Helmholtz Metadata Collaboration | Conference 2023

Report of Contributions

https://events.hifis.net/e/HMC-Conference_2023

Contribution ID: **1** Type: **Talk**

Project MEMAS: ontology-based database system for manufacturing and simulation data in the field of composite materials

Wednesday 11 October 2023 11:30 (20 minutes)

Simulation of aerospace or automotive structures can be ultimately improved by reflecting the actual manufacturing status of the produced parts in detail. This is especially the case for composite structures in view of the complexity of the involved manufacturing processes and their influence on the product reliability. High-fidelity numerical models have to be developed to reflect the actual state of the produced structures and cover their load-bearing capability individually. The more accurate the simulation can describe the actual product, the more it is accepted as a mean to allow the structure certification through Certification by Analysis (CbA) with reduced effort for independent testing. This methodology promises cost reduction and time saving in the product certification programs for aeronautic and automotive structures. To this goal, specific tools should be developed to manage the complex datasets, multiple data structures and data formats produced along the part manufacturing and establish a link to simulation models.

In the recent years, the DLR worked on the development of an integrated data management system (IDMS) called shepard for the storage of research data according to the FAIR principles (Findable, Accessible, Interoperable, and Re-usable). In its first phase, the project MEMAS aims at developing an ontology for the labelling of manufacturing, testing and simulation data to structure and bridge these different fields in composite engineering. The coupling of the IDMS to a multi-field ontology should enable the creation of high-quality and well-documented datasets, which can be converted into predictive numerical models. This presentation will cover the first project phase and present the ontology development and its coupling to the IDMS at DLR.

Please assign your contribution to one of the following topics

Metadata annotation and management close to the research process

Please specify ”other” (stakeholder)

In addition please add keywords.

ontology, manufacturing, simulation, composite parts, database management system

Please assign yourself (presenting author) to one of the stakeholders.

Researchers

Primary authors: VINOT, Mathieu (German Aerospace Center); Mr UNGER, Nicolas (German Aerospace Center); Mr KAMBLE, Pradnil (German Aerospace Center); Dr GLÜCK, Roland (German

Aerospace Center)

Co-author: Dr TOSO, Nathalie (German Aerospace Center)

Presenter: VINOT, Mathieu (German Aerospace Center)

Session Classification: Parallel Track 1

Track Classification: Facilitating connectivity of research data: Metadata annotation and management during and close to the research process

Contribution ID: **2** Type: **Talk**

HARMONise –Enhancing the interoperability of marine biomolecular (meta)data across Helmholtz Centres

Wednesday 11 October 2023 10:50 (20 minutes)

Biomolecules, such as DNA and RNA, provide a wealth of information about the distribution and function of marine organisms, and biomolecular research in the marine realm is pursued across several Helmholtz Centers. Biomolecular metadata, i.e. DNA and RNA sequences and all steps involved in their creation, exhibit great internal diversity and complexity. However, high-quality (meta)data management is not yet well developed and harmonized in environmentally focused Helmholtz Centers. As part of the HMC Project HARMONise, we develop sustainable solutions and digital cultures to enable high-quality, standards-compliant curation and management of marine biomolecular metadata at AWI and GEOMAR to better embed biomolecular science into broader digital ecosystems and research domains. Our approach builds on a relational database that aligns metadata with community standards such as the MIxS (Minimum Information about any (x) sequence) supported by the International Nucleotide Sequence Database Collaboration (INSDC) to promote global interoperability. At the same time, we ensure the harmonization of metadata with existing Helmholtz repositories (e.g. PANGAEA). A web-based hub will enable the standardized export and exchange of core metadata. Alignment with domain-specific standards and relevant data exchange formats (e.g. UNESCO ODIS-Arch specifications) supports connectivity to the Helmholtz knowledge graph as well as global interoperability. Here we will highlight the use of standards and fields in the database scheme that promote interoperability, outline the establishment of a web-based exchange hub for sharing and validating biomolecular metadata across Helmholtz Centers, and present links with high-level international programs such as the Ocean Biomolecular Observing Network (OBON) of the UN Decade of Ocean Science. Enabling sustainable data stewardship, export and publication routines will support researchers in delivering Helmholtz biomolecular data to national European and global repositories in alignment with community standards and the FAIR principles.

Please assign your contribution to one of the following topics

Data interoperability through harmonised metadata and interoperable semantics

Please specify ”other” (stakeholder)

In addition please add keywords.

sequence data management, interoperability, metadata harmonization, FAIR principles

Please assign yourself (presenting author) to one of the stakeholders.

Scientists and technicians who maintain and operate research infrastructure for data generation

Primary author: BIENHOLD, Christina (AWI Helmholtz Centre for Polar and Marine Research)

Co-authors: SIEBERT, Isabell; HARMS, Lars; KOPPE, Roland (AWI); NEUHAUS, Stefan; BAYER, Till (GEOMAR)

Presenter: BIENHOLD, Christina (AWI Helmholtz Centre for Polar and Marine Research)

Session Classification: Parallel Track 1

Track Classification: Facilitating connectivity of research data: Data interoperability through harmonised metadata and interoperable semantics

Contribution ID: **3** Type: **Talk**

Achieving data interoperability through harmonized metadata for joint data analysis: Lessons learnt from ENPADASI, INTIMIC-KP and NFDI4Health

Wednesday 11 October 2023 11:10 (20 minutes)

Joint data analyses may overcome challenges of traditional literature-based meta-analysis owing to the use of harmonized exposure and outcome definitions as well as statistical modelling. It also allows to re-use existing data for other research purposes in more flexible ways to increase scientific impact. Achieving FAIR (meta)data by generating interoperable, harmonized, high quality metadata and data harmonization is mandatory for joint data analysis. DataSHIELD is a software tool allowing remote federated analysis of harmonized datasets across studies without physically sharing individual-level data, thereby substantially reducing burdens and challenges for data sharing that are especially common in ongoing observational studies.

As part of the European Nutritional Phenotype Assessment and Data Sharing Initiative (ENPADASI), and the Knowledge Platform Intestinal Microbiomics (INTIMIC-KP) within the Joint Programming Initiative a Healthy Diet for a Healthy Life (JPI-HDHL) as well as the National Research Infrastructure for Personal Health Data (NFDI4Health) we are collecting and harmonizing metadata on observational studies in the field of nutrition, biomarkers, omics, and chronic diseases. We also established a searchable Mica database to make harmonized metadata publicly available (https://mica.mdc-berlin.de). Based on study-level metadata, studies eligible for federated DataSHIELD analyses of multiple study data can be identified and consent for participation in a federated analysis can be requested from the principal investigators. We provide standard operating procedures (SOPs) for Opal/DataSHIELD infrastructure installation, data upload and setting permissions for eligible studies. Alternatively, harmonized datasets can be hosted at the MDC Opal database. We also provide SOPs for semi-automated data harmonization using the R package harmonizR developed by Maelstrom Research.

Federated analysis of studies is performed centrally at the MDC with DataSHIELD. Currently, we are extending this work as part of NFDI4Health to provide a central access point at the MDC for interested researchers to conduct federated data analysis of multiple epidemiological studies. In addition, the handling of credentials will be optimized by central access solutions (keycloak) in NFDI4Health. Our experiences show that harmonized and searchable study-level metadata are useful for identifying eligible studies for a federated analysis of a specific research question, yielding successful publication. Lessons learnt are also that data harmonization as well as the set-up of Opal and DataSHIELD needs time and resources at the study-level. With these lessons learnt we are contributing to shaping the research infrastructures being built by NFDI4Health.

Please assign your contribution to one of the following topics

Data interoperability through harmonised metadata and interoperable semantics

Please specify ”other” (stakeholder)

In addition please add keywords.

metadata; data harmonization; data re-use; FAIR principles; DataSHIELD; federated analysis

Please assign yourself (presenting author) to one of the stakeholders.

Researchers

Primary author: NIMPTSCH, Katharina

Co-authors: Dr SCHWEDHELM, Carolina; SIAMPANI, Sofia M.; Dr PINART, Mariona; Prof. PISCHON, Tobias

Presenter: NIMPTSCH, Katharina

Session Classification: Parallel Track 1

Track Classification: Facilitating connectivity of research data: Data interoperability through harmonised metadata and interoperable semantics

Contribution ID: **4** Type: **Poster**

HMC Dashboard on Open and FAIR Data in Helmholtz

Tuesday 10 October 2023 14:30 (15 minutes)

The Helmholtz Metadata Collaboration (HMC) has developed the HMC dashboard on Open and FAIR Data in Helmholtz. The dashboard allows users to monitor and interactively analyze statistics on open and FAIR data produced by researchers in the Helmholtz Association. It can be used to analyze in which repositories Helmholtz researchers make their data publicly available, to monitor progression over time and to understand how you can improve the FAIRness of your data. The dashboard is publicly available at https://fairdashboard.helmholtz-metadaten.de.

Please assign your contribution to one of the following topics

Research data in the FAIR data landscape

Please specify ”other” (stakeholder)

In addition please add keywords.

FAIR data, Open data, Dashboard, HMC, Helmholtz

Please assign yourself (presenting author) to one of the stakeholders.

Data professionals and stewards

Primary authors: Mr SEDEQI, Mojeeb R. (HZB, HMC); Mr SCHMIDT, Alexander (HZB); Ms SERVE, Vivien (HZB, HMC); Ms GILEIN, Astrid (HZB); Ms GLODOWSKI, Tempest (HZB); Mr PREUSS, Gabriel (HZB, HMC); Ms MANNIX, Oonagh (HZB, HMC); Mr KUBIN, Markus (HZB, HMC)

Presenters: Mr SEDEQI, Mojeeb R. (HZB, HMC); Mr KUBIN, Markus (HZB, HMC)

Session Classification: Poster session

Track Classification: Poster session

Contribution ID: **5** Type: **Hands-on session**

HMC Dashboard on Open and FAIR Data in Helmholtz

Wednesday 11 October 2023 14:12 (3 minutes)

The Helmholtz Metadata Collaboration (HMC) has developed the HMC dashboard on Open and FAIR Data in Helmholtz. The dashboard allows users to monitor and interactively analyze statistics on open and FAIR data produced by researchers in the Helmholtz Association. It can be used to analyze in which repositories Helmholtz researchers make their data publicly available, to monitor progression over time and to understand how you can improve the FAIRness of your data. The dashboard is publicly available at https://fairdashboard.helmholtz-metadaten.de.

Please assign your contribution to one of the following topics

Research data in the FAIR data landscape

Please specify ”other” (stakeholder)

In addition please add keywords.

FAIR data, open data, dashboard, HMC, Helmholtz

Please assign yourself (presenting author) to one of the stakeholders.

Data professionals and stewards

Primary authors: Mr KUBIN, Markus (HZB, HMC); Mr SEDEQI, Mojeeb R. (HZB, HMC); Mr SCHMIDT, Alexander M. (HZB, HMC); Ms SERVE, Vivien (HZB, HMC); Ms GILEIN, Astrid (HZB); Ms GLODOWSKI, Tempest (HZB); Mr PREUSS, Gabriel (HZB, HMC); Ms MANNIX, Oonagh (HZB, HMC)

Presenter: Mr KUBIN, Markus (HZB, HMC)

Session Classification: HMC Hands-on Session

Contribution ID: **6** Type: **Talk**

Towards metadata descriptors to support interoperability and data reuse in circadian and sleep science

Tuesday 10 October 2023 10:50 (20 minutes)

ABSTRACT

Introduction
Meaningful metadata are essential for the description, retrieval and reuse of data, in particular in multi-centric cooperative research projects. For the Integrative Human Circadian Daylight Platform (iHCDP, https://ihcdp.org/), a collaborative, transnational project between the University of Basel (Switzerland), the Technical University of Munich and the Max Planck Institute for Biological Cybernetics in Tübingen (Germany), we are developing the Circadian Data Hub (CDH). The CDH enables data upload and exchange to support interactivity and reuse of study data within each iHCDP module and across the platform. Expressive metadata are needed to describe the data stored in the CDH. The aim was to identify the expressive metadata needed to describe the data stored in the CDH.

Methods
The CDH is designed according to the FAIR principles formulated by Wilkinson et al. 1 to make data findable, accessible, interoperable and reusable. In several workshop meetings we have harmonized aspects of the data collection effort across institutions.
To create a catalog of metadata, we first integrated information from all data which are currently being collected within the iHCDP project teams. For this purpose, a spreadsheet for data entry was completed, where study modality, variable names, units, devices, sampling methods and frequencies were entered. We then clustered the same and similar information together. From these metadata clusters an initial set of variables for the CDH metadata descriptor was created. Next, we developed a pilot data collection tool in the openBIS system [2]. When uploading data to the CDH, this data collection tool collects information about the data collected in a particular project.

Results
We created a pilot project in the openBIS system. To build the project structure and to consider the specific data sets and variables in circadian and sleep studies we created objects for projects and for participants. For each of these objects we implemented a metadata descriptor with general and project specific metadata. General metadata for projects include: project ID, project title and total number of participants. Participant metadata are participant ID, associated project ID, age, sex and gender. As project-specific metadata, we implemented the variables we created from the metadata clusters, as Boolean variables in the metadata descriptor for projects. We grouped the collected information according to how the data was collected and into metadata clusters. For example, one cluster is: data collected by study personnel - physical/vital signs. In this cluster, we have the following variables: core body temperature, blood pressure, heart rate, weight and height.

Discussion
With the implemented metadata clusters and the variables, projects in the CDH can be described and other researchers could reuse the data in the future. The metadata descriptor can be revised at later times to respond to the addition of novel data collection modalities.

Conclusion
We developed a novel metadata descriptor, which captures the data collected within a specific project for use within the CDH.

Funding
VELUX Stiftung, Switzerland (Proj. No. 1636)

References
[References are in the Comments field.]

Please assign your contribution to one of the following topics

Data interoperability through harmonised metadata and interoperable semantics

Please specify ”other” (stakeholder)

In addition please add keywords.

Metadata, FAIR principles, Harmonization, Sleep science, Circadian studies

Please assign yourself (presenting author) to one of the stakeholders.

Researchers

Primary author: Mrs LINDOERFER, Doris (Chronobiology & Health, TUM Department of Sport and Health Sciences (TUM SG), Technical University of Munich, Munich, Germany)

Co-authors: Dr GARBAZZA, Corrado (Centre for Chronobiology, Psychiatric Hospital of the University of Basel, Basel, Switzerland); Prof. CAJOCHEN, Christian (Centre for Chronobiology, Psychiatric Hospital of the University of Basel, Basel, Switzerland); Dr MÜNCH, Mirjam (Centre for Chronobiology, Psychiatric Hospital of the University of Basel, Basel, Switzerland); Prof. SPITSCHAN, Manuel (Chronobiology & Health, TUM Department of Sport and Health Sciences (TUM SG), Technical University of Munich, Munich, Germany)

Presenter: Mrs LINDOERFER, Doris (Chronobiology & Health, TUM Department of Sport and Health Sciences (TUM SG), Technical University of Munich, Munich, Germany)

Session Classification: Parallel Track 2

Track Classification: Facilitating connectivity of research data: Data interoperability through harmonised metadata and interoperable semantics

Contribution ID: **7** Type: **Talk**

Wearable light logger and dosimetry data: Harmonizing a heterogenous field and enabling novel research in the MeLiDos project

Tuesday 10 October 2023 11:30 (20 minutes)

Personalized light exposure data is progressively gaining importance in various sectors, including research, occupational affairs, and fitness tracking. Data are collected through a proliferating selection of wearable loggers and dosimeters, varying in size, shape, functionality, and output format. Despite or maybe because of numerous use cases, the field lacks a unified framework for collecting, validating, and analyzing the accumulated data. This issue increases the time and expertise necessary to handle such data and also compromises the FAIRness (Findability, Accessibility, Interoperability, Reusability) of the results, especially in meta-analyses.
MeLiDos is a joint, EURAMET-funded project involving sixteen partners across Europe, aimed at developing a metrology and a standard workflow for wearable light logger data and optical radiation dosimeters. Its primary contributions towards fostering FAIR data include the development of a common file format, robust metadata descriptors, and an accompanying open-source software ecosystem. The software ecosystem will encompass tools for:

- Generation of data and metadata files
- Conversion of popular file formats
- Validation of light logging data
- Verification of crucial metadata
- Calculation of common parameters
- Semi-automated analysis and visualization (both command-line and GUI-based)
- Integration of data into a unified database for cross-study analyses

This presentation will provide a concise overview of light logging and dosimetry, including its inherent complexity concerning the produced data and the current fragmented approaches to managing this complexity. It will also introduce the MeLiDos project. The core of the talk will concentrate on presenting a proposed metadata descriptor for personalized light exposure data, which will be implemented as a JSON Schema encapsulating all aspects at the study, participant (wearer), dataset, and device levels. The discussion will conclude with a forecast of the project timeline and the integration of the metadata descriptor within the broader software ecosystem.

Please assign your contribution to one of the following topics

Data interoperability through harmonised metadata and interoperable semantics

Please specify ”other” (stakeholder)

In addition please add keywords.

light logging; dosimetry; JSON Schema; metadata descriptor

Please assign yourself (presenting author) to one of the stakeholders.

Researchers

Primary authors: Dr ZAUNER, Johannes (Technical University of Munich); Dr SPITSCHAN, Manuel (Technical University of Munich)

Presenter: Dr ZAUNER, Johannes (Technical University of Munich)

Session Classification: Parallel Track 1

Track Classification: Facilitating connectivity of research data: Data interoperability through harmonised metadata and interoperable semantics

Contribution ID: **8** Type: **Poster**

Discovery and Access of data in EOC Geoservice using STAC

Tuesday 10 October 2023 14:30 (15 minutes)

The management of spatial data is facing ever greater challenges. In addition to the high number of data and products, technical aspects such as data size and efficient workflows play an increasingly important role for data users and providers. In addition, open access using FAIR principles is also becoming increasingly important in the field of research data management. Data should be made easier to find, accessible, more interoperable, and reusable. To meet the needs of users, we provide various services at the EOC to share the diversity of satellite data and products. In addition, our goal is to provide a platform for scientists to present their data in a modern way.
To make the data accessible to a broad public, we offer a STAC-based catalog service in addition to the established download and visualization services. It helps finding and accessing data more dynamic and efficient. As a provider, we are able to make our valuable data and products available to a wide audience without complex infrastructure or inefficient data transfer. Users can access data simultaneously without having to download entire data sets, thus avoiding longer computing times and saving storage capacity.
The STAC catalog is divided into several specifications. The STAC API provides a RESTful endpoint that enables search of STAC Items, specified in OpenAPI, following OGC's WFS 3. The STAC Catalog is a simple, flexible JSON file of links that provides a structure to organize and browse STAC Items. The collection is an extension of the STAC Catalog including additional information such as the extents, license, keywords or providers, that describe STAC Items that fall within the Collection. The STAC Item, which represent a single spatio-temporal asset as a GeoJSON feature plus datetime and links as a central unit. In addition, further attributes can be defined in the properties for each item.
To fetch the available collections and items, the connection to the STAC API endpoint is required. This can be done via a STAC browser or by using a Jupyter notebook. Using various Python libraries (e.g. pystac), a query can be started and data can be loaded into a xarray-dataset (data cube). The data is made available to the user so that he can visualize the data or analyze it further with the right tool.

Please assign your contribution to one of the following topics

Technological solutions for findable and machine-readable metadata

Please specify "other" (stakeholder)

In addition please add keywords.

geospatial data, stac, fair, geoservice, eoc

Please assign yourself (presenting author) to one of the stakeholders.

Data professionals who provide and maintain data infrastructure

Primary authors: FECKLER, Felix; HAUG, Jan-Karl

Presenter: HAUG, Jan-Karl

Session Classification: Poster session

Track Classification: Poster session

Contribution ID: **9** Type: **Poster**

Leveraging FAIR Data practices through the HMC Information Portal

Tuesday 10 October 2023 13:35 (15 minutes)

Since its establishment the Helmholtz Metadata Collaboration (HMC) has compiled a multitude of information on the state of the research data communities and practices within the Helmholtz Association and beyond.

The Information Portal is a web application for capturing FAIR data practices across all Helmholtz domains, offering a unified user interface for collecting and exploring results.
The development was initiated to structure this data, provide accessible information for multi-level decision-making, and create a curated knowledge base for research data managers, scientists, and other stakeholders. It thereby provides a map of the Helmholtz data landscape across all Helmholtz research domains.

Developed through a top-down approach, 18 categories, and associated metadata schemas were defined and aligned throughout the association. Collected data is curated based on these cross-domain aligned metadata schemas these cross-domain aligned metadata schemas.

Built using state-of-the-art technologies, including Python, JavaScript, and Docker, the Information Portal leverages GitLab as a database. Git-based systems offer advantages, such as raw data accessibility, flexible data curation, easy synchronization, and customizable repositories. The Information Portal offers public read-only access for stakeholders and a personal instance for data curation purposes. Both versions are synchronized via GitLab.

The single-page web application is user-friendly and developed in multiple iterations for an intuitive and flexible interface. The Information Portal is important for creating a sustainable, distributed, semantically enriched Helmholtz data space, promoting seamless data sharing and reuse.

Please assign your contribution to one of the following topics

Resource in the FAIR data landscape

Please specify ”other” (stakeholder)

Please assign yourself (presenting author) to one of the stakeholders.

Expert panels, strategists and administrative stakeholders

In addition please add keywords.

Helmholtz Metadata Collaboration, Information Portal, FAIR, Metadata

Primary authors: PATIL, Akhil Jayant (DLR); LEMSTER, Christine (Geomar); SÖDING, Emanuel

(GEOMAR); BRÖDER, Jens (Forschungszentrum Jülich GmbH (IAS-9)); STUCKY, Karl-Uwe (KIT); WALTER, Konstantin Pascal (HZB (Hub-Matter)); STEINMEIER, Leon (Helmholtz Institute Freiberg); KULLA, Lucas (DKFZ); KUBIN, Markus (Helmholtz-Zentrum Berlin für Materialien und Energie, Helmholtz Metadata Collaboration); ARNDT, Witold (DLR)

Presenter: KULLA, Lucas (DKFZ)

Session Classification: Poster session

Track Classification: Poster session

Contribution ID: **10** Type: **Hands-on session**

Knowledge Graph Development as a Collaborative Process

Wednesday 11 October 2023 14:03 (3 minutes)

Establishing semantic data and knowledge graphs in scientific working groups is no easy feat. In most cases there is neither a user friendly tool chain nor experience with ontologies for the respective research field. But without a start, said experience can never be gained. The same is true for individuals that want to start into the field.

We thus see knowledge graph development not as a task of expert individuals that already know everything, but as a collaborative (learning) process of working groups and organisations. At the start of this process the right ontologies are not known and the individuals do not yet have experience with expressing information in knowledge graphs. Thus, a tool chain must provide basic knowledge to help newcomers to get started. It must also support the learning process and the selection of terms and ontologies, while users are already working with their own data and metadata. Additionally, the tool chain must support cooperation and lateral transfer of knowledge within organisations and working groups as well as between working groups world wide.

We therefore propose to establish a data infrastructure in every research organisation consisting of the following elements: An organisational knowledge graph, integration of (global) ID services, links to FAIR ontologies, policies, and a graph editing tool. This editing tool must support simultaneously the input of graph data, the extension of ontologies, the development of data structures, and finding and reusing existing ontologies and data structures not only from other persons inside the organisation but also from globally emerging metadata standards. While searching for a fitting term from a predefined set of ontologies, the tool would also allow for the creation of an internal term, when no fitting one is found. While trying to create a new term, fitting ones are automatically searched and proposed. The here proposed graph editing tool would provide the possibility to refactor existing data to newly selected ontologies, e.g. through replacing terms or whole structures, while keeping the original history in a git+GitLab like structure. This would also allow for access control and cooperation within the organisation and beyond. Such refactoring translations would also be described in terms of graph data and be published, so that others considering the same transition could use them without much effort.

We think that in the presented infrastructure users could establish processes that would foster harmonization and convergence of ontologies and data structures, while not impeding the collection of data and learning processes of individuals before harmonization is achieved.

Please assign your contribution to one of the following topics

Infrastructure and common practices for consolidating (meta)data

Please specify ”other” (stakeholder)

In addition please add keywords.

knowledge graph
collaborative data development

metadata harmonization
research data interoperability

Please assign yourself (presenting author) to one of the stakeholders.

Data professionals and stewards

Primary authors: Prof. VAN DEN BOOGAART, Karl Gerald (Helmholtz Institute Freiberg); STEINMEIER, Leon (Helmholtz Institute Freiberg); SCHALLER, Theresa

Presenter: STEINMEIER, Leon (Helmholtz Institute Freiberg)

Session Classification: HMC Hands-on Session

Contribution ID: **11** Type: **Poster**

Helmholtz Open Science Policy: Implementing the UNESCO Recommendation on Open Science

Tuesday 10 October 2023 14:30 (15 minutes)

Open science promotes innovation, improves the transfer of knowledge to society and the economy, and ensures quality and transparency in research. The Helmholtz Association, Germany's largest research performing organization, has thus adopted an Open Science Policy in September 2022 1.

This policy supports openness as a central endeavor of science and makes open science the standard for scholarly publications, research data, and research software in Helmholtz. Essentially, the policy takes up the "UNESCO Recommendation on Open Science"[2] as well as the EU Commission's open science funding policy in the current Horizon Europe research framework program. Furthermore, the policy is guided by the principles of transparency, quality assurance, and sustainability.

Currently, already 82 percent of Helmholtz scholarly articles are accessible via open access. Open research data are published; the Helmholtz Centers e.g., participate in the European Open Science Cloud (EOSC) and National Research Data Infrastructure (NFDI). Central, too, is the open publication of research software to promote the reproducibility of scientific results.

Our contribution presents a comprehensive and practically informed perspective on how to implement open science, working together with numerous stakeholders within research performing organizations such as Helmholtz to guide the cultural change towards openness.

This process is not only inspired and supported by the UNESCO Recommendation on Open Science; also, the cultural change towards openness is enabled in a multifaceted and global manner. In this contribution and the ensuing discussion, there will thus be ample opportunity to discuss open science and its implementation; the target audience are researchers, administrators, and infrastructure professionals, as well as all further interested persons.

References
1 Helmholtz Association (2022): Helmholtz Open Science Policy.
[2] UNESCO (2021): UNESCO Recommendation on Open Science. https://unesdoc.unesco.org/ark:/48223/pf0000379949.loca

Please assign your contribution to one of the following topics

Bringing recommendations closer to practice

Please specify "other" (stakeholder)

In addition please add keywords.

Helmholtz Open Science Policy, Implementation, UNESCO

Please assign yourself (presenting author) to one of the stakeholders.

Expert panels, strategists and administrative stakeholders

Primary authors: FERGUSON, Lea Maria (Helmholtz Association, Helmholtz Open Science Office); MEISTRING, Marcel (Helmholtz Association, Helmholtz Open Science Office); Dr PAMPEL, Heinz (Helmholtz Association, Helmholtz Open Science Office); WEISWEILER, Nina Leonie (Helmholtz Open Science Office); BERTELMANN, Roland

Co-authors: Dr SCHULTZE-MOTEL, Paul (Helmholtz Open Science Office); SCHRADER, Antonia (Helmholtz Open Science Office); MESSERSCHMIDT, Lena; Dr GENDERJAHN, Steffi (Helmholtz Open Science Office); BRUCH, Christoph (Helmholtz Open Science Office)

Presenter: WEISWEILER, Nina Leonie (Helmholtz Open Science Office)

Session Classification: Poster session

Track Classification: Poster session

Contribution ID: **12** Type: **Hands-on session**

FAIR Data Management Workflow for MRI Data

Wednesday 11 October 2023 14:09 (3 minutes)

We present a workflow to improve the management of Magnetic Resonance Imaging data and to increase its compliance with the FAIR principles. This involves using the JSON Metadata Mapping Tool we have developed to map metadata from a domain-specific file format to a JSON schema based format, and storing the data and the mapped metadata in repositories. Some steps in the workflow are automated, while others require human intervention, facilitated by Graphical User Interfaces for each service. We assessed the compliance of our curated data to the FAIR principles, both manually and using the F-UJI tool. We obtain a FAIR assessment score of 79% for both datasets, which is the highest compared to similar ones in the same field. According to these results, we conclude that the workflow we have implemented can provide a significant improvement towards FAIR data management.

Please assign your contribution to one of the following topics

Data interoperability through harmonised metadata and interoperable semantics

Please specify ”other” (stakeholder)

In addition please add keywords.

Metadata Mapping, FAIR Assessment, Materials Science

Please assign yourself (presenting author) to one of the stakeholders.

Researchers

Primary authors: BLUMENROEHR, Nicolas (Karlsruhe Institute of Technology, Steinbuch Centre for Computing); Dr AVERSA, Rossella (KIT)

Co-author: Dr MACKINNON, Neil (KIT)

Presenter: BLUMENROEHR, Nicolas (Karlsruhe Institute of Technology, Steinbuch Centre for Computing)

Session Classification: HMC Hands-on Session

Contribution ID: **13** Type: **Poster**

Human-Centric Understanding of the FAIR Data Landscape: HMC Community Survey

Tuesday 10 October 2023 14:20 (15 minutes)

In 2021 HMC conducted its first community survey to align its services with the needs of Helmholtz researchers. A question catalogue, with 49 (sub-)questions based on an expertise-adaptive approach, was designed and disseminated among researchers in all six Helmholtz research fields. 631 completed survey replies were obtained for analysis.

The HMC Community Survey 2021 provides insight into the management of research data as well as the data publication practices of researchers in the Helmholtz Association. The characterization of research-field-dependent communities will enable HMC to further develop targeted, community-driven support for the documentation of research data with metadata.

An outlook to our future survey strategies will be discussed.

Please assign your contribution to one of the following topics

Understanding people in the FAIR data landscape

Please specify ”other” (stakeholder)

In addition please add keywords.

HMC; Survey; Community; FAIR data practices

Please assign yourself (presenting author) to one of the stakeholders.

Data professionals and stewards

Primary authors: GERLICH, Silke (HMC); KUBIN, Markus (HMC, HZB); KULLA, Lucas (DKFZ); LEMSTER, Christine (Geomar); SCHWEIKERT, Jan (KIT); SHANKAR, Sangeetha (German Aerospace Center)

Co-author: SÖDING, Emanuel (GEOMAR)

Presenters: KUBIN, Markus (HMC, HZB); LEMSTER, Christine (Geomar); SHANKAR, Sangeetha (German Aerospace Center)

Session Classification: Poster session

Track Classification: Poster session

Contribution ID: **14** Type: **Poster**

Developing (semi)automatic analysis pipelines and technological solutions for metadata annotation and management in high-content screening (HCS) bioimaging

Tuesday 10 October 2023 13:20 (15 minutes)

Bioimaging is an important methodological procedure widely applied in life sciences. Bioimaging unites the power of microscopy, biology, biophysics and advanced computational methods allowing scientists to study different biological functions at the level of the single molecules and up to the complete organism. In parallel, high-content screening (HCS) bioimaging approaches are powerful techniques consisting of the automated imaging and analysis of large numbers of biological samples, to extract quantitative and qualitative information from the images. HCS bioimaging plays a crucial role in advancing our understanding of cellular processes, disease mechanisms, and drug development by enabling the rapid analysis of large-scale biological data.

However, HCS still presents several bottlenecks restraining these approaches from exerting their full potential for scientific discoveries. As major example, a huge amount of metadata is generated in each experiment, capturing critical information about the images. The efficient and accurate treatment of image metadata is of great importance, as it provides insights that are essential for effective image management, search, organisation, interpretation, and sharing. It is vital to find ways to properly deal with the huge amount of complex and unstructured data for implementing Findable, Accessible, Interoperable and Reusable (FAIR) concepts in bioimaging.

In the frame of NFDI4BioImaging (the National Research Data Infrastructure focusing on bioimaging in Germany), we want to find viable solutions for storing, processing, analysing, and sharing HCS data. In particular, we want to develop solutions to make findable and machine-readable metadata using (semi)automatic analysis pipelines. In scientific research, such pipelines are crucial for maintaining data integrity, supporting reproducibility, and enabling interdisciplinary collaboration. These tools can be used by different users to retrieve images based on specific attributes as well as support quality control by identifying appropriate metadata.

In the present study, we proposed an automated analysis pipeline for storing, processing, analysing, and sharing HCE bioimaging data. The (semi)automatic workflow was developed by taking as a case study a dataset of zebrafish larvae images previously obtained from an automated imaging system generating data in an HCS fashion. In our workflows, zebrafish images are automatically enriched with metadata (i.e. key-value pairs, tags, raw data, regions of interest) and uploaded to the UFZ-OME Remote Objects (OMERO) server using python scripts embedded in workflows developed with KNIME or GALAXY. The workflows give the possibility to the user to intuitively fetch images from the local server and perform image analysis (i.e. annotation) or even more complex toxicological analyses (dose response modelling). Furthermore, we want to improve the FAIRness of the protocol by adding a direct upload link to the Image Data Resource (IDR) repository to automatically prepare the data for publication and sharing.

Please assign your contribution to one of the following topics

Metadata annotation and management close to the research process

Please specify ”other” (stakeholder)

In addition please add keywords.

bioimaging, NFDI, zebrafish, KNIME, GALAXY, OMERO, automatic workflows

Please assign yourself (presenting author) to one of the stakeholders.

Researchers

Primary author: MASSEI, Riccardo

Co-authors: BUMBERGER, Jan; Mr BOHRING, Hannes (Helmholtz Centre for Environmental Research - UFZ); Mr SCHNICKE, Thomas (Helmholtz Centre for Environmental Research - UFZ); Dr BUSCH, Wibke (Helmholtz Centre for Environmental Research - UFZ); Dr SCHOLZ, Stefan (Helmholtz Centre for Environmental Research - UFZ)

Presenter: MASSEI, Riccardo

Session Classification: Poster session

Track Classification: Poster session

Contribution ID: **15** Type: **Poster**

Community Building for Research Data Repositories in Helmholtz

Tuesday 10 October 2023 14:05 (15 minutes)

The Helmholtz Metadata Collaboration (HMC) and the Helmholtz Open Science Office have launched a joint initiative at the end of 2022 to strengthen and connect research data repositories in the Helmholtz Association, and to increase their visibility in the international research landscape. Research data repositories form central hubs for metadata on the Road to FAIR: They generate, consolidate and maintain metadata, thus ensuring that valuable data generated in the course of research projects can be systematically reused over the long term.

The Helmholtz Open Science Office has been involved in re3data, the "Registry of Research Data Repositories", for over 10 years. The multi-disciplinary registry indexes repositories for searching and publishing research data and is the most comprehensive directory of research data repositories worldwide. Currently, about 100 repositories listed in re3data can be assigned to Helmholtz Centers.

Based on the information in re3data and in close collaboration with the domain-specific HMC metadata hubs, the joint initiative will map the research data repositories in Helmholtz and aim to build a networked community to further develop these infrastructures in Helmholtz, also taking into account a successful implementation of the FAIR principles.

The initiative makes use of existing synergies by linking the activities of re3data, HMC and the Helmholtz Open Science Office in a meaningful way. The results achieved will be systematically documented to support HMC, the Working Group Open Science, and other bodies in Helmholtz in practical and strategic activities in the field of research data infrastructures.

Please assign your contribution to one of the following topics

Infrastructure and common practices for consolidating (meta)data

Please specify ”other” (stakeholder)

In addition please add keywords.

Research Data Repositories, Helmholtz Association, Community Building, Metadata, Standards, Infrastructures

Please assign yourself (presenting author) to one of the stakeholders.

Expert panels, strategists and administrative stakeholders

Primary authors: WEISWEILER, Nina Leonie (Helmholtz Open Science Office); BERTELMANN, Roland; CURDT, Constanze (Helmholtz Metadata Collaboration (HMC), GEOMAR Helmholtz Centre

for Ocean Research Kiel)

Presenter: WEISWEILER, Nina Leonie (Helmholtz Open Science Office)

Session Classification: Poster session

Track Classification: Poster session

Contribution ID: **16** Type: **Poster**

Publication cost transparency within the Helmholtz Community and beyond: openCost-metadata for different publication types –from APCs up to research data.

Tuesday 10 October 2023 13:50 (15 minutes)

The openCost project aims to contribute to a fair reform of the scientific publishing system by establishing comprehensive cost transparency in the publishing process.

To this end, openCost creates the required technical infrastructure to freely access publication costs and exchange these data via automated, standardized interfaces and formats.
Standardized recording and open provision of publication costs associated with articles or contracts is vitally important to improve cost transparency both within institutions as well as inter-institutional and on larger levels (e.g. consortium or national contracts).

openCost proposed for easy distribution of data to use the well established and widely available OAI-PMH interface in conjunction with an XML format. This will be implemented exemplarily at the services of the project partners, the Universities of Bielefeld and Regensburg and by DESY for the shared repository infrastructure JOIN², effectively enabling six research centres and another university to provide data.

Via OAI-PMH openCost will also enable service providers (e.g. aggregators and research funding agencies) to harvest data directly from institutions. As a sample aggregator within the project's framework, this is demonstrated by the OpenAPC service (Bielefeld).
By implementing a technical interface, the services of the Electronic Journals Library EZB (Regensburg) will also be extended to include detailed information on OA publication costs, to be finally used as a central information platform for communicating OA information to researchers.

The project's view on publication costs not only include article processing charges (APCs), but all publication-related fees, such as submission or color fees, as well as costs resulting from transformation contracts or memberships.
With the help of international expertise, openCost develops a standardized, structured metadata schema for publication costs. This schema, including standardized vocabularies, identifiers or consensual definitions, will be presented.
To prove the viability of our approach it is constantly integrated in its latest version into the repositories of the project partners and definitions are cross checked with feedback from the community.
The openCost metadata schema already exists for costs based on individual scientific articles, while a similar schema for contracts is currently under development. Additionally, metadata schemas for books and book chapters are to be added.

Beyond the classical publishing outputs openCost will also tackle costs related to the publication of research data.
The publication of data is becoming increasingly important and raises many questions in the publication process, and related costs also require a transparent, standardized approach right from the start.
We assume that many of the carefully thought out criteria of the publication types discussed so far are already connectable, hence the extension of the openCost schema seems to be an immediate goal for the near future.
This extension, however, will require close collaboration with the
community early on to identify relevant parameters. By providing insight into our work at hand, highlight parallels and differences between research data and literature publications, we want to engage in this discussion.

Funded by the German Research Foundation (457354095), openCost was launched in 2022 by the project partners Bielefeld University Library, Regensburg University Library and DESY Central Library.

Please assign your contribution to one of the following topics

Infrastructure and common practices for consolidating (meta)data

Please specify ”other” (stakeholder)

Scientists in scientific libraries engaged in research infrastructure …

In addition please add keywords.

publication costs, fair publishing, transparency, openCost metadata schema

Please assign yourself (presenting author) to one of the stakeholders.

other (please specify)

Primary author: STEIN, Lisa-Marie (DESY)

Co-author: WAGNER, Alexander (DESY)

Presenter: WAGNER, Alexander (DESY)

Session Classification: Poster session

Track Classification: Poster session

Contribution ID: **17** Type: **Poster**

PATOF: From the Past To the Future: Legacy Data in Small and Medium-Scale "PUNCH"Experiments - a Blueprint for PUNCH and Other Disciplines

Tuesday 10 October 2023 14:30 (15 minutes)

The PATOF project builds on work at MAMI particle physics experiment A4. A4 produced a stream of valuable data for many years which already released scientific output of high quality and still provides a solid basis for future publications. The A4 data set consists of 100 TB and 300 million files of different types (hierarchical folder structure and file format with minimal metadata provided create vague context). Recent work with consulting support from the HMC hub "Matter"helped to identify problems and potential solutions for a FAIRification of A4 data. We would like to go beyond and build a FAIR Metadata Factory that can be used across research fields. The first focus will be on creating machine-readable XML files containing metadata from the logbook and other sources and to further enrich them, other challenges will be an automatised treatment of personalised logbook information.

In this project, we intend to conclude the work on A4 data, to extract the lessons learned there in the form of a cookbook, and to apply them to four other experiments: The ALPS II axion and dark matter search experiment at DESY is expected to collect 1 TB of data per week. The PRIMA experiment at MAMI in Mainz for measuring the pion transition form factor is taking data of 3 TB per week in 2023. The upcoming nuclear physics experiment P2 at MESA in Mainz is expected to collect 3 TB of data per week. These are real data mixed with calibration data and polarimetry data. Finally, the LUXE experiment at DESY planned to start in 2026 and will collect 1.5 PB of data per year.

The focus of PATOF is on making the data of A4 (and ALPS II, PRIMA, P2, and LUXE) fully publicly available. We refer to these four future experiments jointly as "APPLe". In order to achieve this, a "metadata factory" will be implemented, the concept as follows:

- DESY library, provide a "cookbook"capturing the methodology for making individual experiment-specific metadata schemas FAIR and describing a "FAIR Metadata Factory", i.e. a process to create a naturally evolved metadata schema by extending the DataCite schema without discarding the original metadata concepts.

We first consult the domain experts from the concrete experiments (e.g., what data must be in the metadata) and design the metadata schema which partially follows the DataCite metadata schema as the core of it, plus experiment-specific metadata fields. Based on the consultation and experience that we have, we cross-reference the metadata of different experiments to find out the best strategies for automatically developing metadata schemas that can be used for different experiments, and even newly developing experiments.

The objectives of the project are i) a FAIR Metadata Factory (i.e. a cookbook of (meta)data management recommendations), and ii) the FAIRification of data from concrete experiments. Both aspects are inherently open in nature so that everybody can profit from PATOF results. The cookbook is expected to be further enhanced with contributions from other experiments even after PATOF ("living cookbook").

Please assign your contribution to one of the following topics

Metadata annotation and management close to the research process

Please specify ”other” (stakeholder)

In addition please add keywords.

Metadata, Scientific Data Management, FAIR

Please assign yourself (presenting author) to one of the stakeholders.

Data professionals and stewards

Primary author: HU, Ding-Ze (Deutsches Elektronen-Synchrotron DESY)

Co-authors: Dr ENKE, Harry (Leibniz-Institut für Astrophysik Potsdam (AIP)); STEIN, Lisa-Marie (DESY); Dr KÖHLER, Martin (Deutsches Elektronen-Synchrotron DESY); Dr SCHOERNER, Thomas (Deutsches Elektronen-Synchrotron DESY)

Presenter: HU, Ding-Ze (Deutsches Elektronen-Synchrotron DESY)

Session Classification: Poster session

Track Classification: Poster session

Contribution ID: **18** Type: **Poster**

Fundamentals of scientific metadata - didactic course design and material for a hands-on training course on metadata

Tuesday 10 October 2023 14:30 (15 minutes)

In their endeavor to generate and share FAIR research data, scientist face various challenges. High-level recommendations such as the FAIR principles [^1] require prior knowledge and a set of technical skills which are typically not part of the academic education. Therefore, the successful implementation of FAIR research data guidelines stands in grave need for well-trained, data-literate and technically skilled scientific staff. As a result, the general demand for the implementation of the FAIR data principles, goes hand in hand with the demand for good educational resources that can help researchers meet those.

The HMC Community Survey 2021 [^2] revealed that more than 45 % of Helmholtz researchers have little to no prior knowledge on FAIR principles and metadata handling. However, their interest in training in this field was astoundingly high (91,7 %).
We therefore created a metadata training course covering the fundamental aspects of metadata annotation, targeted towards early-career researcher (PhD Students and Postdoctoral Researchers) from any scientific domain. The didactic concept encourages and motivates participants to start and sustainably adapt their (meta)data annotations through hands-on exercises focusing on familiar problems.

Our material is designed in a modular fashion to allow easy adaptation both to the audiences skill level or to different scientific domains.

The course was taught on 5 different occasions so far. The general interest in such training as well as the post-hoc evaluation among participants attests both high interest as well as high quality of our material.

[^1]: Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
[^2]: Arndt, W. , Gerlich, S. C. , Hofmann, V. , Kubin, M. , Kulla, L. , Lemster, C. , Mannix, O. , Rink, K. , Nolden, M. , Schweikert, J. , Shankar, S. , Söding, E. , Steinmeier, L. and Süß, W. and Helmholtz Metadata Collaboration (HMC) Working Group "Taskforce Survey" (2022) A survey on research data management practices among researchers in the Helmholtz Association. Open Access , ed. by Lorenz, S., Finke, A., Langenbach, C., Maier-Hein, K., Sandfeld, S. and Stotzka, R.. HMC Report, 1 . HMC-Office, GEOMAR Helmholtz Centre for Ocean Research, Kiel, Germany, 39 pp. DOI 10.3289/HMC_publ_05.

Please assign your contribution to one of the following topics

Enabling and incentivising the research community

Please specify "other" (stakeholder)

HMC staff

In addition please add keywords.

Training
Metadata
HMC

Please assign yourself (presenting author) to one of the stakeholders.

other (please specify)

Primary author: GERLICH, Silke (HMC)

Co-authors: Prof. SANDFELD, Stefan; HOFMANN, Volker

Presenter: GERLICH, Silke (HMC)

Session Classification: Poster session

Track Classification: Poster session

Contribution ID: **19** Type: **Talk**

A Data-driven Approach to Characterizing the Spectral, Temporal and Spatial light Variations in the real world

Tuesday 10 October 2023 11:10 (20 minutes)

Light exposure significantly impacts various aspects of human psychology and physiology, including cognition, mood and circadian rhythm. The light in the real world has sophisticated characteristics; it is spatially articulated and temporally varying, even Changing the head direction and eye movement alter it. How the visual and non-visual light-mediated brain pathways encode these spatial and temporal variations is currently unknown.
A comprehensive multimodal measurement campaign is essential to gain a better mechanistic understanding of the sophisticated light patterns people encounter, thereby creating valuable information for diverse scientific and professional disciplines, including chronobiology, environmental psychology, architectural lighting design and building science.
This project aims to develop an extensive multi-dimensional dataset to map out indoor and outdoor natural scenes' spectral, spatial and temporal properties across a wide range of geographical and seasonal contexts. We collect a setup consisting of an imaging radiometer, Spectroradiometer, colorimeter, RGB and Depth cameras, Light loggers and a temperature logger. We carry on measurements throughout a day in measurement blocks with 30 minutes intervals. The critical point to make the best advantage of this dataset is creating an in-detailed metadata file comprising the list of all measurements, related IDs, and an organized set of environmental and categorical information.
A project in the RedCap platform makes it possible to gather all the valuable information as metadata. The measurement protocol employed in this study spans an entire day, commencing from morning and extending through evening hours. This approach involves the systematic acquisition of data, with each 30-minute interval constituting a discrete block of measurement encompassing all mentioned devices. A rich metadata entry is meticulously compiled for each of these measurement blocks. This metadata repository encapsulates a diverse array of essential information, including date and time stamps, geographical coordinates represented by longitude and latitude, records of the location, categorical descriptions of the scene view, lighting conditions, and an account of the prevailing weather dynamics. This meticulous assembly of metadata not only facilitates the contextual interpretation of measurement outcomes but also augments the reliability and robustness of the acquired data, ultimately enhancing the depth and accuracy of our analytical endeavors.
In order to ensure standardized and consistent categorization for critical factors such as weather conditions, scene views, and lighting, a structured approach has been adopted. This method involves the utilization of a curated dropdown selection mechanism, which draws from a list of conditions derived from authoritative scientific literature within the relevant field. The overarching objective is to streamline and enhance data classification's homogeneity, paving the way for more robust and insightful subsequent analyses. This selection of categories aims to minimize the potential for bias and subjectivity, and the resulting dataset becomes an even more valuable resource for comprehensive and accurate interpretation.

Please assign your contribution to one of the following topics

Metadata annotation and management close to the research process

Please specify ”other” (stakeholder)

In addition please add keywords.

metadata, categorization, reproducibility, accessibility

Please assign yourself (presenting author) to one of the stakeholders.

Researchers

Primary author: TABANDEH, Niloufar (Max Planck Institute for Biological Cybernetics)

Co-author: Prof. SPITSCHAN, Manuel (Max Planck Institute for Biological Cybernetics)

Presenter: TABANDEH, Niloufar (Max Planck Institute for Biological Cybernetics)

Session Classification: Parallel Track 1

Track Classification: Facilitating connectivity of research data: Metadata annotation and management during and close to the research process

Contribution ID: **20** Type: **Poster**

The Helmholtz Digitization Ontology (HDO): Harmonized semantics for the Helmholtz digital ecosystem

Tuesday 10 October 2023 14:30 (15 minutes)

The Helmholtz digital ecosystem connects diverse scientific domains with differing (domain-specific) standards and best practices for handling metadata. Ensuring interoperability within such a system, e.g. of developed tools, offered services and circulated research data, requires a semantically harmonized, machine-actionable, and coherent understanding of the relevant concepts. Further, this needs to be aligned and harmonized with European and global initiatives to ensure an open and interoperable flow of data and information. Accordingly, the Helmholtz Metadata Collaboration develops the "Helmholtz Digitization Ontology"(HDO), which contains machine-actionable descriptions of digital assets and processes relevant to this ecosystem. Containing consistent and carefully curated semantics, it is intended to serve as an institutional reference thereby supporting the integrity of HMC developments internally as well as externally.

HDO is aligned to practices and conventions of the Open Biological and Biomedical Ontologies (OBO): we create coherent and precise definitions in the OBO recommended genus-differentia form (i.e. for each term we define a Genus as well as its differentia). Class labels and definitions are developed bilingually in both English and German. Additionally, classes have further information, including synonymy, singular, plural, gloss, comments as well as micro-credits of contributions. To ensure the sustainable development of HDO, we implemented it based on the Ontology Development Kit (ODK).

Follow the current development status and engage in discussions on our public git repository 1. Preliminary documentation of HDO is available online [2].

References

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

Please assign your contribution to one of the following topics

Resource in the FAIR data landscape

Please specify ”other” (stakeholder)

In addition please add keywords.

Ontology, OWL, OBO, Digital assets, Semantics

Please assign yourself (presenting author) to one of the stakeholders.

Researchers

Primary author: FATHALLA, Said

Co-authors: GUENTHER, Gerrit (Helmholtz-Zentrum Berlin); STEINMEIER, Leon (Helmholtz Institute Freiberg); LEMSTER, Christine (Geomar); HOFMANN, Volker; BUTTIGIEG, Pier (GEOMAR Helmholtz Centre for Ocean Research)

Presenter: FATHALLA, Said

Session Classification: Poster session

Track Classification: Poster session

Contribution ID: **21** Type: **Poster**

Automating Metadata Handling in Research Software Engineering

Tuesday 10 October 2023 14:30 (15 minutes)

Automating Metadata Handling in Research Software Engineering
Mustafa Soylu^ 1
Anton Pirogov^ 1
Volker Hofmann 1
Stefan Sandfeld 1

^ The authors contributed equally to this work

Institute for Advanced Simulation - Materials Data Science and Informatics (IAS9), Forschungszentrum Jülich, Jülich, Germany

Modern research is heavily dependent on software. The landscape of research software engineering is evolving at a high pace, and the effective handling of metadata plays a pivotal role in ensuring software discoverability, reproducibility, and general project quality. Properly curating metadata can, however, become a time-consuming task, while manual curation is error-prone at the same time. This poster introduces two new tools for streamlining metadata management: somesy and fair-python-cookiecutter.

Somesy (**so**ftware **me**tadata **sy**nchronization) provides a user-friendly command-line interface that assists in the synchronization of software project metadata. Somesy supports best-practice metadata standards such as CITATION.cff and CodeMeta and automatically maintains metadata, such as essential project information (names, versions, authors, licenses), consistently across multiple files. This ensures metadata integrity and frees additional time for developers and maintainers to focus on their work.

The fair-python-cookiecutter is a GitHub repository template which provides a structured foundation for Python projects. The template provides researchers and RSEs with support in meeting the increasing demands for software metadata during development of Python tools and libraries. By cloning and applying the template to their projects, developers can benefit from the incorporated best practices, recommendations for software development, and software project metadata to ensure quality and facilitate citation of their work. The fair-python-cookiecutter is aligned with and inspired by standards like DLR Software Engineering Guidelines, OpenSSF Best Practices, REUSE, CITATION.cff, CodeMeta. Furthermore, it uses **somesy** to enhance software metadata FAIRness. The template comes with detailed documentation and thus offers an accessible framework for achieving software quality and discoverability within academia.

https://pypi.org/project/somesy/
https://github.com/Materials-Data-Science-and-Informatics/fair-python-cookiecutter

Please assign your contribution to one of the following topics

Technological solutions for findable and machine-readable metadata

Please specify "other" (stakeholder)

HMC core staff

In addition please add keywords.

FAIR, metadata, python

Please assign yourself (presenting author) to one of the stakeholders.

other (please specify)

Primary authors: SOYLU, Mustafa (Forschungszentrum Jülich); PIROGOV, Anton (Forschungszentrum Jülich)

Co-authors: HOFMANN, Volker; SANDFELD, Stefan

Presenters: SOYLU, Mustafa (Forschungszentrum Jülich); PIROGOV, Anton (Forschungszentrum Jülich)

Session Classification: Poster session

Track Classification: Poster session

Contribution ID: **22** Type: **Poster**

Metador: A metadata-centric framework for enabling FAIR research (meta)data handling

Tuesday 10 October 2023 14:30 (15 minutes)

To be sustainable and useful, scientific data should be FAIR. These goals can only be achieved by definition and adoption of metadata standards and implementation of tools and services that support these standards. Unfortunately, the diversity of needs with respect to scientific (meta)data leads to a large gap between the scope and pace of large-scale standardization efforts and the day-to-day work of domain researchers.

This poster gives an overview of Metador - a metadata-centered framework emphasizing the I and R in FAIR. With Metador we try to address these issues using an incremental, bottom-up approach. It is based on a lightweight technical meta-standard for packaging JSON-serialized metadata objects alongside the research data in archives, a simple API implementing this standard, and a plugin-based metadata schema system. On top of this, it provides automatic generation of dashboards for compatible data archives that can be embedded in different settings.

To showcase the versatility of the framework, we recently implemented the dashboard functionality of Metador as a general-purpose InvenioRDM extension. In future work, we plan to extend this integration to enable improved metadata-driven search capabilities, powered by the dynamic Metador metadata schema system, as well as integrating use cases demonstrating how the framework can be applied to satisfy concrete metadata needs for research data handling.

Please assign your contribution to one of the following topics

Technological solutions for findable and machine-readable metadata

Please specify ”other” (stakeholder)

HMC core staff

In addition please add keywords.

FAIR, metadata, Python, framework

Please assign yourself (presenting author) to one of the stakeholders.

other (please specify)

Primary authors: PIROGOV, Anton (Forschungszentrum Jülich); D'MELLO, Fiona (Forschungszentrum Jülich); SOYLU, Mustafa (Forschungszentrum Jülich)

Co-authors: SANDFELD, Stefan; HOFMANN, Volker

Presenter: PIROGOV, Anton (Forschungszentrum Jülich)

Session Classification: Poster session

Track Classification: Poster session

Contribution ID: **23** Type: **Poster**

Using EVOKS to Build Controlled Vocabularies

Tuesday 10 October 2023 14:30 (15 minutes)

Controlled vocabularies are used to describe knowledge within a particular domain, encompassing a comprehensive collection of domain specific terms. Using controlled vocabularies not only mitigates the challenge of data ambiguity, but also offers several advantages, including references to term definitions, particularly within metadata schemas. Additionally, they foster semantic interoperability and facilitate the seamless integration of ontologies.

EVOKS, the Editor for Vocabularies to Know Semantics, is a general-purpose vocabulary service which allows data stewards and scientists to easily create or import, edit, curate and publish controlled vocabularies using the W3C recommended SKOS data model 1. Access to published vocabularies is effectively ensured through the implementation of SKOSMOS [2] as a dedicated vocabulary browser instance.

To illustrate the usability of EVOKS, we set up the NFDI-MatWerk Acronyms Vocabulary [3], consisting of 67 terms from the NFDI-MatWerk [4] proposal, in a structured data model as a practical example.

The poster shows basic usage and benefits of using EVOKS. In particular:

- Creating and editing controlled vocabularies
- Collaboratively working on vocabularies
- Assigning persistent URLs to the vocabulary and its terms.

The EVOKS interface is designed to be very intuitive and user-friendly: users can quickly get acquainted with the platform and navigate its features without significant time investment.

As an application use case, the poster also demonstrates the integration of controlled vocabularies into metadata schemas. Specifically, it illustrates how the narrower terms of a given term from a controlled vocabulary appear as selectable options in the schema's drop-down menu, when a metadata editor interface is configured.

Setting up a dedicated EVOKS instance is easily possible for HMC members who want to create controlled vocabularies that adhere to the FAIR principles. Using EVOKS, member organizations can create and publish controlled vocabularies tailored to their needs. As shown in our example use-case, integrating these vocabularies with existing applications and projects is easily possible.

This work has been supported by NFDI-MatWerk (DFG –n. 460247524), NFDI4Ing (DFG –n. 442146713), NFFA-Europe-Pilot (EU H2020 –n. 101007417), CRC 980 Episteme in Motion (DFG - n. 191249397), the Helmholtz Research Association with the research program 'Engineering Digital Futures'and the Helmholtz Metadata Collaboration (HMC) platform.

1 http://www.w3.org/TR/skos-reference
[2] https://skosmos.org/
[3] https://purls.helmholtz-metadaten.de/evoks/MatWerkAcronyms/
[4] https://nfdi-matwerk.de

Please assign your contribution to one of the following topics

Data interoperability through harmonised metadata and interoperable semantics

Please specify ”other” (stakeholder)

In addition please add keywords.

vocabulary service, data interoperability, HMC, NFDI, vocabularies, NFDI-MatWerk

Please assign yourself (presenting author) to one of the stakeholders.

Data professionals and stewards

Primary author: Mrs ABDILDINA, Gulzaure (KIT)

Co-authors: Mrs ERNST, Felix (KIT); Mrs OST, Philipp-Joachim (KIT); Dr AVERSA, Rossella (KIT)

Presenter: Mrs ABDILDINA, Gulzaure (KIT)

Session Classification: Poster session

Track Classification: Poster session

Contribution ID: **24** Type: **Talk**

A prototype platform for mapping heterogeneous metadata of the three domains health, environment and earth observation: the MetaMap^3 project

Wednesday 11 October 2023 10:10 (20 minutes)

The environment plays an increasingly important role for human health and efficient linkage with environmental and earth observation data is crucial to quantify human exposures. Currently, there are no harmonized metadata standards for automatic mapping. This project aims to facilitate the linkage of data of different research fields by generating and enriching interoperable and machine-readable metadata for exemplary data of our three domains Health (HMGU), Earth & Environment (UFZ), and Aeronautics, Space & Transport (DLR) and by mapping these metadata so that they can be jointly queried, searched and integrated into HMC.

After reviewing several standards, strategies and tools, we developed an approach to align our metadata to a common structure and format. We identified spatial and time coverage as the main mapping criteria. For the environmental metadata and the epidemiological metadata that have a spatial component (study centers, recruitment districts) we converged to the standard ISO 19115 and to the eXtensible Markup Language (XML).

Additionally for the health domain, we reviewed several metadata standards for health data but currently available standards were set up mostly for medical or clinical data. Thus, they only fit our epidemiological cohort data in a very limited way. Nevertheless, we started to standardize and enrich the metadata of the LISA birth cohort study by supplementing more than 15,000 variables with Maelstrom categories and subcategories. Of these, approximately 900 entries include an ICD-10 code.

Moreover, we compiled a decision matrix to guide the selection of a suitable application to upload and store the harmonized metadata. Building on this conceptual work, we identified a catalog (GeoNetwork) to store a subset of our cross-domain metadata and set up a test instance on a HMGU server and uploaded the first metadata.

We are currently populating the mapping platform with further metadata and testing the full functionality of the tool, especially the filtering and search tools options of the application to enable the intended mapping. By the end of the project, we plan to release the platform to HMC and other researchers working in thematically related fields.

Please assign your contribution to one of the following topics

Data interoperability through harmonised metadata and interoperable semantics

Please specify ”other” (stakeholder)

In addition please add keywords.

Health, environment, earth observation, metadata mapping, metadata catalog

Please assign yourself (presenting author) to one of the stakeholders.

Researchers

Primary authors: Dr DALLAVALLE, Marco (Helmholtz Zentrum München GmbH - German Research Center for Environmental Health, Institute of Epidemiology, Neuherberg, Germany and Chair of Epidemiology, IBE, Faculty of Medicine, LMU Munich); GEY, Ronny (Helmholtz Centre for Environmental Research –UFZ, Research Data Management (RDM), Smart models and Monitoring, Leipzig); STAAB, Jeroen (German Aerospace Center (DLR), German Remote Sensing Data Center, Geo-Risks and Civil Security, Oberpfaffenhofen, Weßling, Germany and Geography Department, Humboldt-University Berlin, Berlin, Germany); Dr STANDL, Marie (Helmholtz Zentrum München GmbH - German Research Center for Environmental Health, Institute of Epidemiology, Neuherberg, Germany); Dr BUMBERGER, Jan (Helmholtz Centre for Environmental Research –UFZ, Research Data Management (RDM), Smart models and Monitoring, Leipzig); Dr TAUBENBÖCK, Hannes (German Aerospace Center (DLR), German Remote Sensing Data Center, Geo-Risks and Civil Security, Oberpfaffenhofen, Weßling, Germany and Institute for Geography and Geology, Julius-Maximilians-Universität Würzburg, Würzburg, Germany); Dr WOLF, Kathrin (Helmholtz Zentrum München GmbH - German Research Center for Environmental Health, Institute of Epidemiology, Neuherberg, Germany)

Presenter: Dr WOLF, Kathrin (Helmholtz Zentrum München GmbH - German Research Center for Environmental Health, Institute of Epidemiology, Neuherberg, Germany)

Session Classification: Parallel Track 2

Track Classification: Facilitating connectivity of research data: Data interoperability through harmonised metadata and interoperable semantics

Contribution ID: **25** Type: **Talk**

ELN-DIY-Meta: Creating Interoperability for ELNs

Wednesday 11 October 2023 09:30 (20 minutes)

Electronic lab notebooks (ELNs) are essential for gathering analog metadata, including challenging-to-digitize experimental parameters. However, interdisciplinary research institutions often employ various systems, creating barriers to metadata exchange. Addressing this interoperability gap, we're developing an API-based data exchange to enhance interoperability between the ELNs Herbie and Chemotion. Here, the project partners contribute their experience from the NFDI4Ing and NFDI4Chem consortia.

Initiating this effort with a particular use case in membrane research, we defined discipline-specific metadata in both ELNs and mapped corresponding data fields, resolving conflicts by expanding each ELNs'frontend and backend. The communication between the ELNs is implemented through an adapter that can handle both RESTful APIs. To maintain development efficiency and expandability, we decided not to implement the communication directly in the individual ELNs. This decision is largely justified by the fact that extending the communication tool to other ELNs or new data structures can be easily achieved.

The communication tool, developed in Python, is a server-based browser application that utilizes Python API packages from the respective ELNs. These API packages of Chemotion and Herbie were developed in the course of this project and can be seen as simple wrappers for the RESTful APIs of the ELNs. An administrator has to define a synchronization key –a key used to determine pairs of elements in both ELNs to be synchronized. Another noteworthy feature is that the administrator can then establish the data field mapping for the synchronized elements through a user-friendly click-and-point interface.

Thus, we have taken another step towards linked ELNs and breaking obstacles between disciplines. This effort will facilitate coherent experimental research datasets in the future.

Please assign your contribution to one of the following topics

Data interoperability through harmonised metadata and interoperable semantics

Please specify "other" (stakeholder)

In addition please add keywords.

Electronic Laboratory Notebook, Interoperability, Membrane, Chemotion, Herbie

Please assign yourself (presenting author) to one of the stakeholders.

Researchers

Primary authors: Mr KIRCHNER, Fabian (Helmholtz-Zentrum Hereon); Mr STARMAN, Martin

(KIT); Mr SAHIM, Sayed Ahmad (Helmholtz-Zentrum Hereon); ESCHKE, Catriona (Helmholtz-Zentrum Hereon); HELD, Martin (Hereon); Dr JUNG, Nicole (KIT)

Presenters: Mr KIRCHNER, Fabian (Helmholtz-Zentrum Hereon); Mr STARMAN, Martin (KIT)

Session Classification: Parallel Track 2

Track Classification: Facilitating connectivity of research data: Data interoperability through harmonised metadata and interoperable semantics

Contribution ID: **27** Type: **Poster**

ADVANCE: Advanced metadata standards for biodiversity survey and monitoring data for supporting research and conservation

Tuesday 10 October 2023 14:30 (15 minutes)

In an ever-changing world, field surveys, inventories and monitoring data are essential for prediction of biodiversity responses to global drivers such as land use and climate change. This knowledge provides the basis for appropriate management. However, field biodiversity data collected across terrestrial, freshwater and marine realms are highly complex and heterogeneous. The successful integration and re-use of such data depends on how FAIR (Findable, Accessible, Interoperable, Reusable) they are. ADVANCE aimed to underpin rich metadata generation with interoperable metadata standards using semantic artefacts, facilitating integration and reuse of biodiversity monitoring data across terrestrial, freshwater and marine realms. To this end, we revised, adapted and expanded existing metadata standards, thesauri and vocabularies. We focused on the most comprehensive database of biodiversity monitoring schemes in Europe (DaEuMon) as the base for building a metadata schema that implements quality control and complies with the FAIR principles. We also created a vocabulary with the most common terms used in biodiversity datasets. We tested and refined the strength of the concept in real use cases, and made both the FAIR metadata schema and the vocabulary available for reuse in open access platforms. Moreover, the ADVANCE metadata schema is being integrated in the new UFZ Biodiversity Platform BioMe, allowing data providers to make their metadata available as well as users to search for data comprehensively described with metadata and reuse them, enabling assessments of the relationships between biodiversity across realms and associated environmental conditions.

Please assign your contribution to one of the following topics

Metadata annotation and management close to the research process

Please specify ”other” (stakeholder)

In addition please add keywords.

metadata schema, interoperability, biodiversity data, FAIR, biodiversity monitoring

Please assign yourself (presenting author) to one of the stakeholders.

Researchers

Primary author: SILVA MENGER, Juliana (UFZ, AWI)

Co-authors: GRIMM-SEYFARTH, Annegret (UFZ); HARPKE, Alexander (UFZ); HENLE, Klaus (UFZ); FRENZEL, Mark (UFZ); RICK, Johannes (AWI); WILTSHIRE, Karen (AWI)

Presenter: SILVA MENGER, Juliana (UFZ, AWI)

Session Classification: Poster session

Track Classification: Poster session

Contribution ID: **28** Type: **Talk**

Join the new Earth System SciencesTerminology Service (ESS TS)

Wednesday 11 October 2023 11:50 (20 minutes)

Using terminologies can empower scientists and infrastructure providers to realise a machine-processable expression of the information contained in their research data and other academic outputs. In the academic world, the ambiguity of terms and the lack of appropriate keywords is tedious and annoying to both, scientists and machines. In addition, there is a lack of controlled vocabularies in many scientific fields. In some cases, the selection of the most appropriate terminology is also difficult. On the other hand, repositories try to promote the use of terminologies as they offer building blocks for (meta-)data schemata and data annotations and allow the persistent reference to concepts and terms by assigning identifiers like Uniform Resource Identifiers (URIs or handles such as Digital Object identifiers (DOIs).
The BITS project (BluePrints for the Integration of Terminology Services in Earth System Sciences) is trying to find solutions to these problems. As a first step, BITS builds **a Terminology Service (TS) for subfields of climate science and geodiversity collections** (Earth's diversity of i.a. rocks, fossils, soils, sediments). For this, the project leverages the existing Terminology Service of the TIB –Leibniz Information Centre for Science and Technology, which currently features more than 160 ontologies, 1.1 million terms, and over 22,000 properties from a range of domains such as architecture, chemistry, computer science, mathematics, and physics. The TS will then be integrated into the two different data repositories of the German Climate Computing Center (DKRZ) and the Senckenberg - Leibniz Institution for Biodiversity and Earth System Research (SGN). In close collaboration with NFDI4Earth and the wider ESS community and TS4NFDI as the NFDI base service project for Terminology Services, the experience gained in building the TS and integrating it into the repositories at DKRZ and SGN will be used to create blueprints that can later be used to connect other Earth System Science repositories to the TS.
This is why we want to join forces with the ESS community. Tell us:

- What are your needs for terminology?
- What do you expect from such a Terminology Service?
- Which terminologies should be part of this TS?
- To use this TS, what tools do you need?
- How should this TS work together with other collections and TSs, e.g. for Biological Data?
- Which further semantic artefacts (like semantic mappings) are of interest for such a service?

Join us in building and transferring into your research community the ESS TS so that it can serve as a valuable resource for researchers, students, professionals, and developers, providing them with accurate and consistent terminology to enhance their work, improve communication and advance knowledge in their respective fields. Want to know more? Get in touch at Info.BITS@tib.eu.

Please assign your contribution to one of the following topics

Enabling and incentivising the research community

Please specify "other" (stakeholder)

In addition please add keywords.

Semantic Artefacts, Terminology Service, Earth System Sciences, BITS

Please assign yourself (presenting author) to one of the stakeholders.

Data professionals who provide and maintain data infrastructure

Primary author: Dr GANSKE, Anette (TIB –Leibniz-Informationszentrum Technik und Naturwissenschaften)

Co-authors: Dr KRAFT, Angelina (TIB –Leibniz Information Centre for Science and Technology]); Dr STOCKER, Markus (TIB –Leibniz Information Centre for Science and Technology])

Presenter: Dr GANSKE, Anette (TIB –Leibniz-Informationszentrum Technik und Naturwissenschaften)

Session Classification: Parallel Track 1

Track Classification: Facilitating connectivity of research data: Data interoperability through harmonised metadata and interoperable semantics

Contribution ID: **29** Type: **Poster**

The Helmholtz Knowledge graph and the unified Helmholtz Information and data exchange (unHIDE)

Tuesday 10 October 2023 13:20 (10 minutes)

Research across the Helmholtz Association is based on inter- and multidisciplinary collaborations across its 18 Centres and beyond. However, the (meta)data generated through Helmholtz research and operations is typically siloed within institutional infrastructures and often within individual teams. The result is that the wealth of the association's (meta)data is stored in a scattered manner, hard to find and consequently cannot be used to its full value to scientists, managers, strategists, and policy makers.
To address this challenge, the Helmholtz Metadata Collaboration (HMC) launched the unified Helmholtz Information and Data Exchange (unHIDE) in 2022. We are creating a lightweight and sustainable interoperability layer to interlink data infrastructures; and increase visibility and access to the Helmholtz Association's (meta)data and information assets. Using proven and globally adopted knowledge graph technology, within unHIDE we develop a comprehensive association-wide knowledge graph (KG) the Helmholtz-KG: a solution to connect (meta)data, information, and knowledge.
A first prototype of the Helmholtz KG was released in April 2023. This includes a comprehensive web front end for manual search of resources 1, a stable and documented 2 backend with a tested data ingestion and integration pipeline, and machine accessible endpoints 3.
In this poster we present an overview of the first release of the Helmholtz KG, how it integrates metadata from heterogeneous sources to make it visible and findable, and how further data providers can be connected. We will illustrate the lessons learned during our development process and give an outlook into the next steps and future release versions of the KG.In the future we aim to move to a user driven co-design process for prioritization of features based on the needs of the KG stakeholder. For this we are organizing a related workshop at the conference - you are welcome to attend.

1 https://search.unhide.helmholtz-metadaten.de/

2 https://docs.unhide.helmholtz-metadaten.de/

3 https://sparql.unhide.helmholtz-metadaten.de/

Please assign your contribution to one of the following topics

Technological solutions for findable and machine-readable metadata

Please specify ”other” (stakeholder)

In addition please add keywords.

Metadata unHIDE Knowledge-Graph Helmholtz FAIR

Please assign yourself (presenting author) to one of the stakeholders.

Data professionals and stewards

Primary authors: BRÖDER, Jens (Forschungszentrum Jülich GmbH (IAS-9)); HOFMANN, Volker

Co-authors: D'MELLO, Fiona (Forschungszentrum Jülich); PREUSS, Gabriel (HZB); MANNIX, Oonagh (HMC matter/HZB); BUTTIGIEG, Pier (GEOMAR Helmholtz Centre for Ocean Research); FATHALLA, Said; SANDFELD, Stefan; SERVE, Vivien (HZB, HMC)

Presenters: BRÖDER, Jens (Forschungszentrum Jülich GmbH (IAS-9)); HOFMANN, Volker

Session Classification: Poster session

Track Classification: Poster session

Contribution ID: **30** Type: **Poster**

An Interactive FAIR-Data Publishing System for Time-Series Data

Tuesday 10 October 2023 14:30 (15 minutes)

An important point for the widespread dissemination of FAIR-data is the lowest possible entry barrier for preparing and providing data to other scientists according to the FAIR criteria. If scientists have to manually extract, transform and annotate the data according to the FAIRcriteria and then export it to make it available to the public, this requires a significant investment of time that does not primarily reward the scientists who prepares and provides the data.

The Energy-Lab at KIT is running a large cluster of an Influx database management system in which energy related time-series data being stored in a variety of individual databases for over 10 years. In order to increase the willingness to make data available to the scientific public, we have developed a web application that greatly supports and automates the publication and annotation process of time-series-data stored in Influx databases.

The web application consists of a backend service and an interactive frontend. The backend provides arbitrary, predefined and annotated time-series-data of a measurement (an Influx database structure that corresponds to a table in a relational database) via an URL without requiring any further information for access. The specification of the data is done via HTTP-GET parameters. These include the desired time interval and specific conditions on the attributes as well as a configuration file in which the Influx-server access information are stored. The actual request is made by a series of REST API calls to the InfuxDB. In order to be able to extract arbitrarily large amounts of data, a stream-based approach was chosen. The data is returned as an RO-Crate dataset, using the CSVW Namespace to describe tabular data. The column data-types are extracted from metadata calls to the Influx database. Further information about the attributes (like unit, quantity) can be specified in the configuration file.

The frontend implements the interactive construction of the URL for reading out the time-series-data. The first step is to select the specific configuration file stored for a particular measurement, which contains the information (see above) for accessing a specific database. This information is used, to access the measurement and determine the time interval for which data is available. Meta information of the measurement is read out including the attributes with their data types. In addition, for attributes which act as tags (descriptive attributes), the existing tag-values are extracted. These attributes can be used to interactively formulate extraction conditions (e.g. only data of certain buildings, devices,…). Finally, the time interval of the data to be extracted must be specified. The result of this step is a URL, conforming to the backend API, to export the data in the appropriate format. Currently all attributes (tags & fields) are returned.

In a future version we also plan that it is possible to specify which attributes should be returned. Further extensions include a cache mechanism, where only dynamic data must be retrieved on every request while historical data requests can be fulfilled from a cache.

Please assign your contribution to one of the following topics

Technological solutions for findable and machine-readable metadata

Please specify ”other” (stakeholder)

In addition please add keywords.

Data-exporter, CO-crate, Export-configurator, time-series data

Please assign yourself (presenting author) to one of the stakeholders.

Researchers

Primary author: SCHMIDT, Andreas (KIT)

Co-authors: Mr KOUBAA, Mohamed Anis (Institute for Automation and applied Informatics); SCHWEIKERT, Jan (KIT); STUCKY, Karl-Uwe (KIT); SUESS, Wolfgang (KIT)

Presenter: SCHMIDT, Andreas (KIT)

Session Classification: Poster session

Track Classification: Poster session

Contribution ID: **31** Type: **Poster**

DISOS: An Ontology Suite for Modeling Dislocations in Crystalline Materials

Tuesday 10 October 2023 14:30 (15 minutes)

Materials Science and Engineering (MSE) is concerned with the design, synthesis, properties, and the performance of materials. Metals and semiconductors are an important type of crystalline materials that usually have defects. One of the common types of line defects is the "Dislocations" which strongly affect numerous material properties, including strength, fracture toughness, and ductility.

The past few years have seen a significant effort in understanding dislocation behavior at different length scales, using both experimental and simulation techniques. However, there is still a lack of common standards to represent and connect dislocation-related knowledge across different communities. An ideal solution to this problem is to represent this data using a formal language with unique and well-defined concepts that is also accessible to machines. Formal knowledge representation through ontologies allows for data interoperability and sharing between related MSE domains in a machine-readable format, thereby enabling machine actionability.

We develop the Dislocation Ontology Suite (DISOS), an ontology suite that comprises four modules describing scientific concepts of materials, representations of dislocations, and different simulation models in the dislocation domain: 1) the dislocation ontology (DISO) 1 represents dislocation-related concepts such as Burgers vectors, slip planes, and slip systems; 2) the Crystallographic Defect Ontology (CDO) describes the physical entity of crystalline materials and crystallographic defects concepts; 3) the Crystal Structure Ontology (CSO) describes the crystal model/representation and coordinate system, and 4) Simulation Data (SIM) models data-related aspects of some simulation frameworks in the context of dislocation simulations. Furthermore, DISOS is used to describe dislocation-related data (see, e.g., the RDF dataset 2 of dislocation data) with the goal of increasing the interoperability in dislocation use cases. We believe that efforts made are important for establishing FAIR (Findable, Accessible, Interoperable, and Reusable) 3 data in the MSE domains.

References
1 A. Z. Ihsan; S. Fathalla; Stefan Sandfeld. In The 38th ACM/SIGAPP Symposium on Applied Computing (SAC '23), 2023. https://doi.org/10.1145/3555776.3578739
2 https://purls.helmholtz-metadaten.de/disos/rdfdata
3 Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016).

Please assign your contribution to one of the following topics

Data interoperability through harmonised metadata and interoperable semantics

Please specify "other" (stakeholder)

In addition please add keywords.

Ontology, Dislocation, Crystallographic Defects, Semantic Web

Please assign yourself (presenting author) to one of the stakeholders.

Researchers

Primary author: IHSAN, Ahmad Zainul

Co-authors: FATHALLA, Said; SANDFELD, Stefan

Presenter: IHSAN, Ahmad Zainul

Session Classification: Poster session

Track Classification: Poster session

Contribution ID: **32** Type: **Poster**

User-Oriented, Reusable Components and Tools for the Integration of FDOs into the Daily Research Routine

Tuesday 10 October 2023 14:30 (15 minutes)

In the last two years, the endeavor of realizing FAIR Digital Objects (FDOs) took a huge leap on the international as well as on the national level in Germany and in particular within HMC. By finding consensus on a common Helmholtz Kernel Information Profile (1) defining basic kernel metadata attributes each FDO must provide to serve as top-level commonality across all research fields, the way was paved to bring FDOs into practice. The initial focus was on technical aspects on how to create and manage FDOs, which resulted in documents (2), infrastructure components (3), and tools (4) addressing these aspects.

Early adoptions were carried out in different contexts, e.g., in NFDI-MatWerk to create a set of reference datasets for different participant projects (5) (6), in collaboration with the Helmholtz incubator platform on AI to represent a complex dataset of annotated images (7), and in different HMC Hubs and projects.

It quickly became apparent that a growing number of FDOs also has an impact on their design and handling. A better understanding of common practices of creating FDOs in different domains also increased their complexity, which brought up a strong demand on additional tooling to fill existing gaps with regard to the creation, management, retrieval, and representation of FDOs.

On this poster we present our results on addressing the creation, retrieval, and representation of FDOs. For the creation of FDOs the FDO Creator was implemented to ease the manual creation of small to mid-size FDO graphs. Realized on top of the Typed PID Maker, the FDO Creator simplifies the creation of validated FDOs and their linking to each other. Once created, the retrieval of FDOs becomes relevant for scientific users. With this in mind, the FDO Search (8) is taking the first steps to implement a user-friendly search for FDOs. Based on PID Records indexed by the Typed PID Maker in an Elastic instance, full text as well as faceted search are offered to the user. Single results can then be visualized either in FAIR-DOscope or by one of the new reusable Web Components for visualizing FDOs as graphs or rendered as text components, offering multiple ways for further interaction.

The strong focus on the reusability aspect of the created Web Components will ease the integration of FDOs into existing software, as well as increase their acceptance by making them tangible for scientific users. Allowing FDOs to be visualized in basically every Research Software Environment will improve their visibility and will foster a natural interaction as part of a researcher's daily business, hiding technical details of a FDO's implementation.

This research has been supported by the Helmholtz Metadata Collaboration (HMC) Platform, the German National Research Data Infrastructure (NFDI) and the German Research Foundation (DFG).

Please assign your contribution to one of the following topics

Technological solutions for findable and machine-readable metadata

Please specify "other" (stakeholder)

In addition please add keywords.

FAIR Digital Objects, Web Components

Please assign yourself (presenting author) to one of the stakeholders.

Scientists and technicians who maintain and operate research infrastructure for data generation

Primary authors: Mr KIRAR, Ajay (Karlsruhe Institute of Technology); Mr INCKMANN, Maximilian (Karlsruhe Institute of Technology); JEJKAL, Thomas

Presenter: JEJKAL, Thomas

Session Classification: Poster session

Track Classification: Poster session

Contribution ID: **33** Type: **Talk**

The HMC-STAMPLATE-Project: Our journey towards an interlinked research data infrastructure for environmental sciences

Wednesday 11 October 2023 09:50 (20 minutes)

Time-series data are crucial sources of reference information in all environmental sciences. Publishing such data consistently and timely for monitoring and warning purposes becomes more and more important. In this context, the Helmholtz-Centers from the research field Earth and Environment (E&E) operate some of the largest measurement-infrastructures worldwide (e.g., TERENO, DANUBIUS or MOSES). To ensure consistency and comparability of (meta)data from these infrastructures according to the FAIR-principles, we have to ensure standardized interfaces and metadata conventions. But a state-of-the-art and community-driven framework for time-series- and sensor-(meta)data, that is jointly adopted across different scientific fields and communities, is still missing. In this context, the Open Geospatial Consortium (OGC) recently proposed the SensorThings API (STA) as an open, geospatial-enabled and unified way to interconnect Internet of Things (IoT) devices, data, and applications over the Web.

Within our STAMPLATE-Project, that is funded by the Helmholtz Metadata Collaboration (HMC), all seven Centers from the research field E&E (AWI, FZJ, Geomar, GFZ, Hereon, KIT, UFZ) as well as the Fraunhofer IOSB hence joined forces to develop the technical and semantic foundations for establishing STA as a flexible and interoperable standard interface for making time-series data, enriched with comprehensive and standardized metadata, available over the internet.

The project is carried out by a highly experienced consortium across the seven E&E Centers. This consortium forms the core of an enhanced user community, that also includes potential end-users of our STA implementations and other interested parties.

In this presentation, we now want to show the current status of our project, demonstrate typical use-cases and also discuss challenges on our journey towards an interlinked research data infrastructure. We also want to advertise the SensorThings API as a generic interface for time-series data beyond our research field.

Please assign your contribution to one of the following topics

Infrastructure and common practices for consolidating (meta)data

Please specify ”other” (stakeholder)

In addition please add keywords.

Metadata, SensorThings API, environmental sciences, research data infrastructure, FAIR

Please assign yourself (presenting author) to one of the stakeholders.

Data professionals who provide and maintain data infrastructure

Primary authors: LORENZ, Christof (Karlsruhe Institute of Technology); KLEEBERG, Ulrike (Hereon)

Co-authors: LEHMANN, Andreas (Geomar); LOUISOT, Benjamin (KIT); FABER, Claas (Geomar); SCHÄFER, David (UFZ); VAN DER SCHAAF, Hylke (Fraunhofer IOSB); HANISCH, Marc (GeoForschungsZentrum Potsdam GFZ); RYAN, Marie (Hereon); KUNKEL, Ralf (FZJ); KOPPE, Roland (AWI); BARTHLOTT, Sabine (KIT)

Presenter: LORENZ, Christof (Karlsruhe Institute of Technology)

Session Classification: Parallel Track 1

Contribution ID: **34** Type: **Talk**

Enhancing Transparency and Reproducibility in AI Model Training through Provenance-Enabled Data Preprocessing and Workflow Documentation

Wednesday 11 October 2023 11:10 (20 minutes)

As the scale and complexity of AI models continue to grow, the demand for vast amounts of data, including unlabelled and uncurated datasets, has become increasingly prevalent. To address this challenge, the role of data preprocessing, filtering, augmentation, and curation using automated methods has gained increasing significance in ensuring optimal model performance. However, while AI models themselves are increasingly documented, the transparency surrounding the preprocessing techniques applied often remains incomplete, impeding reproducibility.

This paper proposes a novel approach aimed at improving transparency and reproducibility in training scenarios by capturing concrete provenance information throughout the data preprocessing workflow. By recording the sequence of transformations applied to the data, researchers and developers gain the ability to recreate workflows and analyze sample provenance, improving informed decision-making. Simultaneously, this approach offers a pathway for gathering metadata, which serves debugging, monitoring, and development purposes, exposing developers to valuable insights.

The proposed solution is realized through the introduction of a lightweight data pipeline library designed for seamless chainable stream operations. Focused on this common use case, this library enables the systematic capture of structured provenance information, reducing the overhead of subsequent analysis. Demonstrated in the context of a compact computer vision use case, the methodology not only exposes useful training metrics but also comprehensively documents the workflow with negligible performance implications.

Further work involves expanding the application of this approach to encompass larger and more complex use cases. Additionally, the integration of intermediate dataset versioning, facilitated by a dedicated DVC plugin, allows for including intermediate data versioning and traceability, thereby improving on the overall reproducibility and transparency of AI model training workflows.

Please assign your contribution to one of the following topics

Metadata annotation and management close to the research process

Please specify ”other” (stakeholder)

In addition please add keywords.

Provenance, Metadata collection, transparent AI, ML Training workflows

Please assign yourself (presenting author) to one of the stakeholders.

Researchers

Primary author: HOFFMANN, Nils

Presenter: HOFFMANN, Nils

Session Classification: Parallel Track 2

Contribution ID: **35** Type: **Poster**

Metadata Extraction Tool and Schema Mapper for Scanning Electron Microscopy (SEM) images

Tuesday 10 October 2023 14:30 (15 minutes)

Standardized metadata and its proper storage are essential for effective management of scientific research data. The challenge lies in manually compiling such metadata, a process which can be both tedious and prone to human error. To address this problem, we introduce the Mapping Service, developed within the framework of HMC.

The Mapping Service helps to streamline the process of metadata extraction and mapping according to existing community-agreed schemas. This tool has been designed as an adaptable and extensible service suitable across various research disciplines. It allows for add-ons which facilitate the extraction of metadata from otherwise proprietary and non-standard file formats and the mapping to schemas, which strengthens the metadata's interoperability and reusability.

The Mapping Service functions as a platform hosting a diverse suite of plugins. These plugins, each equipped with two primary components—a reader and a mapper—are instrumental in achieving metadata extraction and mapping for various experimental techniques. In many research environments, accessing metadata is a bottleneck due to proprietary formats that demand specific software, often leading researchers to manually transcribe unstructured and poorly-documented metadata embedded within, for example, research images. This manual approach, especially for large datasets comprising hundreds of files, is not only time-consuming but also introduces the potential for human errors. The Mapping Service elegantly addresses these challenges: the reader retrieves metadata from a set of diverse research data, such as images or metadata files, while the mapper discerningly selects key variables prescribed by the user-selected schema from the extracted metadata. These variables are then mapped to their respective schema names, resulting in a systematically formatted JSON metadata document.

Through this poster, we showcase one use case via the mapping of Scanning Electron Microscopy (SEM)/Focused Ion Beam (FIB) tomography metadata to our published schema. This functionality is available as a plugin or "Mapping Component"on the Mapping Service and works to emphasize how a large and complex dataset containing a large amount of research images may be easily and efficiently transformed into a single metadata document using an intuitive user interface.Though the poster showcases the use case of SEM/FIB tomography, the Mapping Service has been designed to be a general-purpose tool. Additional plugins tailored to map from one arbitrary schema to another can easily be integrated, and a suite of such plugins is currently in development. Additionally, the service's standalone web service architecture and user interface ensures ease of adoption without any local dependencies or installations required by end users.

In summary, researchers across various fields seeking a streamlined approach to consistent metadata processing will find the Mapping Service to be a useful tool. With this poster, we aim to illuminate the advantages of the Mapping Service's automated extraction and mapping capabilities as well as the necessary considerations and prerequisites for its implementation in a research environment.

This work is supported by NFDI-MatWerk (DFG –n. 460247524), NFDI4Ing (DFG –n. 442146713), NFFA-Europe-Pilot (EU H2020 –n. 101007417), the research program 'Engineering Digital Futures' and the Helmholtz Metadata Collaboration (HMC) platform by the Helmholtz Research Association.

Please assign your contribution to one of the following topics

Metadata annotation and management close to the research process

Please specify ”other” (stakeholder)

In addition please add keywords.

mapping, metadata, service, extraction

Please assign yourself (presenting author) to one of the stakeholders.

Data professionals who provide and maintain data infrastructure

Primary authors: KIRAR, Ajay (Karlsruhe Institute of Technology); VITALI, Elias Giulio Georg (Karlsruhe Institute of Technology); INCKMANN, Maximilian (Karlsruhe Institute of Technology); BLUMENROEHR, Nicolas (Karlsruhe Institute of Technology, Steinbuch Centre for Computing); STOTZKA, Rainer; JOSEPH, Reetu Elza (Karlsruhe Institute of Technology); AVERSA, Rossella (Karlsruhe Institute of Technology); JEJKAL, Thomas; HARTMANN, Volker (KIT)

Presenters: VITALI, Elias Giulio Georg (Karlsruhe Institute of Technology); JEJKAL, Thomas

Session Classification: Poster session

Track Classification: Poster session

Contribution ID: **36** Type: **Poster**

Towards Research-Object-Crate v1.2, with ro-crate-java

Tuesday 10 October 2023 14:30 (15 minutes)

Research Object Crate (RO-Crate) is an open, community driven data package specification to describe all kinds of file-based data, as well as entities outside the package. In order to do so, it uses the widespread JSON-format, representing Linked Data (JSON-LD), allowing to link to external information. This makes the format flexible and machine-readable. These packages are being referred to as (RO-)crates.

Similar to other formats, RO-Crates is based on files and folders and has a single metadata file to describe the whole package. Therefore, such packages are easy to share between different computer systems and software.

In order to create such crates, the RO-Crate community developed libraries written in different programming languages like Python, Ruby, JavaScript, and Java. With Describo, there is also a graphical user interface available.

We developed the ro-crate-java library, which allows creating, modifying and validating crates using the Java Programming Language. The focus of development was the ease of use: We aimed to make it intuitive and easy to create valid crates, without knowing the specification too well. Our implementation can be used for integration into repositories or other services or tools. The library was introduced in the HMC conference 2022 poster session. This follow-up poster will give a preview on a draft feature which is available in the RO-Crate 1.2-DRAFT specification and has been requested a lot: the ability to specify the conformance with multiple profiles within one crate.

Profiles are "a set of conventions, types and properties that one minimally can require and expect to be present in that subset of RO-Crates"(RO-Spec 1.1). They may be used to validate the crate against institutional constraints or to guarantee required information for further processing or visualization.
The new specification includes the possibility to create crates with multiple profiles being specified. As this is an often requested feature, this is now a supported feature since ro-crate-java v1.1.0. The library now makes a difference between stable and unstable features and will update the specification version accordingly.

This research has been supported by the Helmholtz Metadata Collaboration (HMC) Platform, the German National Research Data Infrastructure (NFDI) and the German Research Foundation (DFG).

Please assign your contribution to one of the following topics

Technological solutions for findable and machine-readable metadata

Please specify "other" (stakeholder)

In addition please add keywords.

linked data metadata files folders

Please assign yourself (presenting author) to one of the stakeholders.

Data professionals who provide and maintain data infrastructure

Primary author: PFEIL, Andreas (Karlsruhe Institute of Technology (KIT))

Presenter: PFEIL, Andreas (Karlsruhe Institute of Technology (KIT))

Session Classification: Poster session

Track Classification: Poster session

Contribution ID: **37** Type: **Poster**

Enhancing Usability in Linked Data Editing in Web Applications

Tuesday 10 October 2023 14:30 (15 minutes)

Editing Linked Data documents represents an enormous challenge to users with limited technical expertise. These users struggle with language rules, relationships between entities, and interconnected concepts. These issues can result in frustration and low data quality. In order to respond to this challenge, we introduce a new editor, designed to facilitate effortless editing of JSON-LD documents, catering to both newcomers and advanced users. It is made for easy and seamless integration into other web-based applications and can be used similar to an HTML tag.

The complexity of Linked Data arises from its graph-like structure, where entities are connected through relationships, forming a complex web of semantic connections. While this is advantageous for data integration and cross-platform compatibility, this effort presents significant barriers for those not well-versed in technical aspects. Even with the rise of user-friendly interfaces, manually modifying JSON-LD documents can lead to mistakes in structure and unintended disruptions to valuable linkages.

Our proposed solution is a reusable web component based on modern browser technologies. It offers a view on the data which is easier to perceive than typical graph visualizations. This view shows the data as a list of named entities and their properties to simplify the visual complexity, without giving up on the conceptual graph structure. The list view brings the conceptual entities to the front, but still supports more technical structure elements like blank nodes, as they still exist as properties.

Using schema.org's machine-readable definitions, the editor understands how entities may or may not be connected. This is used to offer autocomplete functionality and avoid the invalid use of the schema.org vocabulary. This functionality can be extended using the integrated schema loader concept.

From a technical point of view, the web component is an HTML Element which takes a (possibly empty) JSON-LD document. It then provides the modified document as a callback as soon as the user saves the document from within the editor. It is therefore easily integrable into existing projects based on arbitrary web frameworks and does not require any special interface implementations. The component is based on StencilJS, which allows generating wrappers for popular frameworks, for tighter integration.

In conclusion, our web component empowers both new and experienced users to edit Linked Data seamlessly, overcoming the inherent challenges associated with manual JSON-LD modification. By simplifying the view on the graph structure and providing an intuitive and supporting interface, the component enhances the ease of use and accessibility of Linked Data editing. This holds significant potential for expediting data curation, collaboration, and integration, thus fostering a more inclusive and dynamic Linked Data ecosystem.

This research has been supported by the Helmholtz Metadata Collaboration (HMC) Platform, the German National Research Data Infrastructure (NFDI) and the German Research Foundation (DFG).

Please assign your contribution to one of the following topics

Enabling and incentivising the research community

Please specify ”other” (stakeholder)

In addition please add keywords.

linked data metadata web editor

Please assign yourself (presenting author) to one of the stakeholders.

Data professionals who provide and maintain data infrastructure

Primary author: MAJER, Lorenz (Karlsruhe Institute of Technology (KIT))

Co-author: PFEIL, Andreas (Karlsruhe Institute of Technology (KIT))

Presenter: PFEIL, Andreas (Karlsruhe Institute of Technology (KIT))

Session Classification: Poster session

Track Classification: Poster session

Contribution ID: **38** Type: **Talk**

Design of metadata schemas for ion chromatography in applied plasma sciences

Tuesday 10 October 2023 11:30 (20 minutes)

Ion chromatography (IC) is an analytical method that separates ions in liquid samples according to their chemical and physical properties. This analytical method is widely used in different scientific fields such as environmental science (for example wastewater or soil analysis), food technology (food extract analysis), applied plasma science (characterization of plasma treated liquids) and protein purification.
To date publications often lack method-relevant parameters, such that a full adaptation of the published methods is not possible. This requires additional testing and hinders the reuse and comparison of methods for IC. We observe that no open, standardized metadata schema exists for IC and related analytical methods, e.g. for high-performance liquid chromatography and gas chromatography. Therefore, we propose a standardized metadata schema to increase the interoperability and reproducibility of scientific investigations in IC. The metadata schema furthermore is a step towards introducing the FAIR principles to the scientific community as a starting point to discuss modern practices in scientific research data management.

This work focuses on the design of two standardized metadata schemas, one for anions and one for cations. These schemas built upon the JSON standard, which sets rules and guidelines for the structure and format of JSON files. Their design is based on the ASTM E 1151 norm for terms and relationships in IC and user-feedback resulting from a coupled data stewardship in applied plasma science. Both schemas are similar in structure and contain relevant metadata for the description of the instrument method (the device settings), processing method (detection windows for automatic labeling of peaks), sample parameters, column properties, quality assurance parameters. Differences are in the specific fields, e.g. different columns for anions/cations and the definition of detection windows for the peak labeling is different. The mandatory fields according to the DataCite metadata schema v4.4 are used for publishing data in repositories and for data citations. The metadata schemas were designed and tested in a laboratory workflow at the INP with a research data management tool called Adamant 1, which is a user-friendly JSON schema based metadata editor. Users without any programming experience can edit and collect metadata in an HTML form, enabling them to properly collect and validate their metadata (a benefit from the JSON schema standard). The presented use-case demonstrates the paradigms of modern research data management for research in the laboratory. The collected metadata can also be sent directly from Adamant to the institute's instance of the electronic lab notebook (ELN) "eLabFTW"2 to enable a smoother transition from the traditional handwritten laboratory notebooks to a digital solution, which increases both the interoperability and accessibility of the collected metadata. The ELN also allows the further description of the experiments and linkage with other parts of the related experiment, e.g. sample preparation, treatment and data post processing.

1. Chaerony Siffa, I., J. Schäfer, and M. Becker, Adamant: a JSON schema-based metadata editor for research data management workflows. F1000Research, 2022.
2. Nicolas Carpi, A.M.a.M.P., eLabFTW: An open source laboratory notebook for research labs. The Journal of Open Source Software, 2017.

Please assign your contribution to one of the following topics

Metadata annotation and management close to the research process

Please specify ”other” (stakeholder)

In addition please add keywords.

Research data management, ion chromatography, FAIR principles

Please assign yourself (presenting author) to one of the stakeholders.

Researchers

Primary author: WAGNER, Robert (Leibniz Institute for Plasma Science and Technology (INP), Greifswald)

Co-authors: Prof. WALTEMATH, Dagmar (Medical Informatics Laboratory, University Medicine Greifswald); Mr CHAERONY SIFFA, Ihda (Leibniz Institute for Plasma Science and Technology (INP), Greifswald); Dr BECKER, Markus (Leibniz Institute for Plasma Science and Technology (INP), Greifswald)

Presenter: WAGNER, Robert (Leibniz Institute for Plasma Science and Technology (INP), Greifswald)

Session Classification: Parallel Track 2

Track Classification: Facilitating connectivity of research data: Metadata annotation and management during and close to the research process

Contribution ID: **39** Type: **Talk**

Advancing Research Data Management with Python-Flask Applications

Wednesday 11 October 2023 10:50 (20 minutes)

Data acquisition (DAQ) systems continue to advance in power, but manual data input will remain required as experiments necessarily check for the unforeseen. Researchers often use electronic lab books or fallback solutions like Excel or Google Docs to record actions and events, highlighting the need for an intuitive interface that enables live- and post-processing while remaining linkable to DAQ systems. A major challenge lies in the dynamic nature of the incoming information, rendering fixed-structure databases like SQL impractical. To tackle this issue, we have developed intuitive Python-Flask applications that harness the inherent flexibility of document-based databases, particularly MongoDB, for storing curated and query-ready data. Although these applications were initially tailored for the laser-particle acceleration group at HZDR as part of the DAHPNE4NFDI project, the intention is to generalize their utility.

Our three interrelated apps are currently as follows:

1. This app facilitates manual data entry during experiments like in traditional table schemas but via a database form. The form can be easily –and on-the-fly –changed according to the needs and allows also to store the DAQ configuration per entry. Choices can be preconfigured or pulled from other sources like a MediaWiki lab documentation system. Entries are directly written to a MongoDB.
2. ZeroMQ Relayer: This extracts metadata from the experiment's drive laser (via zeroMQ) and forwards this in real-time to the experiments. This enables harmonized metadata like ID's and timestamps, either appended to data as well as logged for post-hoc reconstruction.
3. KafkaWatcher: Functioning as an intermediary, KafkaWatcher receives data from the Relayer and various experiment software agents and publishes it to MongoDB. This app is hosted on a gateway machine and allows consumption of incoming Kafka messages and subsequent storage in MongoDB. Flask-SocketIO is used for real-time reception of Kafka messages.

From a technical perspective, the Python-Flask web development microframework was chosen due to its extensive support and wide range of extensions. Python, being a widely adopted programming language among physicists, was a natural choice. Form validation is accomplished using WTForms, while Jinja2 handles template management. Dynamic form modifications are achieved primarily using JavaScript, while Bootstrap ensures a consistent form layout. MongoDB serves as the backend database for storing form fields and the various collections. Finally, Waitress is employed to serve the applications, offering compatibility with both Linux and Windows environments.

Looking ahead, enhancements like online analysis tools, data path logging, and advanced visualizations are on the horizon. While Shotsheet and KafkaWatcher offer native data search and visualization, integration of MongoDB with Grafana amplifies these capabilities. A harmonization with DAQ and SciCat databases is anticipated, facilitating sophisticated analyses.

For simulations, backing up the experiments, scripts have been created for metadata extraction from SMILEI, WarpX, and PIConGPU codes, tailored for SciCat. The convergence of electronic notebook-databases with simulation data and DAQ inputs holds promise for expanded opportunities in experimental analyses, aligning with the principles of Findable, Accessible, Interoperable, and Reusable (FAIR) data.

Please assign your contribution to one of the following topics

Metadata annotation and management close to the research process

Please specify ”other” (stakeholder)

In addition please add keywords.

MongoDB, Mediawiki, Python, Flask, WebApp

Please assign yourself (presenting author) to one of the stakeholders.

Scientists and technicians who maintain and operate research infrastructure for data generation

Primary authors: SCHLENVOIGT, Hans-Peter (HZDR); TIPPEY, Kristin Elizabeth (HZDR)

Presenter: TIPPEY, Kristin Elizabeth (HZDR)

Session Classification: Parallel Track 2

Track Classification: Facilitating connectivity of research data: Metadata annotation and management during and close to the research process

Contribution ID: **40** Type: **Talk**

HELPMI: Helmholtz Laser-Plasma Metadata Initiative

Tuesday 10 October 2023 11:10 (20 minutes)

Metadata is the foremost element in data management strategy when taking account of the F.A.I.R.(findable, accessible, interoperable and reusable) principles, what is becoming increasingly important within the scientific community. Additionally, there is a strong need for better data integration and enrichment in the field of high-intensity laser-plasma physics in an international context: at many laser-plasma (LPA) research labs, experimental data lacks a standardized metadata format and a way to coherently combine and store it.

A reason for this is the various levels of data origins, ranging from calibration data before an experimental campaign, to detector and machine data during the run. In addition, the diagnostic configuration and experimental setup are often subject to changes throughout one campaign, implying consequences on the experimental results and subsequently increasing the complexity of the data. Setting out from this status quo and given their worldwide leading expertise in laser-driven experiments, Helmholtz-Zentrum Dresden –Rossendorf, Helmholtz-Institut Jena and GSI Helmholtzzentrum für Schwerionenforschung have formed HELPMI within HMC to develop a metadata standard for LPA experiment data.

Currently, a standard only exists for LPA simulations, namely the Open Particle Mesh Data (openPMD). openPMD is a hierarchical format, originally developed and mainly used for –but not limited to –laser-plasma simulations. It supports several file formats (HDF5, JSON) as well as streaming techniques (ADIOS) to avoid data rate limitations on high-performance computing systems.

Within the synchrotron radiation and neutron research community, "NeXus is developed as an international standard by scientists and programmers representing major scientific facilities in Europe, Asia, Australia, and North America in order to facilitate greater cooperation in the analysis and visualization of neutron, x-ray, and muon data". NeXus is also a hierarchical data model built on top of HDF5, able to cope with experimental setup geometry description and incrementally adding analysis results to the original (raw) data. In fact, the NeXus base classes form a set of available data structures, whereas the application definitions define rules which classes are mandatory for certain applications. With contributed definitions, further use cases can be defined and tested before review for official implementation.

Thereby it should be possible to make openPMD compatible or interoperable with NeXus, while first maintaining both standards.

Within HELPMI we will examine, based on a real-world data example, how far the NeXus format definitions can carry laser-plasma experimental data, how the standard's structure elements map to domain-specific structures and if there are limitations, requiring an extension of the standard. Additionally, domain-specific terms will be identified and collected into a glossary of laser-plasma experimental data, which will be done in close contact to the LPA community in order to have well-accepted definitions. The resulting glossary or ontology should then be implemented in a both human- and machine-readable fashion.

Therefore, building on analogies to existing standards in similar domains and in close collaboration with the international community, developing a standard within the HELPMI project would enable F.A.I.R. data with easy access, allowing for automated analyses and cross-comparisons in the research field of laser-plasma interaction.

Please assign your contribution to one of the following topics

Technological solutions for findable and machine-readable metadata

Please specify ”other” (stakeholder)

In addition please add keywords.

data standard, metadata, Nexus, openPMD

Please assign yourself (presenting author) to one of the stakeholders.

Researchers

Primary authors: DEBUS, Alexander (Helmholtz-Zentrum Dresden-Rossendorf); KESSLER, Alexander; POESCHEL, Franz (CASUS/HZDR); SCHLENVOIGT, Hans-Peter (HZDR); Mr HORNUNG, Johannes (GSI); KALUZA, Malte Christoph (Helmholtz-Institut Jena); BUSSMANN, Michael (HZDR); Mr EISENBARTH, Udo (GSI); Mr BAGNOUD, Vincent (GSI)

Presenter: SCHLENVOIGT, Hans-Peter (HZDR)

Session Classification: Parallel Track 2

Track Classification: Facilitating connectivity of research data: Technological solutions for findable and machine-readable metadata

Contribution ID: **41** Type: **Poster**

Unified metadata handling for reproducible simulation workflows

Tuesday 10 October 2023 14:30 (15 minutes)

Computer simulations are an essential pillar of knowledge generation in science. Understanding, reproducing, and exploring the results of simulations relies on tracking and organizing metadata describing numerical experiments. However, the models used to understand real-world systems, and the computational machinery required to simulate them, are typically complex, and produce large amounts of heterogeneous metadata. Capturing and structuring these metadata along the processing chain is a vital requirement, for example, to make numerical experiments reproducible, to enable systematic benchmarking and validation of simulation software and models, to assess the reliability of simulations, and to foster data exploration and comparison [1,2]. Providing the ability to search, share, and evaluate metadata from heterogeneous simulations and environments is however a major challenge. The availability of a common metadata management framework, which can be adopted by scientists from different scientific domains, would therefore be highly desirable and foster the meta-analysis of HPC simulation workflows 3.

Here, we present a general concept for acquiring and handling metadata that is agnostic to software and hardware, and highly flexible for the user. It consists of two steps: 1) recording and storing raw metadata, and 2) selecting and structuring metadata in a configurable manner. We implement this concept in tools that can be attached to existing simulation workflows, and demonstrate it by applying our tools to distinct high-performance computing use cases from hydrology and neuroscience.

1. Guilyardi, E., et. al. (2013) doi: 10.1175/BAMS-D-11-00035.1
2. Manninen, T., et. al. (2018) doi: 10.3389/fninf.2018.00020
3. Ivie, P., & Thain, D. (2018). doi: 10.1145/3186266

Acknowledgments

The authors would like to thank Jan Bumberger, Helen Kollai, Michael Denker, Dennis Terhorst, Rainer Stotzka, Guido Trensch, and Stefan Sandfeld for ongoing Aug 21, 2023, 12:00 PM fruitful discussion. This project was funded by Helmholtz Metadata Collaboration (HMC) ZT-I-PF-3-026, EU Grant 945539 (HBP), Helmholtz IVF Grant SO-092 (ACA), and Joint lab SMHB; compute time was granted by VSR computation grant JINB33, Jülich. The work was carried out in part within the HMC Hub Information at the Forschungszentrum Jülich.

Please assign your contribution to one of the following topics

Infrastructure and common practices for consolidating (meta)data

Please specify ”other” (stakeholder)

In addition please add keywords.

Simulation workflow; metadata management framework.

Please assign yourself (presenting author) to one of the stakeholders.

Researchers

Primary authors: VILLAMAR, Jose (Institute of Neuroscience and Medicine (INM-6) and Institute for Advanced Simulation (IAS-6) and JARA-Institute Brain Structure-Function Relationships (INM-10), Jülich Research Centre, Jülich, Germany; RWTH Aachen University, Aachen, Germany); KELBLING, Matthias (Dept. Computational Hydrosystems, Helmholtz-Centre for Environmental Research, Leipzig, Germany); MORE, Heather (Institute of Neuroscience and Medicine (INM-6) and Institute for Advanced Simulation (IAS-6) and JARA-Institute Brain Structure-Function Relationships (INM-10), Jülich Research Centre, Jülich, Germany; Institute for Advanced Simulation (IAS-9), Jülich Research Centre, Jülich, Germany); TETZLAFF, Tom (Institute of Neuroscience and Medicine (INM-6) and Institute for Advanced Simulation (IAS-6) and JARA-Institute Brain Structure-Function Relationships (INM-10), Jülich Research Centre, Jülich, Germany); SENK, Johanna (Institute of Neuroscience and Medicine (INM-6) and Institute for Advanced Simulation (IAS-6) and JARA-Institute Brain Structure-Function Relationships (INM-10), Jülich Research Centre, Jülich, Germany); THOBER, Stephan (Dept. Computational Hydrosystems, Helmholtz-Centre for Environmental Research, Leipzig, Germany)

Presenter: VILLAMAR, Jose (Institute of Neuroscience and Medicine (INM-6) and Institute for Advanced Simulation (IAS-6) and JARA-Institute Brain Structure-Function Relationships (INM-10), Jülich Research Centre, Jülich, Germany; RWTH Aachen University, Aachen, Germany)

Session Classification: Poster session

Track Classification: Poster session

Contribution ID: **42** Type: **Poster**

Ontology-Based Literature Classification and Knowledge Extraction in the domain of perovskite-based solar cell manufacturing

Tuesday 10 October 2023 14:30 (15 minutes)

The important increase in efficiency of perovskite-based solar cells (PSCs) in the last decade is a result of scientific work, which produced a huge quantity of literature and data-sets (between 2014 and 2022 almost 30,000 reports were published). The aim of this work is to elaborate an ontology which can primarily be used to classify literature paragraphs according to the subject discussed within. Furthermore, the annotation of data-sets using terms of the envisaged ontology can help making them more understandable, interoperable, and reusable.

The approach is to firstly develop an ontology with a team of domain experts and then elaborate knowledge extraction concepts from literature based on it.

From a technical point of view an ontology is a representational artifact, comprising a taxonomy as proper part, whose representations are intended to designate some combinations of universals, defined classes, and certain relations between them. A taxonomy on its role is a hierarchy definition between terms denoting types (or classes) linked by subtype relations. The knowledge extraction process consists of the following steps: **(1)** acquiring relevant text resources, **(2)** processing the text into individual terms (also known as tokenization), **(3)** document segmentation and paragraph classification, **(4)** recognizing tokens as classes of information, **(5)** entity relation extraction, and **(6)** named entity linking. The ontology being developed will facilitate some of the tasks in the knowledge extraction process (mainly entity recognition and disambiguation (4) and predicate determination (to build *(subject, predicate, object)*-triplets) by using link prediction during realization of (5)) by strengthening the context to an ontological commitment and by making the correspondence between *sign*, *concept*, and *thing* (or *symbol*, *thought or reference* and *referent*) more robust, especially when named entities and their corresponding classes in the ontology should be precisely recognised.

On the other hand both metadata and content of several scientific articles are represented as RDF triplets. Metadata of newly published scholarly papers are almost structured and based on unique identifiers (such as ORCID or ROR), which make the task of retrieving them much easier. But the knowledge acquisition from textual content reserve some problems and pitfalls during its transformation to RDF triplets. Using an ontology for this purpose helps overcoming ambiguity issues by applying scales of similarity between graphs defined through ontological facts, and ones extracted from texts by means of token-neighbourhood criteria.

The ontology being developed and the content, augmented with its metadata, are then represented in one property graph. In property graphs it is allowed a set of property–value pairs, called attributes and a label to be associated with nodes and edges, offering additional flexibility when modelling data.

The result is a knowledge graph (KG), a graph of data whose nodes represent entities of interest and whose edges represent potentially different relations between them. Both factual and ontological knowledge are existing side-by-side in this KG. In this data-space, classification using link prediction (CULP) transforms recognition problems to those of link prediction (where we try to find the link between an unlabeled node and the proper class node for it).

Please assign your contribution to one of the following topics

Metadata annotation and management close to the research process

Please specify ”other” (stakeholder)

In addition please add keywords.

Ontology, Knowledge Extraction, Knowledge Graph, Perovskite

Please assign yourself (presenting author) to one of the stakeholders.

Data professionals and stewards

Primary author: Mr KOUBAA, Mohamed Anis (Institute for Automation and applied Informatics)

Co-authors: SCHMIDT, Andreas (IT4EDM/IAI); STUCKY, Karl-Uwe (KIT); SUESS, Wolfgang (KIT)

Presenter: Mr KOUBAA, Mohamed Anis (Institute for Automation and applied Informatics)

Session Classification: Poster session

Track Classification: Poster session

Contribution ID: **43** Type: **Poster**

Good research needs good metadata –why we set out to collect, connect, and correct metadata for physics

Tuesday 10 October 2023 14:30 (15 minutes)

FAIR research data and the adoption of semantic technologies hold a great promise to improve the quality, openness, and efficiency of research in the physical sciences. However, the FAIR building we wish to constructs rests on foundations that are still shaky: Metadata often lack the quantity and quality to harness the full potential of advanced search functionalities, knowledge graphs, and AI applications for scientific libraries.

Physics, as compared to chemistry or the life sciences faces a particular challenge due to the lack of a widely accepted controlled vocabulary. The physics research community would benefit considerably if it were to agree on a sound terminological basis on top of which innovative semantic methods can be stacked. For instance, if (automated) annotation from a controlled vocabulary resulted in more and better fitting keyword, a scientist doing a literature search would enjoy a more accurate but also more concise list of references to follow up. Simultaneously, the author would gain in visibility of their work, especially outside their own field of research.

Hence, TIB –Leibniz Information Centre for Science and Technology, together with partners at Physikalisch-Technische Bundesanstalt (PTB) and INP –Leibniz Institute for Plasma Science and Technology are going to propose a "Fachinformationsdienst Physik", a specialised information service. This infrastructure supporting research aims to facilitate researchers to access specialised literature und research-specific information and to provide services based on high quality physics metadata. We are going to follow a holistic approach, taking into account all forms of media, including, but not limited to research data.

Our mission will be to collect, connect and correct (meaning improving and maintaining the quality) of metadata for physics. A centrepiece of our proposed activities is going to be awareness raising within the physics research community in order to contribute to the groundworks of FAIR physics. We envision the Fachinformationsdienst Physik to foster a sustained engagement of the relevant stakeholders, informing the subsequent development of metadata-driven research services tailored to cutting-edge research in physics.

Please assign your contribution to one of the following topics

Enabling and incentivising the research community

Please specify "other" (stakeholder)

In addition please add keywords.

metadata
physics
libraries
terminologies
research-supporting infrastructures

Please assign yourself (presenting author) to one of the stakeholders.

Data professionals who provide and maintain data infrastructure

Primary author: ISRAEL, Holger (TIB - Leibniz-Informationszentrum Technik und Naturwissenschaften)

Co-authors: Dr TOBSCHALL, Esther (TIB - Leibniz-Informationszentrum Technik und Naturwissenschaften); HOFFMANN, Julia (TIB - Leibniz-Informationszentrum Technik und Naturwissenschaften)

Presenter: HOFFMANN, Julia (TIB - Leibniz-Informationszentrum Technik und Naturwissenschaften)

Session Classification: Poster session

Track Classification: Poster session

Contribution ID: **44** Type: **Talk**

Recent metadata enhancements in the RSpace digital research platform

Wednesday 11 October 2023 09:50 (20 minutes)

Summary

Our presentation reflects the topic 'Facilitating connectivity of research data', in particular the subtopics: 'Metadata annotation and management during and close to the research process', and 'Data interoperability through harmonised metadata and interoperable semantics'.

The presentation describes significant advances in incorporating metadata into the RSpace digital research platform, which comprises an electronic lab notebook integrated into a sample management system. The presentation will present highlights from the just completed Interoperability Guideline, which reports on the Enhancing Interoperability through Incorporation of PIDS in Tools project done jointly with Research Space and DataCite under an EOSC Future –RDA grant. In addition to describing incorporation of the new IGSN ID's into the RSpace sample management system, the report also includes general guidelines for incorporating IGSN IDs into research tools. Thus the presentation also contributes to the third conference topic, 'Transforming (meta)data recommendations into implementations'.

We also cover recent enhancements for controlled vocabularies/ontologies in RSpace.

Overview

The EOSC Future RDA project involved interaction with researchers and research administrators from UiT the Arctic University of Norway and Rothamsted Resarch to understand their workflows and requirements for use of IGSNs in the context of RSpace, and touched on pain points related to interoperability, PIDs and collaboration in general.

The project work was informed by the following design principles:

- Ensure shared understanding of roles and responsibilities that comes with the realisation of interoperability - particularly regarding metadata creation and management
- Define PID integration goals based on use cases
- Reuse existing metadata frameworks, local and general
- Leverage the open infrastructure to fortify data management workflow
- Work with the disciplinary communities to define best practices

In the presentation we describe the support for a basic IGSN workflow implemented in RSpace, including how we integrated with DataCite for handling IGSN registering and publishing actions. This resulted in a fully working prototype that enables IGSN registration, metadata entry, and publishing all within RSpace Inventory. RSpace now supports:

- Identifier section present on each sample, subsample and container
- Register an IGSN for a sample
- Delete a draft IGSN
- Fill in IGSN mandatory metadata fields
- Fill in IGSN recommended fields: subject, description, date, alternate identifier
- Preview landing page
- Publish an IGSN
- Generate public RSpace landing page with metadata
- Re-publish with updated metadata
- Retract a published item
- Publish a retracted item

We also briefly describe ongoing work to improve the supported workflows.

Finally, we cover enhancements for controlled vocabularies/ontologies in RSpace, including associating metadata with ontologies, enforcing ontologies, and importing domain standards ontologies, e.g. from BioPortal Ontologies. Also, we have enhanced interoperability of metadata and tools by automatically including tags as vocabulary term URIs when exporting an RSpace data archive to a repository such as Dataverse or Zenodo. This ensures that richness of metadata is preserved between tools used in the active research phase and archival phase, enhancing findability.

Please assign your contribution to one of the following topics

Metadata annotation and management close to the research process

Please specify ”other” (stakeholder)

In addition please add keywords.

metadata, ontologies, interoperability, IGSNs

Please assign yourself (presenting author) to one of the stakeholders.

Data professionals who provide and maintain data infrastructure

Primary authors: MACNEIL, Rory (Research Space); PLANKYTE, Vaida (Research Space)

Presenter: MACNEIL, Rory (Research Space)

Session Classification: Parallel Track 2

Track Classification: Facilitating connectivity of research data: Metadata annotation and management during and close to the research process

Contribution ID: **45** Type: **Poster**

Towards a Helmholtz Data Space - adjusting responsibilities for metadata data by the use of PIDS

Tuesday 10 October 2023 14:30 (15 minutes)

Persistent identifiers (PIDs) are an integral element of the FAIR principles (Wilkinson et al. 2016) as they are recommended to refer to data sets and metadata. They are, however, also considered to be used to refer to other data entities, like people, organizations, projects, laboratories, repositories, publications, vocabularies, samples, instruments, licenses, methods and others. Consistently integrating these PIDs into data infrastructures can create a high level of interoperability allowing to build connections between data sets from different repositories according to common meta information.

Enhanced data acquisition and maintenance, however, requires new models of responsibility for these datasets. The roles of data maintainers extend far beyond the current actors, who are researchers, data managers, and librarians. In fact technicians, center administration and management, center employees, and others do have an important role making sure, their metadata is properly referenced, uniform and reliably maintained.

Here we shed some light what different PID systems we recommend to implement within the Helmholtz Association and make suggestions, which stakeholder groups should be included to take responsibility for maintaining them, in order to shape the Helmholtz Data Space.

Please assign your contribution to one of the following topics

Bringing recommendations closer to practice

Please specify ”other” (stakeholder)

In addition please add keywords.

PID, Interoperability, Recommendation, Data Space, Roles

Please assign yourself (presenting author) to one of the stakeholders.

Data professionals and stewards

Primary authors: SÖDING, Emanuel (GEOMAR); PÖRSCH, Andrea (HMC Hub EE at GFZ); Mrs KOTTMEIER, Dorothee (Pangaea / AWI); RAZEGHI, Yousef

Presenter: SÖDING, Emanuel (GEOMAR)

Session Classification: Poster session

Track Classification: Poster session

Contribution ID: **46** Type: **Talk**

Efficiency and Integrity: Harnessing registry.awi.de for Enhanced Metadata Management in Environmental Research

Tuesday 10 October 2023 10:50 (20 minutes)

In the field of polar and marine environmental research, a diverse array of items including instruments, platforms, models, custom-built facilities, and lab equipment are routinely employed. This results not only in substantial volumes of collected data, but also generates a wealth of accompanying metadata.

In this talk, we highlight the practical utility of registry.awi.de —an authoritative repository within the o2a framework —dedicated to managing essential information concerning platforms, devices, and sensors. We showcase registry.awi.de's efficacy through practical applications in everyday tasks, underscoring its role in preserving item provenance, streamlining data processing, and serving as a central knowledge repository.

Moreover, the study underscores the potential benefits derived from automated workflows and various methods of interfacing with the registry. This multifaceted approach not only simplifies administrative aspects but also enhances overall research efficiency, ensuring the integrity and accessibility of critical metadata while harnessing the advantages of data automation.

Please assign your contribution to one of the following topics

Metadata annotation and management close to the research process

Please specify "other" (stakeholder)

In addition please add keywords.

automation
information system
best practice

Please assign yourself (presenting author) to one of the stakeholders.

Data professionals who provide and maintain data infrastructure

Primary author: ANSELM, Norbert (Alfred-Wegener-Institut)

Co-authors: KOPPE, Roland (AWI); IMMOOR, Sebastian; SCHAEFER, Angela (Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung)

Presenter: ANSELM, Norbert (Alfred-Wegener-Institut)

Session Classification: Parallel Track 1

Track Classification: Facilitating connectivity of research data: Metadata annotation and management during and close to the research process

Contribution ID: **47** Type: **Poster**

The Community Platform HMC Earth and Environment - current status

Tuesday 10 October 2023 14:30 (15 minutes)

In this presentation we will introduce to the current HMC activities and outcome in HUB Earth and Environment: Our process for developing a guideline is planned as a coordinated procedure. For every single implementation guide, we go through the same questions, up to tests - based on use cases and definition of abstract test classes, in order to be able to validate the implementation. Our planned results are recommendations and detailed implementation instructions that enable interoperability - not only in the Helmholtz Association, but also in national and international communities.

Please assign your contribution to one of the following topics

Bringing recommendations closer to practice

Please specify ”other” (stakeholder)

HMC Hub E&E

In addition please add keywords.

Guideline

Please assign yourself (presenting author) to one of the stakeholders.

other (please specify)

Primary authors: PÖRSCH, Andrea (HMC Hub EE at GFZ); KOTTMEIER, Dorothee (Pangaea / AWI); SÖDING, Emanuel (GEOMAR); RAZEGHI, Yousef

Presenter: PÖRSCH, Andrea (HMC Hub EE at GFZ)

Session Classification: Poster session

Track Classification: Poster session

Contribution ID: **48** Type: **Hands-on session**

A web based tool for Import, Mapping and Harmonization of Tabular Data (applied to clinical data)

Wednesday 11 October 2023 14:06 (3 minutes)

The field of clinical research data is rich in information that is often underutilized due to the complexity resulting from heterogeneous representations of the data and the lack of suitable tooling for its harmonization. Ineffective data preprocessing hinders potential insights and prevents effective reuse and combination of data that could otherwise drive progress in the scientific field by providing additional evidence. To counter this problem, we propose a web application designed to assist users in harmonizing non-standardized tabular data.
It enables the seamless import of data in common file formats such as CSV, XLSX, or XLS and allows exporting the harmonized data either as CSV or directly uploading it to REST APIs. Because of its versatility, it can also be useful to address the harmonization problem in other scientific disciplines. Imported data can be mapped and validated against a JSON-Schema. As long as the intended data structure can be articulated in the form of a JSON-Schema, it can be incorporated and utilized in the system. Furthermore, complex transformations on the data can be interactively developed and performed during the import process by utilizing JavaScript.
By taking into account the specifics of the research data lifecycle, our tool provides comprehensive support to researchers. In particular, it facilitates essential steps such as pre-processing, validation, and optional data migration. The migration process allows users to map columns of the table to a given JSON schema. Providing these specialized functions not only simplifies data processing but also ensures data longevity that can be effectively adapted to an evolving research environment.

Consequently, this web application is a promising tool for improving data use across a wide range of scientific disciplines. It offers features that serve important functions in a variety of research areas in general. However, it is important to note that while the tool has the potential to meet the use cases and goals outlined, a comprehensive evaluation of its full capabilities in various real-world scenarios has not yet been conducted. Its performance, especially when processing large datasets, the potential security concerns related to JavaScript transformations, and the ability to meet all predefined requirements need further investigation. The next step in the development is to develop use cases for a more detailed evaluation. In its current state, the tool provides a foundation that can help researchers from numerous disciplines to harmonize data and reduce the overhead associated with redundant data collection by leveraging multiple data sources.

Please assign your contribution to one of the following topics

Data interoperability through harmonised metadata and interoperable semantics

Please specify ”other” (stakeholder)

In addition please add keywords.

HMC, Tool, Harmonization, Clinical-data

Please assign yourself (presenting author) to one of the stakeholders.

Researchers

Primary author: KULLA, Lucas (DKFZ)

Co-authors: MAIER-HEIN, Klaus (DKFZ); NOLDEN, Marco (DKFZ); SCHADER, Philipp (DKFZ)

Presenter: KULLA, Lucas (DKFZ)

Session Classification: HMC Hands-on Session

Contribution ID: **49** Type: **Poster**

Understanding the NeXus standard from an information science perspective?

Tuesday 10 October 2023 14:30 (15 minutes)

NeXus is a well established standard for data exchange of neutron, x-ray and muon large scale facilities. Being around for over 20 years with dedicated governance structures it serves as a successful example of a long-lived standard. NeXus as an ecosystem can be difficult to navigate as people refer to its parts using varying terminology and sometimes having different concepts in mind even when using similar terms. The NeXus community can benefit greatly by connecting and later aligning their established semantic and procedural approaches with standardised definitions from information science, particularly around metadata standards and controlled vocabularies.

This Poster connects Nexus' different parts to some of the more consistent taxonomies of metadata standards and vocabularies and highlight NeXus' strength of allowing for top-down and bottom-up evolution of the standard.

Please assign your contribution to one of the following topics

Data interoperability through harmonised metadata and interoperable semantics

Please specify "other" (stakeholder)

In addition please add keywords.

NeXus, Metadata Standard, Data Format Standard, Information Science

Please assign yourself (presenting author) to one of the stakeholders.

Data professionals and stewards

Primary author: WALTER, Konstantin Pascal (HZB (Hub-Matter))

Presenter: WALTER, Konstantin Pascal (HZB (Hub-Matter))

Session Classification: Poster session

Track Classification: Poster session

Contribution ID: **50** Type: **Talk**

Making your samples FAIR –tools and recommendations from the FAIR WISH Project

Wednesday 11 October 2023 09:30 (20 minutes)

FAIR WISH - FAIR Workflows to establish IGSN for Samples in the Helmholtz Association is an HMC funded project of the first cohort 2020. IGSN, the International Generic Sample Number, is a globally unique, citable and persistent identifier (PID) for physical samples with discovery functionality in the internet. IGSNs enable direct links between data, publications and the originating samples and thus close one of the last gaps in the full provenance of research results.
The main outcome of FAIR WISH is the FAIR SAMPLES Template. This modular template, developed for the Earth and Environmental domain, allows users, i.e. individual researchers, to select metadata properties based on their sample type and create customised sample descriptions. It includes a number of linked-data vocabularies for enriching the descriptions in a standardised form and is the basis of the semi-automated XML generation of the IGSN metadata XMLs. These XMLs are used for batch upload to the IGSN/ DataCite server for IGSN registration, and the source for IGSN landing pages. During the project, we collected and registered rich IGSN metadata for more than 14.000 samples using the FAIR SAMPLES Template. These samples represent the large variety in sample types and sub-disciplines across the three project partners GFZ, AWI and Hereon and all states of digitisation.
The development of the FAIR SAMPLES Template is an important step for the further standardisation and harmonisation of sample metadata and includes a number of linked-data vocabularies for different geo-bio sample types in terrestrial and marine environments. It is specifically designed for individual users with hierarchical samples, but can also be used for the generation of IGSN metadata from digital sample management systems, like the marine Expedition database at Hereon.
Further project results are the full documentation of the IGSN metadata schema and a list of linked-data vocabularies (SKOS/RDF) that are recommended to be included in the IGSN metadata enabling further standardisation. Since 2023, IGSNs are registered as DataCite DOIs which required an initial mapping of IGSN metadata to the DataCite Schema.

Please assign your contribution to one of the following topics

Please specify "other" (stakeholder)

In addition please add keywords.

Please assign yourself (presenting author) to one of the stakeholders.

Scientists and technicians who maintain and operate research infrastructure for data generation

Primary authors: ELGER, Kirsten (GFZ German Research Centre for Geosciences); BALDEWEIN, Linda (Helmholtz-Zentrum Hereon); BRAUSER, Alexander (Deutsches GeoForschungsZentrum (GFZ) Potsdam); FRENZEL, Simone (GFZ German Research Centre for Geosciences); HELM, Birgit (Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research); KLEEBERG, Ulrike (Hereon); LEEFMANN, Tim (Helmholtz Zentrum Hereon); WIECZOREK, Mareike (Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung); NORDEN, Ben (GFZ German Research Centre for Geosciences)

Presenter: ELGER, Kirsten (GFZ German Research Centre for Geosciences)

Session Classification: Parallel Track 1

Track Classification: Facilitating connectivity of research data: Data interoperability through harmonised metadata and interoperable semantics

Contribution ID: **52** Type: **Talk**

SECoP@HMC

Wednesday 11 October 2023 10:10 (20 minutes)

The Sample Environment Communication Protocol (SECoP) provides a generalized way for controlling measurement equipment –with a special focus on sample environment (SE) equipment 1. In addition, SECoP holds the possibility to transport SE metadata in a well-defined way.

SECoP is designed to be
- simple to use,
- inclusive concerning different control systems and control philosophies,
- self-explaining providing a machine readable description of the available data and metadata.

Within the project SECoP@HMC, we are developing and implementing metadata standards for typical SE equipment at large scale facilities (photons, neutrons, high magnetic fields). A second focus is the mapping of the SECoP metadata standards to a unified SE vocabulary for a standardized metadata storage. Thus, a complete standardized system for controlling SE equipment and collecting and saving SE metadata will be available and usable in the experimental control systems of the participating facilities. This approach can be applied to other research areas as well.

In this presentation we will report on the current status of the project SECoP@HMC.

Please assign your contribution to one of the following topics

In addition please add keywords.

Please assign yourself (presenting author) to one of the stakeholders.

Please specify ”other” (stakeholder)

Primary author: KIEFER, Klaus (Helmholtz-Zentrum Berlin)

Presenter: KIEFER, Klaus (Helmholtz-Zentrum Berlin)

Session Classification: Parallel Track 1

Contribution ID: **53** Type: **Poster**

Is your data policy findable and accessible? Check these four elements.

Tuesday 10 October 2023 14:30 (15 minutes)

This research poster dives into the important impact of four simple but crucial elements in research data policies: clear titles, persistent identifiers, publication dates, and open availability. These elements, often underestimated in policy, play a pivotal role in enhancing data discoverability, transparency, and collaboration - ultimately strengthening the foundation of modern scientific inquiry.
Upon examining the institutional data policies of institutes in research field matter (both inside and outside Helmholtz) we realized the essential role these four elements played in driving change through policy. We also understood that due to their simplicity they could be easily overlooked.
In a data-centric era, the integration of these elements aligns policies with discovery, transparency, and accessibility principles. As data's importance grows, acknowledging these nuances is key/ becomes pivotal for guiding scientific progress.

Please assign your contribution to one of the following topics

In addition please add keywords.

Please assign yourself (presenting author) to one of the stakeholders.

Please specify ”other” (stakeholder)

Presenter: ÖZKAN, Özlem (Helmholtz Berlin)

Session Classification: Poster session

Track Classification: Poster session

Contribution ID: **55** Type: **Talk**

HMC Project: Metamorphoses

Wednesday 11 October 2023 11:30 (20 minutes)

Currently the amount and diversity of high-quality atmospheric remote sensing observations from satellites is quickly increasing, and their synergetic use offers unprecedented knowledge gaining opportunities. FAIR data are important for this kind of data interoperability and reusability. This project will lead to FAIR satellite data products. It will develop metadata standards for describing the vertical sensitivity of the remote sensing data (essential for sophisticated data reuse like synergetic data merging) and develop and apply a tool for matching different satellite data using time and space metadata. In this presentation we show the progress after the first year.

Please assign your contribution to one of the following topics

In addition please add keywords.

Please assign yourself (presenting author) to one of the stakeholders.

Please specify ”other” (stakeholder)

Presenter: SHAHZADI, Kanwal (KIT)

Session Classification: Parallel Track 2

Contribution ID: **56** Type: **not specified**

1st release of the Helmholtz Knowledge Graph: meet and greet the developers

Wednesday 11 October 2023 14:00 (3 minutes)

Research across the Helmholtz Association is based on inter- and multidisciplinary collaborations across its 18 Centres and beyond. However, the (meta)data generated through Helmholtz research and operations is typically siloed within institutional infrastructures and often within individual teams. The result is that the wealth of the association's (meta)data is stored in a scattered manner, hard to find and consequently cannot be used to its full value to scientists, managers, strategists, and policy makers.
To address this challenge, the Helmholtz Metadata Collaboration (HMC) launched the unified Helmholtz Information and Data Exchange (unHIDE) in 2022. We are creating a lightweight and sustainable interoperability layer to interlink data infrastructures; and increase visibility and access to the Helmholtz Association's (meta)data and information assets. Using proven and globally adopted knowledge graph technology, within unHIDE we develop a comprehensive association-wide knowledge graph (KG) the Helmholtz-KG: a solution to connect (meta)data, information, and knowledge.
A first prototype of the Helmholtz KG was released in April 2023. This includes a comprehensive web front end for manual search of resources 1, a stable and documented 2 backend with a tested data ingestion and integration pipeline, and machine accessible endpoints 3.

1 https://search.unhide.helmholtz-metadaten.de/

2 https://docs.unhide.helmholtz-metadaten.de/

3 https://sparql.unhide.helmholtz-metadaten.de/

Presenter: BRÖDER, Jens (Forschungszentrum Jülich GmbH (IAS-9))

Session Classification: HMC Hands-on Session