our work on the generation of automatic parsing tools to extract metadata from research files. Structuring tools allowed for the conversion of human-readable files into machine readable data objects, in particular json or timeseries, for structured storage in RDMS. The heterogeneous data and metadata stored within the RDMS are systematically structured, ensuring findability and semantic coherence. This developed approach and methods establishes a robust foundation for future advancements such as multi-objective optimization and machine learning-driven insights.

Corresponding Author: Pradnil Kamble, pradnil.kamble@dlr.de

Room Barcelona, 14:30-15:30

Session: Technical Solutions, Semantics and Data Fabric

ID T13

Discovery and Access of data in EOC Geoservice using STAC

Author: Jan-Karl Haug[1]

[1] German Aerospace Centre (DLR)

In order to make the EOC Geoservice data accessible to a wider public, we offer a STAC-based catalog service in addition to the established download and visualization services. This allows data to be found and accessed dynamically and efficiently. Users can access the data without having to download the complete data set, thus avoiding longer transmission times and saving storage capacity. STAC is divided into several specifications and is structured hierarchically. The STAC item is the core atomic unit, representing a single spatiotemporal asset with the associated metadata and forms the basis for STAC. The STAC API provides a RESTful endpoint that enables the search for STAC items. The STAC Catalog provides a structure for organizing and searching STAC items. The STAC collection contains additional information, such as spatialtemporal extent, license, keywords, or providers, of the STAC items that fall into the collection. To retrieve the available collections and items, the STAC API endpoint needs to be approached either via a STAC browser or within arbitrary code, e.g. in a Jupyter notebook. In such a notebook, a query can be started using various Python libraries (e.g. pystac) and data can be loaded into an xarray dataset (data cube). The data can then be visualized or further analyzed. The following presentation introduces the EOC EO Products Service and how you can efficiently access the catalog and its available content.

Corresponding Author: Jan-Karl Haug, jan-karl.haug@dlr.de

ID T14

Semantic x-Lab: How HELIPORT and ALAMEDA Teams are Joining Forces to Improve Knowledge Acquisition from Lab Resources

Author: Oliver Knodel[1]
Co-authors: David Pape[1], Martin Voigt[1], Manja Luzi-Helbing[2], Marc Hanisch[2], Felix Mühlbauer[2], Alexander Kessler[3], Andrew Kishor Mistry[4]

[1] Helmholtz-Zentrum Dresden-Rossendorf (HZDR), [2] German Research Centre for Geosciences (GFZ), [3] Helmholtz Institute Jena (HI JENA), [4] Helmholtz Centre for Heavy Ion Research (GSI)

In modern scientific research, the efficient management and integration of data from various laboratory resources is of crucial importance. The collaborative HMC-funded project 'Semantic x-Lab' is an initiative with the aim to improve knowledge acquisition by combining the advanced platform HELIPORT (HELmholtz SciIentific Project WORkflow PlaTform) and the results of the ALAMEDA (A Scalable Multi-Domain Metadata Management Platform) project. The goal of the Semantic x-Lab project is to interlink information from HELIPORT and ALAMEDA, make it explorable, and even discover knowledge that was previously considered in a different context or research field, in order to generate new insights. The outcome of the project will be a distributed knowledge graph involving laboratory resources and large-scale facilities in different research domains. We work with our laboratory partners in a user-centered co-design process. Use cases are defined collaboratively and feedback on the use of the knowledge graph is continuously incorporated into the development process. Semantic x-Lab exemplifies how the outcomes of previous successful HMC projects can be reused to launch new collective efforts. This presentation will explore the methods used to bring the platforms together, the anticipated challenges in doing so, and the expected benefits for the scientific community. Semantic x-Lab can serve as a model for integrating different data management systems to promote more effective and FAIR-compliant research practices and to discover new insights into previously unknown relationships. HELIPORT is a data management solution designed to provide a FAIR overview of all components and steps of the research experiment. It integrates documentation, workflows, related publications, proposal management, electronic lab notebooks, software repositories and other data sources, and facilitates the automatic exchange of relevant metadata throughout the entire lifecycle of the experiment. ALAMEDA builds upon HELIPORT and extends it with a focus on use cases from the earth and environment research field. It covers five main categories of metadata: observations and measurements, samples and data history, sensors and devices, methods and processing, and environmental properties. The project is carried out by a multi-disciplinary team. We envisage a broad usage in different research fields and the generation of cross-domain insights.

Corresponding Author: Oliver Knodel, o.knodel@hzdr.de

ID T15

LabFriend: improving (meta)data entry in ELNs with semantic support and speech recognition

Author: Marta Dembska[1]

[1] German Aerospace Centre (DLR)

The digitalisation of laboratory environments has made Electronic Laboratory Notebooks (ELNs) essential for recording and managing experimental data. However, manual data entry remains a bottleneck, often leading to incomplete or inconsistent (meta)data due to user resistance and the tedious nature of form-filling. To address these challenges, we are launching LabFriend, an intelligent (meta)data input assistant designed to enhance ELNs by integrating semantic technologies. LabFriend will employ association rule mining to provide context-aware text entry suggestions, ensuring terminology consistency and reducing input errors. Additionally, its speech recognition functionality aims to enable hands-free (meta)data capture, improving accessibility and usability. By integrating with knowledge graphs and ontologies, LabFriend is expected to support FAIR (Findable, Accessible, Interoperable, and Reusable) data practices, enhancing both (meta)data quality and quantity. The system is planned as a stand-alone tool with seamless integration into multiple ELNs, facilitating interoperable and structured data management. As the project is at its early stages, this presentation will outline the key challenges in (meta)data entry, discuss our proposed solutions, and invite feedback on potential implementation strategies.

Corresponding Author: Marta Dembska, marta.dembska@dlr.de

ID T16

Managing, searching and annotating research and production data in shepard

Author: Roland Glück[1]
Co-authors: Patrick Kaufmann[1], Felix Lettowksy[1], Jessica Friedline[2], Wiglef Rehm[2]

[1] German Aerospace Centre (DLR), [2] Xitaso GmbH, Augsburg

The system shepard (storage for heterogenous production and research data) was developed at the Center for Lightweight Production Technology of the German Aerospace Center in Augsburg. Its purpose is to offer structured storage of data created during production and research experiments. Particular focus is attributed towards search functionalities and the possibility of annotating data with the help of ontologies. Shepard consists of a backend running on a server for the infrastructure and a web-based frontend for visualizing the data structure and the data itself. The backend offers its functionality via a REST API; the frontend makes calls to this API and passes the results to the user in a structured visual way. In the implementation, a graph database (here neo4j) is used to store the structure of the stored data and other organizational Elements like users and permissions. The actual payload data is stored, depending on its type, in a timeseries database (currently switching from influx to timescaleDB), a MongoDB for arbitrary files and custom defined data structures, and a spatial database (postgis) which is currently in an experimental phase of integration. Shepard is publicly available at https://gitlab.com/dlr-shepard.Data can be structured in collections with subordinated data objects containing references to the actual data payload or other data objects and collections. Usually, collections correspond to greater entities like entire projects whereas data objects model experiments or process steps and their substeps. Hierarchical and causal or temporal dependences can be modelled by predefined relations between data objects. Additionally, one has the possibility to create custom tailored relationships by means of so called references which represent Relations between data objects, collections and containers for the actual payload data. A system of permissions makes the stored data visible and editable for certain classes of users and thus contributes to accessibility and thanks to the search option also to findability of data. In order to make data interoperable, semantic annotations referring to ontologies can be attached to the data, thus opening the door for the application of reasoning with ontologies and further AI methods. This gives the whole system the flavor of a data lakehouse.

Corresponding Author: Roland Glück, roland.glueck@dlr.de

Room Barcelona, 17:00-18:15

Session: Infrastructure and Common Practices

ID T17

Registry2RDF: Bridging the Gap in Sensor Metadata Integration

Author: Mihir Rambhia[1]
Co-authors: Claas Faber[2], Fabian Kirchner[1], Linda Baldewein[1], Carsten Schirnick[2], Smruthishree Srinivasa[1], Catriona Eschke[1]

[1] Helmholtz-Zentrum Hereon, [2] GEOMAR Helmholtz Centre for Ocean Research Kiel

Public registries, such as the O2A Registry, Sensor Management System (SMS), Persistent IDentification of INSTruments (PIDINST), and others, are increasingly used by the scientific community to store and provide sensor metadata in a FAIR manner. Applications developed to support sensor management operations need to aggregate sensors registered in different registries. However, each registry has a unique API and outputs data in different formats, making cross-registry integration challenging. Currently, there is no standardized approach for aggregating and integrating metadata from multiple registries in applications that manage sensors across different platforms. As part of the MOIN4Herbie project under HMC, Herbie — a digital lab notebook used for digitizing the provenance of scientific data — is being further extended to digitize the storage of sensor maintenance metadata. To achieve this, ontologies were first developed for the sensor maintenance use case. These ontologies were then implemented using the Shapes Constraint Language (SHACL) to collect and validate metadata from users as Resource Description Framework (RDF) graphs. To facilitate metadata integration from registries into Herbie, the Registry2RDF module has been developed to transform registry data from the O2A Registry into RDF graphs. To enhance interoperability, this module will be further extended to support a unified SHACL shape, enabling data collection from multiple registries. It could serve as the missing standardized approach for integrating metadata into other applications. With this contribution, we aim to initiate a discussion on how to generalize our approach for broader applicability in other projects.

Corresponding Author: Mihir Rambhia, mihir.rambhia@hereon.de

ID T18

The Helmholtz Knowledge Graph: driving the transition towards a FAIR data ecosystem in the Helmholtz Association

Author: Volker Hofmann[1]
Co-authors: Mustafa Soylu[1], Gabriel Preuß[2], Fiona D'Mello[1], Said Fathalla[1]. Lucas Kulla[3], Stefan Sandfeld[1]

[1] Forschungszentrum Jülich (FZJ), [2] Helmholtz-Zentrum Berlin (HZB), [3] German Cancer Research Centre (DKFZ)

Research in the Helmholtz Association is carried out in inter- and multidisciplinary collaborations that span between its 18 independently operating research centres across Germany. However, research data and digital assets is heterogeneous in terms of formats, used schemas, record consistency and hosting location. This impairs convergence in a FAIR Helmholtz Data Space. The Helmholtz Metadata Collaboration (HMC) is taking on this challenge by developing the Helmholtz Knowledge Graph (Helmholtz KG) [1] as a lightweight interoperability layer and semantic harmonisation target, that connects Metadata Helmholtz digital assets, which are stored in a decentralised manner. With the KG, we envision (1) providing better cross organisational access to Helmholtz's (meta)data and information assets on an upper semantic level, (2) harmonising and optimising the related metadata across the association, and (3) forming a basis from which the semantic quality and the depths of metadata descriptions is improved and extended into domain and application levels. With recent works [2] we have improved the maturity level of the software to be scaled as more and more data providers are connected and the Graph keeps growing. At the same time we are developing an internal data model with mappings that will allow to aggregate and harmonise data delivered in a from that is consistent with a defined set of semantic standards (Schema.org, DataCite, Dcat, DC). With our effort we want to establish active exchange over which common standards for Helmholtz data providers can emerge and support metadata harmonisation.

Links

- [1] Broeder, J.; Preuss, G.; D'Mello, F.; Fathalla, S.; Hofmann, V.; Sandfeld, S. (2024) The Helmholtz Knowledge Graph: driving the Transition towards a FAIR Data Ecosystem in the Helmholtz Association; The Semantic Web: ESWC 2024, Springer Computer Science Proceedings. doi:10.34734/FZJ-2024-03156
- [2] https://codebase.helmholtz.cloud/hmc/hmc-public/unhideAcknowledgements

This work was supported by (1) the Helmholtz Metadata Collaboration (HMC), an incubator-platform of the Helmholtz Association within the framework of the Information and Data Science strategic initiative

Corresponding Author: Volker Hofmann, v.hofmann@fz-juelich.de

ID T19

A comprehensive metadata descriptor for multimodal light measurements of natural indoor and outdoor scenes

Author: Niloufar Tabandeh[1]
Co-author: Manuel Spitschan[1]

[1] Technical University Munich (TUM)

Environmental light plays a crucial role in human health, with complex spectral, spatial, and temporal characteristics influencing biological and psychological processes. To date, no standardised metadata schema exists to capture the full range of multimodal light properties and relevant environmental information in natural settings. This lack of a structured framework presents a significant challenge for interdisciplinary research and limits data comparability and large-scale analysis. To address this need, we have developed a modular and hierarchical metadata descriptor comprising a total of 35 items designed to standardise the documentation of environmental light measurements. The schema is structured across multiple tiers: (1) general project-level metadata, including project ID, owner, and title (n= four items); (2) measurement-level metadata (n= 22 items); capturing details such as record ID, date and time, devices used, measurement protocols, and environmental conditions (e.g., weather, ambient lighting); and (3) device-specific metadata (n= nine items); incorporating detailed measurement parameters, for example, melanopic equivalent daylight illuminance (melanopic EDI) for spectral irradiance measurements. This metadata descriptor has been tested, refined and validated through a multimodal data collection campaign across different geographical locations and natural environments, yielding a total of 875 independent data points. Its structured approach enhances data interoperability, facilitates large-scale analyses, and supports the FAIR (Findable, Accessible, Interoperable, and Reusable) data principles. The schema is flexible and can be adapted for different measurement devices and research projects. This ensures consistent data collection and improves crossstudy comparisons. An implementation in the Frictionless platform is planned.

Corresponding Author: Niloufar Tabandeh, niloufar.tabandeh@tum.de

ID T20

Integrated research infrastructures: A context and platform for enabling widespread implementation of metadata

Author: Vaida Plankyte[1]

[1] Research Space, Edingburg, UK

Development of commonly accepted metadata standards and schemas is the fundamental underpinning of more effective and widespread adoption of metadata in research communities. However, this is just a starting point. Less attention has been given to how we can use this foundation to encourage broad adoption of metadata standards and schemas in research practices and workflows. This presentation discusses the role integrated research infrastructures can play in enabling broad adoption of metadata standards and schemas in research practices and workflows. It sets the scene with brief introductions to two important recent developments in thinking about research infrastructures. The first is the 'MaLDReTH' map of the landscape of digital research tools used in the research lifecycle, which was developed by an RDA working group. The second is the concept of Vertical Interoperability, defined as Interoperability enabling streamlined passage of data and metadata across research tools involved in every stage of the research lifecycle. The presentation continues with an exploration of how these two complementary constructs can provide a platform for supporting incorporation of commonly accepted metadata standards and schemas into actual research workflows in ways that facilitate widespread adoption across diverse research communities. It addresses design and implementation challenges and means of addressing them.

Corresponding Author: Vaida Plankyte, vaida@researchspace.com

ID T21

Why are there so many metadata schemas and what role does the PIDs play?

Author: Andrea Pörsch[1]
Co-authors: Emanuel Söding[2], Kirsten Elger[1], Dorothee Kottmeier3, Stanislav Malinovschii[2], Sören Lorenz[2]

[1] German Research Centre for Geosciences (GFZ),
[2] GEOMAR Helmholtz Centre for Ocean Research Kiel (GEOMAR), [3] Alfred-Wegener Institute (AWI)

Persistent identifiers (PIDs) are vital components in standardised metadata schemas for referencing and digitally linking an increasing number of entities in the research data landscape (persons, institutions, publications, data, samples, instruments, etc.). They facilitate the enhancement of metadata completeness and are pivotal metadata elements in the creation of knowledge graphs. A survey in the Earth and Environmental research field identified the use of three primary metadata schemas: DataCite, ISO 19115/19139, and Schema.org. Each schema has its unique context, purpose, advantages, and limitations: 1. The DataCite metadata schema is well-established and essential for publishing research data, scientific software, IGSNs and other objects with digital object identifiers (DOI). Its extensive use is granted by the small number of only six mandatory metadata properties that ensure the bibliographic information of an object. In addition, the DataCite Schema supports data discovery by a variety of recommended and optional metadata properties. Users of the DataCite schema are strongly encouraged to include persistent identifiers whenever possible (ORCID, ROR, IGSN, PIDINST, DOIs for citations of related articles, datasets and other sources). 2. The ISO 19115/19139 Schema is an international standard for the description of geographic information and services. The EU INSPIRE Directive mandates the use of ISO 19115, 19119 and 19139 to facilitate the exchange of governmental geospatial data and services, and the analysis of map data. It employs Universally Unique Identifiers (UUIDs) for the identification of resources. 3. Schema.org was developed as a collaborative initiative to provide standardised vocabulary for structuring web data. When embedded in websites, schema.org enables these websites to be found by major internet search engines, such as Google. Schema.org can be mapped from other metadata schemas (e.g., DataCite, ISO) in a flat hierarchy structure that enhances discoverability. This poster compares these three "worlds" and explores ways to bridge them, with a view to enhancing the discoverability and presentation of research results in search portals.

Corresponding Author: Andrea Pörsch, apoersch@gfz.de

Wednesday, 14.05.2025

Room Barcelona, 09:00-10:00

Session: Human Actors and FAIR Metrics

ID T22

Adrift in the DAS - how we can help researchers improve the F in FAIR

Author: Kirsty Merrett[1]
Co-author: Jade Godsall[1]

[1] University of Bristol, England

In 2024 the UK Reproducibility Network ran various pilots to investigate different indicators for measuring open research. The Open Research Indicators Pilot was sector led, with institutions and solution providers working together to develop, test, and evaluate prototype machine learning solutions with valid, reliable, and ethical indicators for Open Research. The University of Bristol was the lead for the 'openness of data' pilot and assessed providers' data to ascertain the usefulness of machine learning for this purpose. The pilot's findings highlight the inherent challenges and limitations of monitoring and assessing published data within a research landscape that prioritises articles as benchmark outputs; the combination of article primacy and existing publisher and repository systems means datasets can only realistically be monitored in Data Availability Statements (DAS), however, statements do not have consistent metadata, terminology, or templates, may be obscured to machine learning through nested HTML drop down menus and if not open access, metadata is frequently excluded from openly accessible metadata of publications. This lack of consistent language, terminology, and templates is compounded by different conventions in disciplines and incongruous data and article publishing workflows, which often results in researchers being left adrift, struggling to understand what is required by the journal's policy, whether data should be cited as a footnote or reference, where a DAS should be located, how it should be structured for findability, and even the demarcation between data as supplementary files and publishing as a dataset. Prototyping innovative machine learning tools confirmed an uncomfortable truth many in the RDM community suspected; we do not have consistent metadata for digital tools to reliably and accurately extract DAS, and we are not doing enough at the human interface with researchers to ensure FAIR data can be found by others. This talk describes how Bristol's Research Data Service bridge that gap; through advocacy, training, DAS generator tools, and set texts for citation for open, restricted and controlled datasets in the data.bris repository. Whilst RDM professionals, publishers, and funders must work together to agree on metadata and language to improve the reliability and ease with which researchers cite, curate, publish and preserve research data the researcher may be the most important link in the chain to drive the process

Corresponding Author: Kirsty Merrett, j.k.merrett@bristol.ac.uk

ID T23

Metadata and the Boss: What management can and needs to do

Author: K. Gerald van den Boogaart[1]
Co-authors: Theresa Schaller[1], Florian Rau[1]

[1] Helmholtz-Zentrum Dresden-Rossendorf (HZDR)

Going forward into the big data and AI age, FAIR metadata is of key importance for visibility, impact and correct use of our research. Standardized and usable metadata is however not something the individual scientists can generate alone without a functioning ecosystem at their research center and their science community. Good Metadata requires the establishment of structures, competences and strategic developments across the research centers, including provision of IT tools, integration across the various parts of the center including e.g. finance and project management and competence management across the organisation. The talk shows various roles and task relevant for metadata production from a science management and institutional perspective and discusses the responisbilities of management at all levels from group leader to center head in the process of developing organizational competences and structures fit for the FAIR and open science.

Corresponding Author: K. Gerald van den Boogaart, boogaart@hzdr.de

ID T24

Enhancing Metadata Quality through Persistent Identifiers: Insights from PID4NFDI and DataCite

Author: Sara El-Gebali[1]

[1] DataCite

Persistent Identifiers (PIDs) and their accompanying metadata play a crucial role in effective research data management and the realization of FAIR principles (Findable, Accessible, Interoperable, Reusable). The PID4NFDI coordination hub is dedicated to creating a sustainable, scalable, and community-driven framework for metadata interoperability across the NFDI landscape. This initiative seeks to harmonize metadata practices, enhance metadata quality and completeness, and ultimately improve the discoverability of research outputs. In this talk, we will share actionable insights derived from recent metadata assessments using metrics from DataCite's DOI metadata records, complemented by key findings from our survey on metadata practices within the NFDI community. These insights highlight critical factors influencing metadata quality and researcher engagement. Drawing from real-world examples, we demonstrate how comprehensive metadata, including bibliographic information and provenance details, supports the discoverability, reusability, and long-term sustainability of research outputs. Additionally, we outline upcoming initiatives within PID4NFDI to address identified metadata needs, emphasizing the critical role of PID service providers in improving metadata management practices. By highlighting strengths and pinpointing gaps in current metadata practices among NFDI consortia, we offer targeted recommendations that inform strategic decision making and enhance organizational metadata workflows. Our collaborative approach aims to reduce administrative burdens, raise awareness among researchers and infrastructure providers about the importance of high-quality metadata, and foster a unified metadata ecosystem, ultimately enhancing the value and impact of research data.

Corresponding Author: Sara El-Gebali, sara.elgebali@datacite.org

ID T25

DDI Adoption Metrics

Author: Knut Wenzig[1]

[1] German Institute for Economic Research, Research Data Centre of the Socio-Economic Panel (DIW Berlin/SOEP)

DDI metadata standards are widely used to describe especially tabular data in depth, including their columns/variables. According to re3data.org, a global registry of research data repositories, they are the most prevalent standards with these capabilities. Additionally, re3data.org identifies OAI-PMH as the most widely adopted protocol for harvesting such metadata, with many endpoints available at re3data.org. Building on this, Wenzig and Han (2024, https://doi.org/10.29173/iq1116) recently analyzed over 250,000 data records in the DDI Codebook format from various sources. Their methods can be adapted for continuous monitoring of metadata usage and availability. Such an approach could help new users identify where and how other institutions implement DDI standards, offering concrete starting points for further research and community engagement. A project funded by KonsortSWD – NFDI4Society is advancing these efforts with the following key deliverables:- Development of metrics on DDI standard adoption, including the number of institutions using DDI, number of institutions providing DDI metadata through standardized methods, and statistics about the availability of records in the different DDI standards (DDI-Codebook, DDI-Lifecycle, DDI-CDI).- Publication of datasets containing raw data for these metrics at each measurement interval.- An experimental dashboard displaying these metrics alongside links to individual metadata records and repositories. This talk will provide an update on the project's progress, which serves two immediate goals: 1. Visualizing the adoption of DDI metadata standards. 2. Encouraging institutions that have not yet made their metadata accessible to do so, enhancing their visibility within the community.

Corresponding Author: Knut Wenzig, kwenzig@diw.de

Room Barcelona, 10:30-11:30

Session: Metadata Annotation and Management III

ID T26

Software CaRD: a curation and reporting dashboard for compliant FAIR software publications

Author: Oliver Bertuch[1]
Co-authors: David Pape[2], Sophie Kernchen[3], Christian Meeßen[4]

[1] Forschungszentrum Jülich (FZJ), [2] Helmholtz-Zentrum Dresden-Rossendorf (HZDR),
[3] German Aerospace Centre (DLR), [4] German Research Centre for Geosciences (GFZ)

As baseline for a satisfaction of the FAIR4RS principles, research software must be published with metadata in publication repositories that assign persistent identifiers and make the metadata accessible. Additionally, published software metadata must be correct, and rich enough to further improve findability, accessibility, interoperability and reusability. Metadata curation for software publication not only safeguards the respective metadata quality, but also assesses compliance with relevant policies in the Helmholtz Association, its centers, and beyond. Furthermore, software metadata can cumulatively be enriched with dynamic metadata (e.g., usage, citations, development) and can thus be used for evaluation and academic reporting, e.g., to contribute to software-related indicators currently developed within the Helmholtz Association. While software publication can be automated, metadata curation, publication approval and evaluation processes usually require human involvement and should be supported by user interfaces that build on automation tools. In our project, we will create "Software CaRD" (Software Curation and Reporting Dashboard), an application that presents software publication metadata for curation. Preprocessed metadata from automated pipelines are made accessible in a structured graphical view. Issues and conflicts are highlighted to allow for easy resolution. Software CaRD also assesses metadata for compliance with configurable policies. For evaluation and reporting, relevant metadata from applicable sources is tracked and visualized.

Corresponding Author: Oliver Bertuch, o.bertuch@fz-juelich.de

ID T27

From FAIR WISH to FAIR AIMS – bringing physical samples to the digital world

Author: Kirsten Elger[1]
Co-author: Alexander Brauser[1], Linda Baldewein[2], Simone Frenzel[1], Birgit Heim[3], Rolf Krahl[4], Ulrike Kleeberg[2], Ben Norden[1], Mareike Wieczorek[3]

[1] Helmholtz Centre for Geosciences (GFZ), [2] Helmholtz Zentrum Hereon (HEREON), [3] Alfred-Wegener-Institute (AWI), [4] Helmholtz-Zentrum Berlin (HZB)

FAIR WISH - FAIR Workflows to establish IGSN for Samples in the Helmholtz Association, Helmholtz Centers AWI, GFZ and Hereon (2022-2023), was the most "physical" of all HMC projects. The International Generic Sample Number (IGSN) is a globally unique and persistent identifier (PID) for physical objects. IGSNs facilitate the digital connection between scientific publications, research data and the samples from which the data was obtained, thereby addressing one of the final gaps in the full provenance of research results. The objective of FAIR WISH was to promote the wider use of IGSNs in the geosciences, with thousands of samples that are often moved between laboratories and institutions. Major project outcomes were: (1) the development of standardised and discipline-specific IGSN metadata profiles for different geo-bio sample types; (2) the full documentation of the GFZ IGSN Metadata Schema; (3) the "FAIR SAMPLES Template"; and (4) SAMIRA – the FAIR SAMPLES Template Processing Software. The FAIR SAMPLES Template, developed for the geosciences, is a modular tool that allows users to select sample type-specific metadata properties and provide standardised descriptions. It contains a set of linked data vocabularies and forms the basis for the semi-automatic generation of the IGSN metadata XMLs, the batch upload to the DataCite API for IGSN registration (both with SAMIRA), and the source for the IGSN landing pages. It offers researchers an easy and flexible approach to describe their samples (and subsamples) with comprehensive IGSN metadata, regardless of the level of digitalisation of sample metadata and the individual researcher's metadata training. The new HMC project FAIR AIMS – Automated IGSN Management System will build on these achievements and develop an online version of the FAIR SAMPLES template with automated workflows for IGSN registration, the integration of linked data vocabularies as drop-down lists, and automatic metadata quality checks during metadata upload. The potential integration of these tools into HZBs sample management system SEPIA, with specific metadata profiles for materials science will be investigated. This presentation highlights the results of the FAIR WISH project and discusses the benefits and limitations of the current FAIR SAMPLES template. FAIR AIMS will address these challenges by making the Template fully online and further automating the IGSN registration workflows.

Corresponding Author: Kirsten Elger, kelger@gfz.de

ID T28

Organizing Open Data for DESY, HIFIS, NFDI and EOSC

Author: Tim Wetzel[1]
Co-authors: Patrick Fuhrmann[1], Armando Bermudez Matinez[1], Johannes Reppin[1], Regina Hinzmann[1]

[1] Deutsches Elektronen-Synchrotron (DESY)

DESY is currently expanding on their Open Data infrastructure by establishing a tool bundle allowing for FAIR publication of data for a multitude of scientific communities. For reference, the experimental offers at DESY include over 200 experimental techniques at 21 synchrotron beamlines which have been made use of by over 3000 scientists during 2019. The amounts of data being produced are generally under the control of the respective experiment's primary investigators and are stored in the facility's storage systems for later re-use after the so-called embargo time has passed. During embargo time, the scientists who conducted the experiment are granted exclusive exploitation rights for the data. The policies of funding agencies and scientific journals foresee publication of scientific data under FAIR principles which is not as common and as extensive as would be necessary to make use of the obtained results. In order to emphasize our notion of importance for open and FAIR data publications, we deployed a metadata catalogue that delivers not only human- and machine-readable formats of metadata including a structured description of the experimental parameters for searchability but also concrete data locations. The latter aspect makes for ease of access to the data by either downloading it or even viewing it with additionally provided web services that allow for data exploration enabling scientists to check the data's usefulness before having to download potentially large datasets. Organizing the metadata in the catalogue is a task tackled by providing schema building blocks to the scientific communities allowing them to seamlessly integrate their community specific data descriptions into the catalogue schema while ensuring that the catalogue contents are up to standards through additional curation by the respective scientific communities themselves. Delivering these building blocks and supporting the construction of fully functional metadata schemata that over more enable a substantial increase in metadata quality by implementing automated validation mechanisms against the schemata is one of the main aspects to be described in the talk. The organizational and functional dependencies of all other deployed services including DOI minting are finally described in an architectural overview that is offered as a blueprint to all institutions interested in establishing this suite of services themselves with the intention to create an added value to their scientists

Corresponding Author: Tim Wetzel, tim.wetzel@desy.de

ID T29

Automated Data Integration from Heterogeneous Sources including electronic lab notebooks using LinkAhead

Author: Florian Spreckelsen[1]
Co-author: Alexander Schlemmer[1], Henrik Tom Wörden[1], Timm Fitschen[1]

[1] IndiScale GmbH, Göttingen

Many scientific projects rely on a multitude of different software systems for data storage and data exchange Keeping data findable and accessible can be challenging, especially if data has to be shared between different sites, working groups and institutes. The open source software LinkAhead provides a powerful framework for managing complex data integrated from heterogeneous data sources. LinkAhead is built in a way that data models can be extended and adapted to future requirements by the researchers at any time. Its extendable crawler framework can be adapted to automatically import data and meta data from different repositories and ELN systems, like elabFTW or PANGAEA. LinkAhead also provides a graphical web-interface, as well as an API for automated queries. The software can be fine-tuned to different scientific disciplines and is already used productively at multiple Helmholtz institutes, including GEOMAR, AWI and KIT.

Corresponding Author: Florian Spreckelsen, f.spreckelsen@indiscale.com

Room Barcelona, 12:00-13:00

Keynote 3

Challenges and potential solutions for making science more open and FAIR

Speaker: Konrad Förstner[1]

[1] ZB MED – Information Centre for Life Science

At its core, science must serve society and contribute to solving its challenges. Enhancing the efficiency, transparency, and collaborative nature of scientific endeavours is essential for achieving this goal. Open Science and the FAIR principles aim to improve the scientific process by promoting openness, accessibility and efficiency. Metadata, the essential information that contextualizes research data, plays a crucial role in unlocking the full potential of scientific data. Translating these ideals into practical implementation, however, presents significant challenges. These include the absence of established standards and tools, as well as the slow adoption of these principles within scientific communities due to a lack of supporting incentives and norms. Additionally, the integrity of (meta)data infrastructure is increasingly threatened by political interference, which can hamper scientific progress globally. Recent events have demonstrated that research infrastructure can be significantly impacted by policy changes and funding cuts. These developments underscore the need to rethink strategies for sustainable research data management and how research infrastructures remain reliable for the global scientific community.

The talk aims provide the foundation for a joint discussion that brings together the perspectives and experiences of different research communities.

Tuesday, 12:30-13:30 & 15:30-17:00

Poster & Demo Sessions – All Demos

ID D01

The InvenioRDM repository platform as a key building block of collaborative and FAIR data infrastructures

Author: Martin Fenner[1]

[1] Front Matter

InvenioRDM is an open source repository platform that powers a growing number of repositories for data, software and digital documents, with the largest and oldest instance being Zenodo at CERN. The software was started at CERN, but now has more than 25 project partners, including several from Germany. The software can become a core building block for FAIR data infrastructures at Helmholtz via one or more instances that focus on specific content types and/or domains. In this demo I will showcase the core features of v12 of the software, released in August 2024, and discuss the deployment and customization of the software. If time permits, I will also share some of the features that are on the development roadmap.

Corresponding Author: Martin Fenner, martin@front-matter.io

ID D02

Hands-on semantic data management with LinkAhead: Increased data findability and reusability

Author: Alexander Schlemmer[1]
Co-author: Florian Spreckelsens[1], Daniel Hornung[1]

[1] IndiScale GmbH, Göttingen

In this hands-on workshop, we introduce the open source software LinkAhead, which promotes agility in semantic data management: LinkAhead is a semantic research data management system, facilitating enhanced data findability and reusability through data embedding into context. Its flexible data model (the data structure can be changed without migration of existing data) allows to leverage existing standard ontologies, promoting transparency, interoperability and collaboration across diverse research domains. LinkAhead can be fine-tuned to different scientific disciplines and is already used productively at multiple Helmholtz institutes, including GEOMAR, AWI and KIT. Data management is essential for the storing, searching, retrieving and analyzing of data sets along with their contextual connections, ensuring their usability for current and future users. Effective data management not only ensures the reuse of valuable data by current and future users but also enhances its discoverability ("Where can I find the training data for sensor X from setup Y?") and utility through contextual embedding ("What were the experimental settings for data collection in project P, and what were the associated challenges?"). Thus, LinkAhead provides an effective Technical Solution for Findable and Machine-Readable Metadata (Topic 5) and supports preparing FAIR open data, enables data collaboration, and fosters knowledge exchange. This workshop consists of a live demonstration of LinkAhead and its Python client, and participants are encouraged to follow along on their own machines. A Jupyter notebook will be made available online before the session.

Workshop participants (max. 30) will learn these LinkAhead skills:

- Understand, create and edit data models
- Semantic queries, also for linked data sets
- Add and retrieve data

Corresponding Author: Alexander Schlemmer, a.schlemmer@indiscale.com

ID D03

Streamlining Sample FAIRification in the active research phase: A Demonstration of IGSN Registration through RSpace

Author: Tilo Mathes[1]

[1] Research Space, UK

This demonstration showcases RSpace's integrated solution for FAIRifying physical samples by registering International Generic Sample Numbers (IGSN) as part of researcher workflows within its electronic laboratory notebook and sample management system. Physical samples represent crucial objects to describe the research process and improve the discoverability and reproducibility of the resulting research outputs. Yet their documentation and identification are often disconnected from digital research workflows, which typically results in poor documentation of research results and poor discoverability of physical sample data and metadata. RSpace addresses this challenge by embedding persistent identifier registration within sample management workflows, regularly used in the active research phase. Our demonstration will introduce RSpace's sample management system, Inventory, and provide a step-by-step walkthrough of the IGSN registration process within RSpace. The demonstration will show how researchers can seamlessly document sample metadata, generate compliant IGSN records, and register identifiers without leaving their sample management tool. We will demonstrate how the system ensures metadata compliance with the IGSN schema, and how RSpace uses the IGSNs to create rich research documentation and to facilitate the discovery of research samples through publicly discoverable landing pages.

Corresponding Author: Tilo Mathes, tilo.mathes@researchspace.com

ID D04

A deep dive into DMPonline integrations to foster semantic and technical interoperability between research tools and domains

Author: Agnes Jasinska[1]
Co-author: Andrea Davanzo[1], Glenys Jacob[1], Marta Nicholson[1], Don Stuckey[1]

[1] Digital Curation Centre (DCC), The University of Edinburgh, UK

Data management planning becomes more effective when DMPs make use of recognized identifiers and metadata standards to enable semantic and technical interoperability. The Digital Curation Centre's DMPonline service follows these principles, allowing active integration across research tools and domains. This commitment aims to enable FAIR data principles and reproducible research. In our demo session, we will showcase several integrations involving DMPonline, highlighting our recent accomplishments and ongoing developments. Key aspects of our presentation will include: - Integration of DMPonline accounts with user identities, such as ORCID, to streamline user experience and data continuity. - Enhancement of DMPonline accessibility through APIs, allowing seamless data exchange and interaction. - Collaborative integrations with external services, including the OECD research domains, the OpenAIRE Graph Search API for the automatic retrieval of research output metadata, and our partnership with RSpace for enhanced research management. Participants will also work collaboratively to identify the challenges to achieving greater DMP interoperability and brainstorm possible solutions that employ automation, sophisticated workflows, and robust metadata management within and across various research domains.

Corresponding Author: Agnes Jasinska, agnes.jasinska@ed.ac.uk

ID D05

Enhancing Metadata Handling in Research Software

Author: Mustafa Soylu[1]
Co-authors: Volker Hofmann[1], Stefan Sandfeld[1]

[1] Forschungszentrum Jülich (FZJ)

Effective management of software metadata is important to ensure their discoverability, reproducibility, and overall quality. This poster introduces two tools designed to simplify and enhance research software development (RSD) metadata management: fair-python-cookiecutter and somesy. FAIR-python-cookiecutter is a git repository template that provides a structured starting point for Python projects, enabling researchers and developers to integrate metadata effortlessly. It promotes best practices in software development while aligning with key standards such as the DLR Software Engineering Guidelines, OpenSSF Best Practices, REUSE, CITATION.cff, and CodeMeta. The template also incorporates somesy to strengthen the FAIR principles (Findability, Accessibility, Interoperability, and Reusability) of software metadata. In response to user feedback, recent updates have introduced Poetry v2 support, dependency upgrades, and pre-commit tool enhancements, ensuring the template remains a relevant and reliable resource. Somesy (Software Metadata Synchronization) is a command-line tool that automates metadata consistency across different project files. By supporting formats like CITATION.cff and CodeMeta, it ensures that critical details such as project names, versions, authors, and licenses remain synchronized. This eliminates the need for manual updates, allowing developers to focus on core development tasks. Designed for cross-platform compatibility, somesy works seamlessly on Linux, Windows, and macOS. The latest updates to Somesy introduce, amongst other features, Poetry v2 integration, enhanced validation options, support for ORCID IDs as plain strings, entity (organization) support. We further streamlined metadata handling by Somesy to facilitate post-hoc integration into existing projects. Both tools are actively maintained based on user feedback and continuously evolving RSD best practices. This way we ensure obust, user-friendly, and up-to-date tools for the community.

Links:

- [1] https://github.com/Materials-Data-Science-and-Informatics/fair-python-cookiecutter2-
  https://pypi.org/project/somesy

Acknowledgements: The presented content was created by the FAIR data commons & Hub Information of the Helmholtz Metadata Collaboration (HMC) at Forschungszentrum Jülich. HMC is an Incubator platform of the Helmholtz Association within the framework of the Information and Data Science strategic initiative.

Corresponding Author: Mustafa Soylu, m.soylu@fz-juelich.de

ID D06

Shepard - the modern way to handle research data

Author: Felix Lettowsky[1]
Co-authors: Roland Glück[1], Patrick Kaufmann[1], Jessica Friedline[2], Wiglef Rehm[2]

[1] German Aerospace Centre (DLR), [2] Xitaso GmbH, Augsburg

This demonstration is related to the talk about the shepard system and serves mainly to demonstrate the usage of the frontend of the shepard system (the purpose of the talk is to introduce its basic structure and functionalities). In developing shepard, significant efforts have been made to enhance the usability of the API and especially the web-based frontend. Recognizing that the usability of such tools is crucial for acceptance - particularly for users who are not as technically versed - comprehensive user research has been conducted and usability criteria have been applied. Keeping in mind that complex tasks involving larger volumes of data and requiring automation will still necessitate the API, the frontend has been designed to mirror the structure and terminology of the backend, while also providing help for the user to understand the underlying concepts, such as the separation between data storage and the contextualization layer. Consequently, the frontend can serve as a gateway to real-world adoption of organized, formalized, and accessible data structuring. Adhering to interaction design, the frontend includes features that will enhance intuitive usage by applying principles from day-to-day learned software use. User research has uncovered possible incentives for users to switch from traditional folder systems on their personal computer to the more structured environment of shepard: For instance, shepard now includes a "lab journal" feature that enables users to document important information alongside their structured metadata and other formalized contexts like semantic annotations. The demonstration aims to present the functionality of the frontend both on data created both locally on the fly during the demonstration and on real data stored on the already running system of the German Aerospace Center in Augsburg. This shows also that shepard gained already considerable acceptance at the German Aerospace Center since also data from other institutes except Augsburg.

Corresponding Author: Felix Lettowsky, felix.lettowsky@dlr.de

ID D07

NeXusCreator – Standardizing Science, Simplifying Data

Author: Hector Perez Ponce[1]
Co-author: Heike Görgiz[1], Rolf Krahl[1]

[1] Helmholtz-Zentrum Berlin (HZB)

NeXusCreator is an open-source tool developed at Helmholtz-Zentrum Berlin (HZB) to simplify the creation of NeXus/HDF5 files without requiring programming expertise. It addresses the challenge of heterogeneous data storage across BESSY II beamlines, where various file formats such as SPEC, TIFF, and DAT, among others are used. By standardizing data structures, NeXusCreator supports FAIR data management principles, ensuring that scientific data is Findable, Accessible, Interoperable, and Reusable, with a particular focus on the last two principles. The tool works by converting structured NeXus definition files (ASCII files with extension .nxd) into NeXus files (HDF5 binary files with the .nxs extension) and can be integrated into Python scripts for automation. Additionally, NeXusCreator facilitates data validation and exchange using Punx, enabling researchers to efficiently verify file correctness and extract structured NeXus definitions from existing NeXus files.NeXusCreator is currently being used to NeXus-compliant instrument and application definitions for multiple beamlines, including FEMTOSPEX, PEAXIS, PINK, and mySpot. The tool can also integrate with automation frameworks like SECoP and Bluesky, enabling real-time data standardization. It supports external datasets through Readers and Converters, allowing seamless transformation of experimental data from various sources. NeXusCreator can handle single scans or multiple scans within a single file or multiple files while organizing outputs in a structured manner for efficient data retrieval. Future developments include a graphical user interface (GUI) and expanded integration capabilities. As an open-source project, researchers and institutions are encouraged to collaborate and contribute to its ongoing development, further enhancing its functionality.

Corresponding Author: Hector Perez Ponce, hector.perez_ponce@helmholtz-berlin.de

ID D08

Open-source application for rich standardized metadata management

Author: Mariana Montes[1]
Co-authors: Paul Borgermans[1], Danai Kafetzaki[1], Joachim Bovin[1], Jef Scheepers[1], Mustafa Dikmen[1], Ingrid Barcena Roig[1]

[1] KU Leuven, Belgium

Metadata plays a crucial role in active research data management, especially when governed in a systematic way. However, metadata management, from metadata creation to consumption, is a challenge, especially without software systematization. In the context of ManGO, the active data management platform of KU Leuven built on top of iRODS, we developed the metadata schema manager, an open source web application to support usage of metadata and assist researchers in collaborative and consistent data documentation. Quality metadata are then promoting contextualization, data findability, and driving of automation. The cornerstone of our application is a rich user interface to pragmatize project metadata needs and seamlessly connect the input to backend specifications. The user interface, built in JavaScript and Bootstrap, consists of templates and methods to create a metadata schema (a form), attach metadata in files and folders, validate, and display. The schema consists of a collection of fields of different types: from different scalar input fields, through multiple-choice fields, to composite fields. The fields are added from scratch, copied from previous work, or added from a predefined library. Considering the active nature of data, a life cycle was designed for the schema, such that its status is "draft", "published", or "archived", and only stable versions are used for metadata annotation, while the schema can evolve into new versions. In the backend, the schema is stored in JSON format, and it serves as documentation and as validator of new inputs using open-source python modules. To accommodate different types of metadata, we enriched the backend flat structure to allow hierarchical structures, using namespaces, such that it is possible to maintain links between metadata fields and generate nested structures. Metadata is stored in an SQL database, where metadata can be queried from, and consumed according to inherent system functionalities and project needs. Via the metadata schema manager, we expect users to design and use schemas to increase usability and uniformity of metadata, thereon facilitating collaborative and consistent data annotation and enrichment during the active data phase.

Corresponding Author: Mariana Montes, mariana.montes@kuleuven.be

ID D09

SciCat Integration at MLZ - Infrastructure and Live Demo

Author: Christian Felder[1]
Co-authors: Alexander Zaft[1], Christian Trageser

[1] Forschungszentrum Jülich (FZJ)

This presentation provides an overview about the current status of SciCat at MLZ and the underlying infrastructure. Data acquisition and metadata capture are decoupled based on the RabbitMQ message broker. Information from various sources, such as the user office system, sample environment and the instrument are aggregated in the Networked Instrument Control System and transmitted in messages to our central Kubernetes cluster. Messages can be processed by secondary user services, e.g. the ingestor. The ingestor populates the SciCat catalogue with relevant metadata information from the experiment. A live demonstration using a virtual instrument will conclude the talk.

Corresponding Author: Christian Felder, c.felder@fz-juelich.de

ID D10

A demo on metadata extraction tool for machine-actionable Software Management Plans

Author: Suhasini Venkatesh[1]
Co-authors: Venkatesh Suhasini[1], Dhwani Solanki[1], Dietrich Rebholz-Schuhmann[1], Leyla Jael Castro[1]

[1] ZB MED - Information Centre for Life Sciences

Research software is core to science reproducibility; however, reproducibility is hindered by the lack of documentation on how to (re)use research software. Documentation for humans in the form of readmes, technical documents, or tutorials commonly lack the richness and depth needed by other researchers to continue any effort built on top of existing research software. Thanks to discussions on good practices for research software (e.g., FAIR for Research Software -FAIR4RS, and Software Management Plans - SMPs), researchers are slowly becoming more conscious of the need to share and document their software. Despite these efforts, machine-readable metadata (aligned to FAIR4RS, in the form of semantically structured metadata) is even more scarce. Codemeta, a vocabulary extending schema.org, has emerged as a common practice for research software metadata, although its adoption is not widespread. Furthermore, it does not cover some of the metadata considered in SMPs and aligned to quality indicators. We provide a vocabulary supporting machine-actionability for SMPs, the maSMP ontology, based on Codemeta, schema.org, and Bioschemas. One of the barriers for metadata adoption is the common extra effort required, many times relying on researchers filling an extra file or form with information that is already present in their code repository. Here we present our metadata extraction tool, aligned to our maSMP metadata schema, and including an API and a web-based end-user interface. It currently supports direct metadata extraction from GitHub repositories. We plan to extend it to GitLab and OpenCode, as well to include AI-based capabilities to extend the metadata coverage to elements that may not be not explicit in the code repository (e.g., deployment instructions). In addition to supporting our maSMP metadata schema, we will also plan to provide a RESTful API to facilitate integration with existing software management systems and workflows (e.g., Research Data Management Organizer -RDMO). This work will help improve documentation, reproducibility, and the overall quality of research software. By automating metadata extraction and aligning it to maSMP and FAIR4RS, we aim to foster better practices in research software management across scientific domains.

This work is part of the NFDI4DataScience project funded by the German Research Foundation (DFG), project number 460234259.

Corresponding Author: Suhasini Venkatesh, venkatesh@zbmed.de

Tuesday, 12:30-13:30 & 15:30-17:00

Poster & Demo Sessions – All Poster

ID P01

regimo: Integration of Electronic Lab Notebooks in the Publication Workflow

Author: Mohamed Anis Koubaa[1]
Co-author: Karl-Uwe Stucky[1]

[1] Karlsruhe Institute of Technology (KIT)

The linking of existing tools for the assessment and handling of metadata is useful for the publication of such metadata on open platforms. The workflow described here uses data from RDMO (a Research Data Management Organiser) and from Kadi (an Electronic Lab Notebook) to facilitate the publication of metadata on the OEP (an Open Platform for registration of linked Data), which are semantically augmented by using the TIB Terminology service. RDMO projects deliver a checklist for metadata input which is mapped by regimo to metadata schemas of the platform on which the data should be registered. By connecting to RDMO, regime supports the reusability of administrative metadata sets within the scope of a research project. Data to be published are contained in Kadi database. These data are structured and can be gathered by regimo. Additionally, field-specific metadata are delivered and can be mapped to terms from the TIB terminology service. It is envisaged to make regimo adaptable to arbitrary metadata platforms in order to increase the outreach by mapping properties of different schemes from the respective platforms.

Corresponding Author: Mohamed Anis Koubaa, mohamed.koubaa@kit.edu

ID P02

The Nuclear, Astro, and Particle Metadata Integration for eXperiments (NAPMIX) project

Author: Andrew Kishor Mistry[1]
Co-author: Ivan Knezevic[1]

[1] Helmholtz Centre for Heavy Ion Research (GSI)

The Nuclear, Astro, and Particle Metadata Integration for eXperiments (NAPMIX) project was recently awarded funding within the scope of the OSCARS call on open science and began work in December 2024.The project aims to facilitate data management and data publication under the FAIR principles for the NAP communities on the European level by developing a cross-domain metadata scheme, input mechanism for end users, and machine, and human readable outputs. A core component of the scheme is its nodal, multi-layered schema structure, allowing metadata enrichment from multiple domains, while establishing overlaps for enhanced versatility. This comprehensive approach supports the unification of data standards across various research institutions, promoting interoperability and collaboration on a European scale. Our efforts include the development of a user-friendly frontend generator to facilitate metadata input and also allow users to specify field-specific attributes, customize generic names to suit their needs, and export schemas in various formats such as JSON and XML, adhering to different nomenclatures. In addition, API's will be developed to enable automated metadata generation. The project involves RIs and ESFRIs, and leverages synergies from existing Open Science initiatives such as, ESCAPE, EURO-LABS, and PUNCH4NFDI. In this contribution, we will present an update on the first stages of the project.

Corresponding Author: Andrew Kishor Mistry, a.k.mistry@gsi.de

ID P03

Beyond Compliance: Human-Centered FAIR Data Tools & Management

Author: Hajira Jabeen[1]

[1] University of Cologne

The FAIR (Findable, Accessible, Interoperable, Reusable) principles are transforming research data management, yet a critical challenge persists—many FAIR-enabling tools are developed without meaningful engagement from the researchers and data stewards who are meant to use them. As a result, these tools face low adoption rates, slow uptake, and usability barriers that hinder their intended impact. Existing Research Data Management (RDM) solutions—such as Dataverse, DMPonline, FAIRsharing, OpenAIRE, RDMkit, CEDAR, RO-Crate, and DataCite Metadata Schema—offer valuable functionalities. However, they often suffer from rigid designs, domain insensitivity, and steep learning curves, making them inaccessible or impractical for diverse research communities. The missing piece? Human-centered design and stakeholder-driven development. This poster highlights the need for a paradigm shift in FAIR tool development, advocating for user-centered strategies that integrate researchers, data stewards, and domain experts throughout the design process. We propose a roadmap for a FAIR metadata tool that prioritizes:- Stakeholder Co-Design - Engaging users early through participatory design and iterative prototyping.- Domain-Specific Adaptability - Allowing customization of metadata standards for different research fields.- Automation & Interoperability - Leveraging AI-assisted metadata suggestions and seamless integrations with existing platforms.- Intuitive & Seamless UX - Simplifying metadata workflows to encourage adoption and compliance.- Scalability & Expandability - Ensuring the tool can grow with evolving research needs and integrate seamlessly with existing RDM platforms. By embedding human actors at the core of FAIRification efforts, we can create tools that are not only FAIR in principle but also fair in practice—usable, adaptable, and embraced by the research community. This poster invites discussions on building a truly user-centered FAIR ecosystem that moves beyond checkboxes toward genuine usability and impact.

Corresponding Author: Hajira Jabeen, hajirajabeen@gmail.com

ID P04

Advancing Supramolecular Data Integration: Automated Metadata Extraction and Binding Affinity Predictions for SupraBank

Author: Frank Biedermann[1]

[1] Karlsruhe Institute of Technology (KIT)

The MetaSupra project enhances SupraBank, a FAIR-compliant repository for physicochemical interaction data, by automating metadata extraction and integrating quantum-chemical molecular properties. Our work addresses two key challenges: (1) efficient data extraction from chemical literature, and (2) accurate binding affinity predictions for host-guest complexes. We developed a Python-based PDF annotation tool that automates metadata extraction using regular expression searches and Large Language Models (LLMs). This tool identifies binding constants, solvents, and other chemical parameters, reducing manual curation efforts. It includes an interactive PDF viewer with categorized highlights and unit conversion for binding affinities (Ka, Kd, logKa). Following beta testing, the script will be openly available on GitHub, facilitating widespread adoption. For binding affinity predictions, we conducted classical and QM/MM molecular dynamics (MD) simulations to refine free energy calculations of cucurbit[7]uril (CB7) with adamantane-based guest molecules. Our GAFF reparameterization, based on quantum mechanical (QM) calculations, improved force field accuracy. Preliminary QM/MM MD simulations further validated these refinements, highlighting differences in binding stability. These methodologies will enable high-throughput machine-learning-assisted predictions, contributing to the enrichment of SupraBank. By bridging computational chemistry and automated data curation, MetaSupra streamlines data integration for supramolecular chemistry. Our next steps include public deployment of our PDF annotation tool and expanding machine-learning applications for molecular binding affinity predictions.

Corresponding Author: Frank Biedermann, frank.biedermann@kit.edu

ID P05

A Description Framework for Research Software and Metadata Publication Policies

Author: David Pape[1]
Co-authosr: Oliver Bertuch[2], Tobias Huste[1], Oliver Knodel[1], Michael Meinel[3], Sophie Kernchen[3], Nitai Heeb[2], Christian Meeßen[4]

[1] Helmholtz-Zentrum Dresden-Rossendorf (HZDR), [2] Forschungszentrum Jülich (FZJ), [3] German Aerospace Centre (DLR), [4] German Research Centre for Geosciences (GFZ)

The curation of software metadata safeguards their quality and compliance with institutional software policies. Moreover, metadata that was enriched with development and usage information can be used for evaluation and reporting of academic KPIs. Software CaRD ("Software Curation and Reporting Dashboard"; ZTI- PF-3-080), a project funded by the Helmholtz Metadata Collaboration (HMC), develops tools to support the curation and reporting steps of the research software publication process. The dashboard will present metadata collected by the HERMES workflow in a graphical user interface, assess compliance with a configurable set of policies, and highlight issues and breaches. It will be usable both standalone and in a CI/CD context. As a first step in the project, and as a foundation for the curation dashboard, a description format for software publication policies had to be developed. Our solution takes an approach that allows for configuration at different levels of abstraction: Low level building blocks describe metadata (e.g., CodeMeta) validation in terms of the Shapes Constraint Language (SHACL). A higher-level configuration language allows users to reuse and parameterize these components. This makes Software CaRD usable for RSEs, management, and policy makers, and it allows for customization that facilitates usage in different research institutions. This poster submission presents our approach, showcases example policies, and gives guidance to users of the application.

Corresponding Author: David Pape, d.pape@hzdr.de

ID P06

Onboarding Guide for Data Stewards in Matter

Author: Özlem Özkan[1]

[1] Helmholtz-Zentrum Berlin (HZB)

Effective data stewardship is crucial for ensuring high-quality, reusable, and interoperable research data. This onboarding guide provides a structured introduction to the roles and responsibilities of data stewards in research field matter, equipping them with essential knowledge on metadata management, Open Science principles, and FAIR data practices. The guide is divided into two sections: 1. Core Concepts of Data Stewardship: covering research data management (RDM), metadata fundamentals, FAIR principles, and data reusability. 2. Matter-Specific Considerations: exploring metadata strategies, NeXus standards, Electronic Lab Notebooks (ELNs), and institutional data policies. By combining theoretical foundations with domain-specific practices, this guide empowers data stewards to support researchers effectively and foster a sustainable data ecosystem in matter sciences.

Corresponding Author: Özlem Özkan, oezlem.oezkan@helmholtz-berlin.de

ID P07

Improving research data management for samples: the SEPIA Sample Database for Metadata Storage and Exchange

Author: Mojeeb Rahman Sedeqi[1]
Co-authors: Rolf Krahl[1], Katherine Rial[1], Heike Görzig[1]

[1] Helmholtz-Zentrum Berlin (HZB)

SEPIA (Sample Essentials, Persistent Identifiers & Attributes System): The overall goal of the SEPIA project is to enrich the research data collected at at matter facilities in Helmholtz by providing a better description of the sample being measured. The key for this endeavour is to track the full history of the sample and to collect all the information generated on the way. To this end, a database is created to identify the individual sample and allow referencing of it unambiguously. This poster will explore the architecture and functionalities of the SEPIA project, highlighting its role in improving data accessibility and interoperability among researchers and labs, and implementation in a first use case at the Helmholtz-Zentrum Berlin (HZB). We will discuss the potential impact of SEPIA on collaborative research efforts, data sharing practices, beamlines and the overall improvement of scientific research through better metadata management.

Corresponding Author: Mojeeb Rahman Sedeqi, mojeeb.sedeqi@helmholtz-berlin.de

ID P08

Adapting Metadata Training for Health Scientists: The Fundamentals of Scientific Metadata for Health

Author: Hamideh Haghiri[1]
Co-author: Özlem Özkan[2], Marco Nolden[1]

[1] German Cancer Research Centre (DKFZ), [2] Helmholtz-Zentrum Berlin (HZB)

Scientific metadata plays a vital role in making datasets more findable, accessible, interoperable, and reusable (FAIR). The Fundamentals of Scientific Metadata course (Gerlich et al, https://zenodo.org/doi/10.5281/zenodo.10091707) provides a structured introduction to key concepts in metadata documentation and annotation, including structured metadata, JSON and JSON Schema, enabling technologies and standards, as well as web location and identifiers. To better support scientists working with health data, we adapted this course into The Fundamentals of Scientific Metadata for Health. Key modifications include replacing generic teaching examples with health-related examples, making the content more understandable and tangible for health researchers. Additionally, the original hands-on tasks, which were based on a non-health dataset, were replaced with exercises using a health-related dataset. In this poster, we describe these adaptations, detailing the rationale behind the changes. We also present participant feedback from the first delivery of the adapted course, offering insights into its impact and identifying areas for future improvements in metadata training for health scientists.

Corresponding Author: Hamideh Haghiri, hamideh.haghiri@dkfz-heidelberg.de

ID P09

Metadata Made Easy: A Helmholtz-Focussed Overlay on Croissant

Author: Sebastian Lobentanzer[1]

[1] Helmholtz Munich

Effective metadata management is fundamental to reproducible and scalable scientific research, yet current practices often demand significant manual effort, which harms adoption. Drawing on the foundation laid by the recent Croissant initiative—backed by Google, Hugging Face, and others—we propose a research-focused overlay tailored to the specific requirements of Helmholtz scientists and the broader academic community. Our goal is to implement robust metadata standards while preserving a user-friendly experience, ensuring that best practices become the default rather than an additional burden. Central to our approach is a cohesive set of tools that streamline metadata application across the entire research workflow. We introduce command-line interfaces (CLI) that not only facilitate dataset retrieval and version control at the bash level, but also seamlessly integrate metadata annotations into our existing workflows. Structured annotation of datasets facilitates reducing redundancies in daily practice by, for instance, preventing the repeated download of large public datasets. Downstream of dataset acquisition and annotation, libraries in Python and R allow researchers to load curated datasets with minimal overhead, taking advantage of well-structured documentation and automatic tagging. This integration extends to machine learning pipelines, including foundational models that rely on consistent, large-scale annotations to ensure performance and reliability. By combining principled metadata standards with intuitive toolchains backed by a large community, our initiative empowers data scientists to focus on discovery rather than the logistics of dataset management, while simultaneously encouraging the implementation of FAIR principles. By drawing on the cumulative experience of Helmholtz scientists, we aim to establish solutions that are well-informed by daily practice; by developing those solutions in close collaboration with active scientists, we hope to make it as user-friendly and streamlined as possible for our community. By contributing back into the greater collaborative effort, we leverage open-science principles, improving metadata implementation beyond our scientific scope.

Corresponding Author: Sebastian Lobentanzer, sebastian.lobentanzer@helmholtz-munich.de

ID P10

Sustainable Research Data Management in Helmholtz - Insights from HMC Data Professionals Survey 2024

Author: Sangeetha Shankar[1]
Co-authors: Oonagh Brendike-Mannix[2], Silke Christine Gerlich[3], Markus Kubin[2], Lucas Kulla[4], Christine Lemster[5], Andreas Schmidt[6], Jan Schweikert[6], Karl-Uwe Stucky[6]

[1] German Aerospace Centre (DLR), [2] Helmholtz-Zentrum Berlin (HZB), [3] Forschungszentrum Jülich (FZJ), [4] German Cancer Research Centre (DKFZ), [5] GEOMAR Helmholtz Centre for Ocean Research Kiel, [6] Karlsruhe Institute of Technology (KIT)

Helmholtz Association and to understand their research data management (RDM), FAIR data practices, gaps and needs. The survey is part of HMC's mission to enhance sustainable management of research data and to more closely align its services to data professionals' needs. The questionnaire focused on gaining insights into RDM-related tasks that data professionals in Helmholtz work on, adherence of data management practices to the FAIR principles, RDM-related tools developed and used, alignment with data policies, challenges faced and needs expressed with respect to RDM-related work in Helmholtz. The survey was answered by 156 data professionals employed at different Helmholtz centers. This poster highlights key findings from the survey and provides an overview of the current state of research data management practices in the Helmholtz Association. Survey results show that data professionals work on various RDM-related tasks and operate at multiple organizational levels. With respect to their primary profession, more researchers were observed to be involved in data management than data professionals and librarians combined. However, more than three fourths of the respondents expressed having had no formal training for their data management-related tasks. Nevertheless, the majority of respondents reported to align their data management practices to policies and guidelines of their Helmholtz center. Adherence to findability- and accessibility-related aspects of the FAIR principles was found to be more pronounced than alignment with respect to interoperability- and reusability related aspects. Additionally, the survey data yields a list of commonly used tools for data management as well as tools that are (co-)developed by Helmholtz staff. As over 90% of the respondents reported to facing challenges in their RDM-related work, they expressed an interest in various service formats and topics. These results guide HMC in developing targeted support and services to improve research data management practices in the Helmholtz Association. A comprehensive survey report will be published soon.

Corresponding Author: Sangeetha Shankar, sangeetha.shankar@dlr.de

ID P11

Sample Management System: SAMS

Author: Maren Rebke[1]
Co-authors: Stefan Pinkernell[1], Roland Koppe[1], Sonja Hänzelmann[1]

[1] Alfred-Wegener Institute (AWI)

Comprehensive sample management is the backbone of field research and requires FAIR (findable, accessible, interoperable and reusable) metadata. However, research groups often have their own internal widely variable solutions. To facilitate the digitalisation of the entire data flow, an integrative, centralised sample management system (SAMS) is being developed as a strategic infrastructure at the Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research (AWI). The SAMS represents a repository for metadata on the collection and storage of samples taken on platforms like vessels, land and ice missions. It will ensure the unique identification of samples, allow registration via DataCite and provide a link to the Nagoya documentation. Storage management with individual loan requests and lending documentation will be offered. Furthermore, the SAMS will be integrated into the overall O2A "Observations to Analysis and Archives" workflow* with underlying metadata provided by PANGAEA, REGISTRY and the DSHIP land system. SAMS is based on an open source software, which is a flexible semantic data management system that provides access via a web-based graphical user interface and programmatic API clients. It can be modified and extended for different applications. The database structure, vocabularies and further development are coordinated with other Helmholtz partner centres and elaborated in collaboration with researchers from different disciplines at the AWI. The metadata is based on the specifications developed in the HMC Project FAIR WISH. It will be provided for overarching data portals like the Earth Data Portal for joint presentation and findability of samples from institutes of the Helmholtz research field Earth and Environment, among others.

*https://ieeexplore.ieee.org/document/7271657

Corresponding Author: Maren Rebke, maren.rebke@awi.de

ID P12

OMExcavator: a tool for exporting and connecting Bioimaging-specific metadata in wider knowledge graphs

Author: Stefan Dvoretskii[1]
Co-authosr: Marco Nolden[1], Christian Schmidt[1], Michele Bortolomeazzi[1], Klaus Maier-Hein[1], Josh Moore[2]

[1] German Cancer Research Centre (DKFZ), [2] German BioImaging - Society for Microscopy and Image Analysis (GerBI-GMB)

Bioimaging data volume has greatly increased in recent years, and this trend is paving the way for future important discoveries in Biology and Healthcare. Even though the volume of generated data is huge, its findability, interoperability, and general reusability needs to be improved. In Bioimaging, the widely used RDM system is OMERO, which stores images and their accompanying metadata in the Open Microscopy Environment Data Model. This data model contains the most important metadata about the images like the resolution, and imaging process details, as well as user defined annotation, with a possibility of interoperable export. In this work, we developed a tool called OMExcavator to make the metadata records of images stored in OMERO servers available for semantic exploration. We used OMERO Python API and omero-marshal library to generate a generic JSON-LD metadata representation of OMERO images metadata. These JSON-LD records can then be linked with other resources, in this case, unHIDE - an overarching knowledge graph of the Helmholtz Association as a part of a Helmholtz FAIR data space. Additionally, the LinkedData representation of the OMERO-powered Imaging Data Resource datasets was converted to the CroissantML format, which allowed to fuse OMERO datasets metadata with HuggingFace and Kaggle datasets in a wider Knowledge Graph. This integration can potentially enhancing the AI-readiness of OMERO datasets. This tool can help the relevant communities in the Bioimaging field in Germany and beyond to share the Bioimaging metadata in wider Knowledge graphs. Furthermore, other scientific domains may face similar problems connected to exporting the domain-specific metadata and find the described approach useful for solving them.

Corresponding Author: Stefan Dvoretskii, stefan.dvoretskii@dkfz-heidelberg.de

ID P13

Knowledge Graphs for Scientific Data: Expanding Metadata Integration and Categorization

Author: Stanislav Malinovschii[1]
Co-authors: Emanuel Söding[1], Dorothee Kottmeier[2], Sören Lorenz[1], Andrea Pörsch[3]

[1] GEOMAR Helmholtz Centre for Ocean Research Kiel, [2] Alfred-Wegener Institute (AWI),
[3] German Research Centre for Geosciences (GFZ)

Knowledge Graphs help connect and organize information from diverse sources and entities, enabling advanced search and filtering techniques on large datasets while revealing hidden connections and dependencies. However, their effectiveness depends on well-harmonized and standardized metadata. To explore options for metadata harmonization in existing repositories, we have worked with a large dataset derived from the GFZ Data Services (https://dataservices.gfz-potsdam.de/portal). To enrich the dataset descriptions, we have significantly expanded the number of categorized keywords by extracting them from titles, descriptions, and metadata fields. This approach has allowed us to capture a more detailed and structured representation of the datasets. Furthermore, we successfully mapped datasets to specific research projects, such as ArboDat+ and GEOFON, using DOI-based metadata. To further enhance metadata quality and improve the findability of publications, we introduced new categorization dimensions, including Event, Geographic Feature, Material & Sample Type, Data & Format, Process & Phenomena, and Human Impact & Land Use, etc. These additions enable more precise metadata representation within knowledge graphs. Additionally, we developed automated techniques to detect and resolve inconsistencies in metadata across repositories, improving data harmonization and interoperability. In this poster, we present our updated methods, challenges, and results, including statistics on harvested metadata, classification accuracy, and insights into repository-specific metadata variations. We also discuss the potential for constructing more scientifically relevant knowledge graphs and provide recommendations for improving metadata quality and interoperability across datasets.

Corresponding Author: Stanislav Malinovschii, smalinovschii@geomar.de

ID P14

Development of an electronic lab notebook at the Helmholtz-Institute Freiberg for sample management and documentation of analytical methods - lessons learned from the official testing phase

Author: Theresa Schaller[1]
Co-authors: Thomas Gruber[1], Florian Rau[1]

[1] Helmholtz-Zentrum Dresden-Rossendorf (HZDR)

At the Helmholtz-Institute Freiberg for Resource Technology (HIF), researchers develop new technologies to improve circular economy. In this context, different types of samples (e.g. rock samples, recycling material) play an important role. The sample passes through different states and labs - starting at the sample preparation, through the analysis of the particular sample to the final storage. With electronic lab notebooks (ELNs) this entire process is digitized, thus improving findability, accessibility, interoperability and reusability (FAIR) of the samples and their corresponding data. Once the sample is registered in the system, every further work on the sample will be connected to this sample, explicitly. Thus, all important metadata can be recorded digitally in a structured way. At the HIF, we are developing an ELN based on semantic MediaWiki. Since the beginning of 2025 we are now in its official beta-testing phase. Scientists use the system to register their samples and file digital preparation requests. Their experience has significantly helped us to further improve the system. In this contribution, we will discuss the challenges in the development of an ELN and our experience during the official beta-testing phase. Furthermore, we will present the updated structure of our ELN.

Corresponding Author: Theresa Schaller, t.schaller@hzdr.de

ID P15

Agentic Multimodal Workflows for Ontology-based Representation and Knowledge Systems

Author: Mohammad J. Eslamibidgoli[1]

[1] Forschungszentrum Jülich (FZJ)

This project proposes a comprehensive framework that combines agentic workflows with AI-driven technologies to build a dynamic, AI-augmented knowledge graph from multimodal data sources, with a particular focus on multidisciplinary fields such as materials science. Our approach emphasizes data preprocessing, feature extraction, and the construction of a robust knowledge graph. This graph is tightly integrated with both a conversational AI interface and an AI-based recommender system, leveraging agentic workflows to facilitate proactive data interactions and autonomous decision-making. Designed to capture and analyze complex data relationships, the knowledge graph supports semantic search and advanced data retrieval. The system utilizes Neo4j for graph database management, Graph-RAG based language models for comprehensive analysis, and vector similarity search for precise feature extraction. To handle materials science data, we incorporate the Elementary Multiperspective Material Ontology (EMMO), providing a powerful and flexible foundation for managing heterogeneous, high-throughput datasets. By integrating agentic workflows with sophisticated ontology mappings, the system can autonomously evolve and adapt over time, steadily improving its functionality and accuracy in delivering actionable insights.

Corresponding Author: Mohammad J. Eslamibidgoli, m.eslamibidgoli@fz-juelich.de

ID P16

Harmonizing NetCDF Metadata Workflows: A Collaborative Initiative for Enhanced Data Integration and Reusability

Author: Romy Fösig[1]
Co-authors: Björn Saß[2], Sabine Barthlott[1], Klaus Getzlaff[3], Tobias Kerzenmacher[1], Dorothee Kottmeier[4], Katharina Loewe[1], Ulrich Loup[5], Christof Lorenz[1], Florian Obersteiner[1], Corinna Rebmann[1], Emanuel Söding[3], Phillip S. Sommer[2]

[1] Karlsruhe Institute of Technology (KIT), [2] Helmholtz-Zentrum Hereon, [3] GEOMAR Helmholtz Centre for Ocean Research Kiel, [4] Alfred-Wegener Institute (AWI), [5] Forschungszentrum Jülich (FZJ)

To ensure FAIR data (Wilkinson et al., 2016: https://doi.org/10.1038/sdata.2016.18), well-described datasets with rich metadata are essential for interoperability and reusability. In Earth System Science, NetCDF is the quasi-standard for storing multidimensional data, supported by metadata conventions such as Climate and Forecast (CF) and Attribute Convention for Data Discovery (ACDD).While NetCDF can be self-describing, metadata often lacks compatibility and completeness for repositories and data portals. The Helmholtz Metadata Guideline for NetCDF (HMG NetCDF) Initiative addresses these issues by establishing a standardized NetCDF workflow. This ensures seamless metadata integration into downstream processes and enhances AI-readiness. A consistent metadata schema benefits the entire processing chain. We demonstrate this by integrating enhanced NetCDF profiles into selected clients like the Earth Data Portal (EDP, https://earth-data.de). Standardized metadata practices facilitate repositories such as PANGAEA (https://www.pangaea.de/) and WDCC (https://www.wdc-climate.de), ensuring compliance with established norms. The HMG NetCDF Initiative is a collaborative effort across German research centers, supported by the Helmholtz DataHub. It contributes to broader Helmholtz efforts (e.g., HMC) to improve research data management, discoverability, and interoperability. Key milestones include: • Aligning metadata fields across disciplines, • Implementing guidelines, • Developing machine-readable templates and validation tools, • Supporting user-friendly metadata entry. This presentation outlines key challenges, solutions, and the anticipated impact on the geoscientific community. We will present a first version of NetCDF metadata attribute guidelines and invite you to join this initiative.

Corresponding Author: Romy Fösig, romy.foesig@kit.edu

ID P17

Advancing Cross-Domain Data Reuse: The CDIF-4-XAS Project for X-ray Absorption Spectroscopy

Author: Markus Kubin[1]
Co-authors: Patrick Austin[2], Heike Görzig[1], Arofan Gregory[3], Simon Hodson[3], Rolf Krahl[1], Leandro Liborio[2], Abraham Nieva De La Hidalga[4]

[1] Helmholtz-Zentrum Berlin (HZB), [2] Rutherford Appleton Laboratory, Science and Technology Facilities Council, UK (RAL-STFC), [3] The Committee on Data of the International Science Council (CODATA), France, [4] Cardiff University, UK

The Cross Domain Interoperability Framework (CDIF) provides a set of implementation guidelines designed to enhance cross-disciplinary reuse of research data. CDIF encompasses standards and methodologies addressing various interoperability levels crucial for cross-domain data utilization. Its initial version comprises five core profiles: Discovery, Access, Controlled Vocabularies, Data Description for Integration, and Universals, which collectively support the cross-disciplinary implementation of the FAIR principles. The OSCARS-funded CDIF-4-XAS project applies CDIF to enhance the interoperability and reusability of X-ray Absorption Spectroscopy (XAS) data. This initiative aims to streamline data exchange between applications, databases, and institutions, addressing the critical need for interoperable XAS data across multiple research disciplines. Recently, the project published its first deliverable: a comprehensive landscape analysis of standards, vocabularies, ontologies, data formats, and practices in XAS. Building on this, the next phase involves developing semantic descriptions of two XAS community standards using a CDIF profile. The CDIF-4- XAS project anticipates that adopting CDIF recommendations for metadata standardization is expected to significantly enhance the reuse potential of XAS data beyond its original disciplinary context. This standardization is expected to facilitate seamless integration of XAS datasets into other infrastructures, promoting data reuse across diverse research domains including energy, chemistry, and environmental sciences. We invite researchers and data professionals to discuss this approach, share insights and help identify additional use cases for CDIF. Your input will be valuable in further improving data interoperability between scientific disciplines and opening new opportunities for interdisciplinary research.

Corresponding Author: Markus Kubin, markus.kubin@helmholtz-berlin.de

ID P18

A scalable data management framework for flow cytometry research using OMERO

Author: Riccardo Massei[1]
Co-authors: Thomas Hornick[2], Heiko Wagner[2], Susanne Dunker[2]

[1] Helmholtz Centre for Environmental Research (UFZ), [2] German Centre for Integrative Biodiversity Research (iDiv)

Ecologists and evolutionary biologists seek to accurately identify and quantify pollen grains to answer fundamental questions about pollinator effectiveness, community networks, and the evolution of floral traits. However, traditional methods have been limited by time-consuming and labor-intensive measurements, as well as proneness to misidentification. Recent advances in flow cytometry have incorporated imaging capabilities, enabling the high-throughput analysis of cellular morphology and subcellular structures. The integration of machine learning and artificial intelligence algorithms has further enhanced data interpretation, facilitating automated classification and feature extraction. However, the large-scale image and metadata information generated by this approach present significant challenges, including substantial storage requirements, standardization issues, and computational demands for processing high-dimensional information. To address these limitations, we developed a robust Data Management Framework that incorporates optimized data pipelines, cloud-based storage solutions, and improved analytical tools. Our framework utilizes OMERO to manage flow cytometry data, associates specific metadata as key-value pairs for effective data filtering, and employs interactive Jupyter with Galaxy for pollen classification. This approach enables efficient data management, facilitating the full potential of image data in flow cytometry research and supporting our understanding of pollen ecology.

Corresponding Author: Riccardo Massei, riccardo.massei@ufz.de

ID P19

Towards a Common Controlled Vocabulary for Device Types in Helmholtz Research Area Earth and Environment

Author: Dorothee Kottmeier[1]
Co-author: Nils Brinckmann[2], Norbert Anselm[1], Paul Remmler[3], Robert Huber[4], Corinna Rebmann[5], Felix Mühlbauer[2], Romy Fösig[5], Linda Baldewin[6]

[1] Alfred-Wegener Institute (AWI), [2] German Research Centre for Geosciences (GFZ), [3] Helmholtz Centre for Environmental Research (UFZ), [4] Centre for Marine Environmental Sciences at the University of Bremen (MARUM), [5] Karlsruhe Institute of Technology (KIT), 6 Helmholtz-Zentrum Hereon

The Helmholtz Metadata Collaboration (HMC) Hub Earth and Environment aims to establish a framework for semantic interoperability across the diverse research data platforms within the Helmholtz research area Earth and Environment (E&E). Standardizing metadata annotation and harmonizing the use of semantic resources are critical for bridging gaps in data sharing, discovery, and reuse. To develop a common community-driven solution, the HMC, in collaboration with the E&E DataHub, has established the formal "Metadata-Semantics" Working Group, bringing together interested data stewards from key Helmholtz research data platforms in the E&E domain. As part of its strategy to standardize metadata annotation in collaboration with the community, the working group will initially focus on harmonizing device-type denotations in two Helmholtz sensor registries: the O2A REGISTRY, developed at AWI, and the Sensor Management System (SMS), operated by the UFZ, GFZ, KIT and the FZJ. The harmonization process includes the development of a common FAIR controlled vocabulary and establishing a peer-reviewed curation approach for the vocabulary. The common vocabulary enables the creation of referenceable ad-hoc terms when necessary, implements version control and quality control mechanisms, and ensures interlinkage with existing terminologies in the field (e.g., NERC L05, L06, ODM2, GCMD).The vocabulary will be developed in collaboration with experts from various disciplines within Helmholtz E&E, with contributors also from the other data infrastructures so that it will eventually be applicable to other research data entities in Helmholtz E&E (e.g., PANGAEA or GFZ data services). Furthermore, a strategy and governance framework will be developed, allowing for hosting and maintaining the vocabulary using terminology services provided by the Base4NFDI consortium. By consolidating expertise, tools, and infrastructures, this initiative effectively bundles resources from various experts distributed at the Helmholtz centers. We envision that the solution outlined for creating, maintaining and hosting terms for measurement device types is also applicable to the creation of other collaboratively used terminologies relevant in the research field. This effort will result in general recommendations for establishing semantic interlinkages across research data infrastructures.

Corresponding Author: Dorothee Kottmeier, dorothee.kottmeier@pangaea.de

ID P20

Establishing Workflows to Engage Stakeholder Groups in PID Metadata Maintenance

Author: Emanuel Söding[1]
Co-authors: Andrea Pörsch[2], Dorothee Kottmeier[3], Stanislav Malinovschii[1], Sören Lorenz[1]

[1] GEOMAR Helmholtz Centre for Ocean Research Kiel, [2] German Research Centre for Geosciences (GFZ), [3] Alfred-Wegener Institute (AWI)

At the Helmholtz Association, we aim to establish a well-structured and harmonized data space that connects information across distributed data infrastructures. Achieving this goal requires the standardization of dataset descriptions using appropriate metadata. It also in-volves defining a single source of truth for much of the metadata, from which different sys-tems can draw. Persistent Identifier (PID) in the metadata enable the reuse of common information from shared sources. Broad adoption of PID types enhances interoperability and supports machine-actionable data. As a first step, within the Helmholtz research field Earth and Environment, we have agreed to adopt several PID types into our data systems. These include ROR, ORCID, IGSN, PIDINST, DataCite DOI, and Crossref DOI. However, to practically record and integrate this information into our repositories, we must first identify the specific places and stakeholders within institutions where this data is generated and maintained. These stakeholders must then be empowered through clearly defined workflows that make them aware of their roles and encourage them to prioritize metadata management. In this presentation, we introduce suggestions for what such workflows could look like. For example, institutional activities such as the hiring process could be leveraged to systematically collect relevant information and feed it into appropriate data workflows. We also identify potential stakeholders who could take responsibility for these processes. Ultimately, all these data streams should converge in the institutional data repository. From there, they can be shared with local, disciplinary, or international partners, contributing to the broader data space.

Corresponding Author: Emanuel Söding, esoeding@geomar.de

ID P21

BeStMeta (Behavioral Standard Metadata): Developing metadata standards and FAIR analysis pipelines for Video Tracking Assays (VTAs) in toxicology and medical sciences

Author: Deborah Schmidt[1]
Co-authors: Riccardo Massei[2], Madhu Nagathihalli Kantharaju[1], Ivan Ezquerra-Romano[1]

[1] Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC),
[2] Helmholtz Centre for Environmental Research (UFZ)

The BeStMeta HMC project was established in 2025 to address the need for standardized metadata schemas and reporting protocols in video tracking assays (VTAs), which are crucial tools for measuring behavioral changes in single or multicellular organisms. Currently, the lack of unified metadata standards hinders reproducibility and consistency in research findings, limiting the potential of VTAs to advance our understanding of assay output and comparability. The BeStMeta objective is to establish standardized metadata formats and reporting schemas to ensure reproducibility of VTAs in toxicological and medical research, thereby improving the transparency and reliability of research outcomes. By increasing the FAIRness (Findable, Accessible, Interoperable, and Reusable) of analysis protocols and algorithm training for VTAs, we will facilitate data sharing and collaboration among researchers, reduce the duplication of efforts, and enable the development of more accurate and robust tracking models. Furthermore, we will develop web-based pipelines and semi-automatic procedures for VTAs metadata enrichment, allowing researchers to easily annotate, manage, and analyze large datasets. This will not only streamline research processes but also broaden the application of VTAs, enabling researchers to explore new possibilities in fields such as neuroscience, pharmacology, and environmental science. BeStMeta will have a great impact on the scientific community, as it will enable to promote the development of innovative solutions for pressing environmental and health challenges. By engaging researchers in all career stages through workshops and sharing events, we will ensure effective use of analysis pipelines and web-based solutions, increase collaboration, and promote the adoption of standardized metadata standards.

Corresponding Author: Deborah Schmidt, deborah.schmidt@mdc-berlin.de

ID P23

Defining Metadata Requirements for a Public Data Center for High-Energy Astroparticle Physics

Author: Victoria Tokareva[1]

[1] Karlsruhe Institute of Technology (KIT)

The KASCADE Cosmic-ray Data Centre (KCDC) grants free unlimited public access to datasets generated by the high-energy astroparticle physics experiment KASCADE, along with several other research projects, since 2013. It also serves as an information and analysis platform in the field of high-energy astroparticle physics for both professional researchers and the general public. The platform provides users with a diverse set of digital objects, including preselected, simulation, and custom-selected datasets, collections of cosmic-ray spectra, publications, manuals, tutorials, and code snippets. The vision of KCDC as a project includes contributing to open data and open science and bringing astroparticle data to humans and machines in a FAIR (Findable, Accessible, Interoperable, Reusable) way. And metadata curation plays a pivotal role in fulfillment of this initiative. This contribution provides an overview on the implementation of the key aspects of metadata modeling for this astroparticle physics use case, such as definition of the user use cases covered, essential elements defined, standards and best practices used, and metadata evaluation and compliance workflows proposed. We will particularly focus on metadata modeling and mapping for experimental datasets and simulations, as well as for digital educational resources. Additionally, we present metadata mapping and harvesting strategies adopted for integration within interdisciplinary data ecosystems. Thus, it contributes to such topics of HMC Conference 2025 as "Metadata Annotation and Management:, "Metrics, Tools, and Workflows for Metadata Assessment", "Technical Solutions for Findable and Machine-Readable Metadata".

Corresponding Author: Victoria Tokareva, victoria.tokareva@kit.edu

ID P24

Metadata extraction, workflows and automation for research data management at KU Leuven

Author: Paul Borgermans[1]
Co-authors: Mariana Montes[1], Jef Scheepers[1], Danai Kafetzaki[1], Joachim Bovin[1], Mustafa Dikmen[1], Ingrid Barcena Roig[1]

[1] KU Leuven, Belgium

At KU Leuven, we advocate for and support sound metadata handling as a key pillar of Research Data Management (RDM) across the whole institution. While working with researchers from a wide range of domains, with diverse types of data and requirements, we strive to identify and develop common (and good) practices, as well as generic processes, from metadata extraction to automation triggers. This constitutes a challenge, as we try to maximize generalizability and customizability in the tools and workflows we offer, and aim to establish a solid framework supporting daily operational capabilities and long term goals in the quality of our research data. With our tools, metadata can be identified and extracted from the earliest stage, i.e. during an (automatic) ingestion phase, when research data is entering our centrally managed data platform. The source of the metadata can take various forms, such as path- and name-based patterns, embedded in domain specific file formats or as related sources such as "sidecar" files. As metadata is extracted, it can be further processed in order to standardize, normalize, validate and filter it, thus increasing its overall quality and consistency. In addition, predefined workflows can be triggered within the data platform to consolidate (e.g. move, combine) and classify research data. By indexing the relevant metadata in a search index, the research data becomes then easily findable within the constraints of various access policies. These tools, written in Python, are mostly developed in-house, with some dependencies on research domain-specific modules and occasional third-party tools, such as Apache Tika for metadata extraction. They are published under an open source license, and we look forward to collaborate with the RDM community at large.

Corresponding Author: Paul Borgermans, paul.borgermans@kuleuven.be

ID P25

Semantic description and integration of Helmholtz digital assets using the Helmholtz Digitization Ontology

Author: Said Fathalla[1]
Co-authors: Volker Hofmann[1], Thomas Jejkal[2], Stefan Sandfeld[1]

[1] Forschungszentrum Jülich (FZJ), [2] Karlsruhe Institute of Technology (KIT)

However, the lack of standardized semantics and interoperable metadata creates heterogeneity that complicates semantic integration and automated processing of these assets. To address and overcome these barriers and to ensure interoperability of the various systems across the Helmholtz Association, we developed the Helmholtz Digitization Ontology (HDO) [1]. This mid-level ontology contains concepts and relationships representing digital assets and processes that appear in the Helmholtz digital ecosystem. The main goal for developing HDO is to serve as a harmonized and machine-actionable institutional reference to represent digital assets and procedures pertinent to their handling and maintenance within Helmholtz. The ontology is aligned with the practices and conventions of the Open Biological and Biomedical Ontologies (OBO). Each class includes rich annotations including labels, definitions, synonyms, grammatical details, comments, and contributor credits partly bilingual in English and German. It is developed sustainably using the Ontology Development Kit (ODK). After the 1st release in late 2024, HDO was adopted and adapted in use cases, bridging the Helmholtz research fields. The latest HDO release [2] includes an extension that introduces classes and properties required to semantically represent FAIR Digital Objects (FDOs) based on the Helmholtz Kernel Information Profile (KIP). This representation provides a foundational semantic framework for FDOs being integrated with semantic web technology such as the Helmholtz Knowledge Graph. With this, we ensure the interoperability of systems in the Helmholtz FAIR data space and the machine-actionability of digital assets within it. A comprehensive HTML documentation of HDO is available online [3].

References

- [1] https://purls.helmholtz-metadaten.de/hob/hdo.owl
- [2] https://codebase.helmholtz.cloud/hmc/ hmc-public/hob/hdo
- [3] https://purls.helmholtz-metadaten.de/hob/HDO_00000000

Corresponding Author: Said Fathalla, s.fathalla@fz-juelich.de

ID P26

Archiving seismic legacy data as part of the MetaSeis Project: establishing a workflow, visualization on maps and linking to the PANGAEA data archive

Author: Estella Weigelt[1]
Co-authors: Daniel Damaske[2], Niklas Selke[2], Stefanie Schumacher[1], Mechita Schmidt-Aursch[1], Janine Felden[1], Janine Berndt[3], Antonie Haas[1], Andreas Walter[1]

[1] Alfred-Wegener Institute (AWI), [2] Centre for Marine Environmental Sciences at the University of Bremen (MARUM), [3] GEOMAR Helmholtz Centre for Ocean Research Kiel

The contribution presents the MetaSeis Project. Aim of the project is to develop a unifying data infrastructure and prepare for future archival of reflection 3D seismic data and active OBS data from recent and future research cruises. We aim to adopt and extend existing standards and interoperable vocabularies in the seismic metadata including metadata quality and validation checks. To ensure long-term archival in the digital library system "PANGAEA" according to the FAIR-principles, a workflow for the integration of future and legacy data sets is established. The contribution presents the way in which the Alfred Wegener Institute makes its seismic data visible and accessible, and the difficulties that arise. Aim is to visualize the marine seismic profiles measured to date on maps via track lines and to make the corresponding data accessible to the geoscientific community via data archive "PANGAEA" as part of the "Open Access Agreement". For a first overview the online portal https://marine-data.de presents the location of the acquired seismic profiles, and displays meta data as expedition data, contact persons, descriptions of the surveys, and a link to the cruise report. JPG-Images are provided for some significant seismic profiles. These maps can be quickly updated after new expeditions to keep the community informed about where data has been collected and where there are still gaps. In a second step the Marine Data Portal is linked to the data archive "PANGAEA". The data itself can be accessed there within the scope of the "Open Access Agreement". The "PANGAEA" archive provides the basis for the data to be permanently visible to the scientific community and to be digitally available for future projects. Another major advantage is that the data is assigned a DOI, which is becoming increasingly important for the submission of publications, research proposals and other applications.

Corresponding Author: Estella Weigelt, estella.weigelt@awi.de

ID P27

Unifying Heterogeneous Medical Image Metadata Using Large Language Models

Author: Elisa Stegmeier[1]
Co-authors: Selen Erkan[1], Rajesh Baidya[1], Philipp Schader[1], Stefan Dvoretskii[1], Constantin Ulrich[1], Klaus Maier-Hein[1]

[1] German Cancer Research Centre (DKFZ)

The reusability of medical image datasets relies heavily on the quality and consistency of their metadata. However, datasets from different sources often adhere to varying standards, leading to inconsistencies that hinderinteroperability. To address this challenge, metadata must be extracted, standardized, and harmonized. As part of the Human Radiome Project, funded by the Helmholtz Foundation Model Initiative, we are collecting datasets from both public sources and clinical partners. The overarching goal is to prepare these data for AI readiness to support the training of a large-scale medical foundation model. To date, we have gathered data from over 1000 distinct sources, each with its own data formats and metadata standards, posing significant challenges for integration and standardization. Manually performing this task is resource-intensive and impractical given the vast amount of data. In our study, we leverage large language models (LLMs) to automate the curation and unification of medical image metadata. We have developed a two-step pipeline. In the first step, Data Curation, LLMs are used to extract and interpret metadata that lack a standardized format. The second step, Metadata Unification, focuses on using LLMs to standardize and map the extracted values across datasets, enabling the creation of structured datasets based on specific parameters. A key challenge is identifying which metadata are available, ensuring their correctness, and selecting the most suitable LLM for extraction and unification. Evaluating the performance of LLMs and ensuring the correctness of extracted metadata remains crucial. To address this, we generate manual ground truths and apply a majority voting mechanism across different LLM outputs, allowing for a more reliable assessment of metadata accuracy. In this poster, we will highlight the importance of standardized and harmonized metadata for medical image datasets, discussing how it enhances dataset interoperability and reusability. We will outline the key challenges posed by heterogeneous data sources, varying metadata standards, and the sheer volume of information that requires processing. Additionally, we will present our approach to addressing these challenges using large language models (LLMs) and share the results we have obtained so far in this ongoing project.

Corresponding Author: Elisa Stegmeier, elisa.stegmeier@dkfz-heidelberg.de

Monday, 10:00-13:00

Workshops

Room tba, 11:00-13:00
ID W01

Search over Multi-Layer Metadata

Author: Carsten Hoyer-Klick[1]
Co-authors: Sebastian Hellmann[2], Jan Forberg[2]

[1] German Aerospace Centre (DLR), [2] University of Leipzig

The SOMMER project (Search over Multi-Layer Metadata for NFDI4Energy Repositories) is an initiative aiming to establish an adaptable, multi-domain metadata catalog to support energy research. Utilizing the existing Databus and MOSS (Metadata Overlay Search System) technologies, the project enhances metadata annotation and search capabilities for diverse datasets in NFDI or HMC contributing significantly to the FAIR (Findable, Accessible, Interoperable, and Reusable) data management goals. The core objectives of SOMMER include refining user interface and experience, fostering community engagement, and enhancing metadata compatibility across various repositories. By incorporating feedback from energy researchers and operators of research data repositories the project intends to create a well-integrated, user-accepted metadata registry. This registry will support various scientific domains while maintaining adherence to recognized standards such as DCAT, SPARQL, and RDF, ensuring interoperability with tools like LDM and TIB Terminology Services. Within the HMC Conference we want to take a specific focus on operators of data repositories and how they can use databus and MOSS to enhance their visibility and searchability of their data. Through the integration of multiple metadata standards, SOMMER addresses the challenges of diverse data sources, enabling researchers to annotate datasets from a central platform. This approach encourages broad adoption within the energy research community by supporting domain-specific and interdisciplinary metadata needs. Ultimately, SOMMER provides a scalable solution with cross-domain search functionalities with the potential to become a reference model for metadata catalogs within the broader NFDI and HMC framework. Within the workshop we want to introduce the current implementation of Databus and MOSS and collect feedback to improve functionalities and the interface to search for research data.

Agenda:

- 30 min:
  Welcome and introduction to Databus and MOSS, demonstration of the current functionalities
- 15 min:
  Time to experiment with Databus and MOSS
- 40 min:
  Discussion on implementation challenges in research data infrastructures
- 5 min:
  Next Step.

Corresponding Author: Carsten Hoyer-Klick, carsten.hoyer-klick@dlr.de

Room tba, 11:00-13:00
ID W02

STAMPLATE & the DataHub Digital Ecosystem: Towards a FAIR Research Data Infrastructure for Environmental Time-Series

Author: Christof Lorenz[1]
Co-author: Ulrich Loup[2], David Schäfer[3], Claas Faber[4], Mihir Rambhia[5], Nils Brinckmann[6]

[1] Karlsruhe Institute of Technology (KIT), [2] Forschungszentrum Jülich (FZJ), [3] Helmholtz Centre for Environmental Research (UFZ), [4] GEOMAR Helmholtz Centre for Ocean Research Kiel,
[5] Helmholtz-Zentrum Hereon, [6] German Research Centre for Geosciences (GFZ)

In environmental sciences, time-series data is crucial for monitoring environmental processes, validating earth system models, training data-driven methods, and understanding climate processes. However, a uniform standard and interface for making such data consistently available according to FAIR principles is still lacking. To address this, seven research centers from Helmholtz Earth & Environment initiated the HMC project STAMPLATE within the DataHub initiative. STAMPLATE aims to establish the Open Geospatial Consortium's SensorThings API (STA) as the central interface, linking it to other community-driven tools and services to foster a digital ecosystem for environmental time-series data. Within STAMPLATE, we developed a thematic metadata profile for STA, enhancing the core data model with domain-specific information. STA has also been successfully integrated into tools for sensor metadata management, time-series management (TSM) systems, and an overarching (meta)data portal for data consolidation and visualization. Particular attention was given to the consistent description of data quality. To achieve this, we integrated the System for Automated Quality Control (SaQC) into our framework and extended the STA data model, enabling interoperable provision of quality information. This session provides an overview of our ecosystem, integrated services, and metadata schema, with hands-on tutorials for selected tools. The highlight will be a bring-yourown- data session, allowing participants to experiment with our tools, use SaQC for flagging their own data, and integrate the results into our infrastructures. Finally, the diverse applicability of our framework is demonstrated through use cases from different communities, such as the Boknis Eck and TERENO observatories. This tutorial is designed for researchers, technicians, and data professionals working with timeseries data from any sensor system.

Main content:

- Introduction to the digital DataHub ecosystem
- STA as generic and modern interface for time-series data
- Hands-on-tuturials of integrated tools and sub-systems
- Bring-your-own-data session

Required previous knowledge for the hands-on and bring-your-own-datasessions:

- Basic skills with Python / Jupyter Notebooks
- Some experience in data processing (e.g, Pandas,...)

Corresponding Author: Christof Lorenz, christof.lorenz@kit.edu

Room tba, 11:00-13:00
ID W03

Publish & utilize SKOS vocabularies with SkoHub

Author: Adrian Pohl[1]
Co-author: Tobias Bülte[1]

[1] North Rhine-Westphalian Library Service Centre (HBZ-NRW)

This hands-on workshop will guide participants through the process of publishing controlled vocabularies encoded in SKOS using [SkoHub Pages] (https://github.com/skohub-io/skohub-pages). SkoHub Pages offers an efficient way to publish vocabularies in both human and machine-readable formats without requiring an additional server using GitHub Pages. After that we will see how to use these vocabularies in reconciliation tasks with [SkoHub Reconcile] (https://github.com/skohub-io/skohub-reconcile)

Agenda:

(1) Introduction to RDF & SKOS (Simple Knowledge Organization System)
- Brief overview of RDF & SKOS and its significance in managing controlled vocabularies
- Examples of SKOS usage in libraries and research

(2) Introduction to SkoHub
- Overview of SkoHub moduls their functionalities
(SkoHub Vocabs, SkoHub Reconcile, SkoHub Shapes and SkoHub Pages)

(3) Hands-on Session: Building and Publishing a Vocabulary
- Step-by-step guide to creating a controlled vocabulary in SKOS format
- Practical exercise: Participants will build their own vocabulary
- Publishing the vocabulary using SkoHub Pages on GitHub in human and machine-readable formats

(4) Reconciliation:
- Introduction to Reconciliation and SkoHub Reconcile
- Practical Exercise: Using the published vocabulary in a simple reconciliation task

(5) Q&A and Discussion
- Addressing participants' questions and discussing potential applications.

Target Audience:

- information professionals, developers interested in Semantic Web technologies, linked data, and vocabulary management, librarians

Prerequisites:

- Participants should have basic knowledge of controlled vocabularies

Corresponding Author: Adrian Pohl, pohl@hbz-nrw.de

Room tba, 10:00-11:30
ID W04

Leveraging the HMC FAIR Data Dashboard: An Interactive Workshop on Enhancing Open and FAIR Data Practices

Author: Gabriel Preuß[1]
Co-authors: Markus Kubin[1], Mojeeb Rahman Sedeqi[1], Oonagh Brendike-Mannix[1], Pascal Ehlers[2]

[1] Helmholtz-Zentrum Berlin (HZB), [2] German Aerospace Centre (DLR)

We propose an interactive workshop focusing on the HMC Dashboard on Open and FAIR Data in Helmholtz, a tool designed to monitor and improve open and FAIR research data practices within the Helmholtz Association. This workshop aims to engage repository managers, data stewards and researchers in exploring the potential of the Dashboard and its role in building a FAIR data space.

The workshop will comprise four key components:

(1) An overview of the Dashboard's capabilities to provide insights into repository usage, trends and data publication statistics across the Helmholtz Association.

(2) An interactive discussion on how repository managers and data stewards can use the dashboard to improve infrastructure, interfaces and data management practices, with a focus on improving FAIR metadata.

(3) An exploration of external factors that can potentially influence the functionality and community adoption of the dashboard, including repository practices, relevant standards, interfaces and technologies, and developments in the broader research data ecosystem.

(4) A collaborative discussion on strategies for assessment of (meta)data quality, addressing potential benefits, challenges, and synergies with existing evaluation frameworks and metrics development.

This workshop will support conference topics focused on human actors in the FAIR data landscape, and metrics, tools and workflows for metadata assessment. Participants will gain practical insights into using the Dashboard to foster community engagement and improve FAIR (meta)data practices, contributing to the realization of a FAIR Data Space. We invite all interested participants to contribute and discuss ideas on leveraging the Dashboard's capabilities effectively and to shape the future of FAIR (meta)data practices in the Helmholtz data ecosystem and beyond.

Corresponding Author: Gabriel Preuß, gabriel.preuss@helmholtz-berlin.de

Room tba, 11:30-13:00
ID W05

CDIF-4-XAS: progress, next steps and invitation to collaborate

Author: Simon Hodson[1]
Co-author: Heike Gorzig[2], Markus Kubin[3], Leandro Liborio[4], Abraham Nieva de la Hidalga[5], Rolf Kraal[2], Arofan Gregory[1]

[1] The Committee on Data of the International Science Council (CODATA), France [2] Helmholtz-Zentrum Berlin (HZB), [3] Helmholtz Metadata Collaboration (HMC), [4] Rutherford Appleton Laboratory, Science and Technology Facilities Council, UK (RAL-STFC), [5] Cardiff University, UK

The Cross Domain Interoperability Framework (CDIF) is a set of practical, implementation-level principles designed to improve data stewardship practices within any community and lower the barriers to cross-domain data reuse. CDIF offers standards and methodologies for achieving different levels of interoperability necessary for reusing data across diverse domains. The first version of CDIF is built around five core profiles that address the essential functions for implementing cross-domain FAIR principles: Discovery, Access, Controlled Vocabularies, Data Description for Integration and Universals. New profiles are now being added, along with further implementation guidelines. The OSCARS CDIF-4-XAS project will use CDIF to enhance the interoperability and reusability of X-ray Absorption Spectroscopy (XAS) data. XAS data is vital for many fields, but its specialised formats and metadata conventions hinder cross-domain use. Data sharing among applications, databases, and facilities is inefficient, leading to the loss of essential experimental information. With increasing data volumes and interdisciplinary collaborations, the need for a more interoperable solution becomes urgent. The first project output, an 'Overview of standards, vocabularies, data formats and practices in the XAS area' has recently been published. The next deliverable will provide a semantic description of two XAS community standards (NXxas for multi-spectra raw and processed data and XDI for single spectra data) using a CDIF profile (XAS-CDIF).

This session will introduce CDIF and the CDIF-4-XAS project; we will describe the findings of the landscape analysis; and describe next steps. These include: exploration of the use of the CDIF Discovery profile and of DDI-CDI data description for variables; characterisation of the HDF5 data structure using DDI-CDI; mappings of key NXxas and XDI concepts. In particular, we will aim to discuss and explore potential use cases: the CDIF-4-XAS project contends that increased standardisation of metadata through following CDIF recommendations will increase the reuse potential of XAS data outside the original experiment. Concrete use cases need to be identified to demonstrate that this is indeed the case.

In the interactive workshop, we welcome interested participants to join the discussion, exchange ideas, and help identify additional use cases for CDIF, to foster a broader community adoption for enhancing data interoperability across domains.

Corresponding Author: Dr Simon Hodson, simon@codata.org

www.helmholtz-metadaten.de