Helmholtz Metadata Collaboration | Conference 2024

Report of Contributions

Contribution ID: **68** Type: **POSTER&PITCH**

The CREATIVE project - customising a generic repository with domain-specific metadata

Monday 4 November 2024 15:00 (1 hour)

The CREATIVE project aims to make the generic repository RADAR4KIT easily accessible and attractive for the domain-specific communities organized in the Climate and Environment Centre (CEC) at the Karlsruhe Institute of Technology (KIT). This aim will be achieved with the help of customized templates and input masks for subject-specific metadata, which enhance the RADAR4KIT usability for the CEC scientists and thus facilitate data publication beyond the generic functionalities of the repository.

At the same time, the subject-specific metadata schemas are harmonised with the schemas used by the NFDI4Earth, the National Research Data Infrastructure (NFDI) for Earth System Sciences (ESS), and the virtual research environment V-FOR-WaTer, which is being developed at KIT in a collaboration between, mainly, hydrologists and computer scientists. This harmonisation effort in combination with corresponding interfaces enable the domain-specific (meta) data to be included in the data base of NFDI4Earth and promotes broader use of the data beyond KIT. Additionally, by implementing the interfaces and the standardized metadata description, the CEC researchers can use V-FOR-WaTer to pre-process, edit and visualize the data, thus accelerating scientific work with the data sets and their interdisciplinary use. RADAR4KIT then functions as an adapted specialist repository for the CEC institutes and connects meaningfully to other initiatives. The adaptation steps for the templates can be transferred to other specialist areas at KIT and via detailed documentation and publication with NFDI4Earth also in other domains in the ESS.

Another key focus of the CREATIVE project is connecting and supporting the data stewards at the CEC institutes. For the most part there are no designated positions for this task, so usually researchers and PhD students take on the role unofficially and out of necessity - often without a clear overview of existing research data management (RDM) structures, tools and support. CREATIVE aims to facilitate the exchange between data stewards using the example of publishing data via RADAR4KIT, supported by the CREATIVE team. Several workshops during the development of the metadata templates and during the testing phase are meant to further strengthen this exchange and establish a network of data stewards, thus, supporting sustainable and future-proof research data management (RDM) at KIT.

Please specify ”other”

In addition, please add 3 to 5 keywords.

metadata template, data stewards, domain-specific, harmonisation

Please specify ”other”

For whom will your contribution be of most interest?

Data professionals and stewards

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

Data professionals and stewards

Primary author: HASSLER, Sibylle (Institute of Meteorology and Climate Research –Atmospheric Trace Gases and Remote Sensing, Karlsruhe Institute of Technology, Karlsruhe, Germany)

Co-authors: Mr ZULETA SALMON, Carlos (Institute of Meteorology and Climate Research –Atmospheric Trace Gases and Remote Sensing, Karlsruhe Institute of Technology, Karlsruhe, Germany); MEYER, Jörg (Scientific Computing Center (SCC) –Data Analytics, Access and Aplications, Karlsruhe Institute of Technology, Karlsruhe, Germany); BRAESICKE, Peter (Institute of Meteorology and Climate Research –Atmospheric Trace Gases and Remote Sensing, Karlsruhe Institute of Technology, Karlsruhe, Germany); ZEHE, Erwin (Institute for Water and Environment –Hydrology, Karlsruhe Institute of Technology, Karlsruhe, Germany)

Presenter: HASSLER, Sibylle (Institute of Meteorology and Climate Research –Atmospheric Trace Gases and Remote Sensing, Karlsruhe Institute of Technology, Karlsruhe, Germany)

Session Classification: Poster Session B

Track Classification: Connecting research data: 4. Metadata annotation and management

Contribution ID: **69** Type: **TALK**

Introducing the Open Data Format: A New FAIR-Compliant Data Format using DDI Codebook

Monday 4 November 2024 12:05 (20 minutes)

Currently, social scientists use different and sometimes proprietary software to analyse data, which processes metadata in diverse ways. Data formats of statistical software packages are only partially compatible and pose an obstacle to replication studies. Proprietary data formats jeopardise the requirement for interoperability enshrined in the FAIR principles. As part of KonsortSWD, we developed an open data format that can be used in common statistical programs to improve the exchange and reuse of research data and enable access to documentation materials directly via the statistical software. The open data format can be enriched with multilingual metadata and links to data portals, enabling direct access to documentation materials through the statistical software itself. In this paper, we present a practical use case of the Open Data Format based on SOEP panel data. We show how datasets in the ODF look, how the SOEP data was converted into our open data format, and how ODF data files can be imported and exported in R and Stata.

In addition, please add 3 to 5 keywords.

Data Format, FAIR, Open Data, DDI Codebook

Please specify ”other”

For whom will your contribution be of most interest?

Data professionals who provide and maintain data infrastructure

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

Data professionals who provide and maintain data infrastructure

Please specify ”other”

Primary authors: WENZIG, Knut (DIW Berlin/SOEP); HARTL, Tom (DIW Berlin); HAN, Xiaoyao (DIW Berlin)

Presenters: HARTL, Tom (DIW Berlin); HAN, Xiaoyao (DIW Berlin)

Session Classification: Session B2

Track Classification: Connecting research data: 4. Metadata annotation and management

Contribution ID: **71** Type: **TALK**

FAIRlead - A Domain Independent, Model Driven Approach to follow the FAIR Data Avenue

Monday 4 November 2024 10:50 (20 minutes)

Enriching data with describing metadata is the key-enabler for the reusability and interoperability of experimental results and thus to further research in a scientific domain. However, in order to be able to use data of former scientific work (both initial data and result data from experiments), a common understanding of the semantics of this data is essential. This understanding is typically achieved by developing or using domain-specific ontologies or conceptual models that describe the relevant entities of a domain and their relationships to each other. However, depending on the domain (photovoltaic, autonomous driving, etc.), the ontologies are different. Accordingly, the metadata in the various domains has a different structure. This also means that for the manual or (semi) automatic acquisition of metadata, each area requires its own tool to record the metadata and link it to the actual data. This is where we want to start with our research by developing a model-driven software development (MDSD) approach that enables us to provide tools for capturing metadata and linking it to the actual data. In the MDSD, source code is generated based on an input model and transformation rules. Specifically, in our case, the code generator receives an ontology and a series of templates describing the mapping of the model information to the target language or platform (PHP, Java, Django, . . .) as input. The functionality of the software to be generated includes an interactive interface for the manual collection of metadata (i.e. experimental setup of a photovoltaic system), a REST-API for the programmatic collection of metadata and as an extension point to embed application logic, an associated persistence component, as well as suitable search and export functionalities. Furthermore, a connection to the respective data is essential, whereby the widest possible range of data formats should be supported here. The non-functional requirements for the software include user-friendliness, easy integration into existing environments, expansion options, and no commitment to a specific target programming language.

Another important aspect of FAIRlead is the integration of existing data towards the domain ontology. FAIRlead shall assist users in both extending the domain ontology and reusing concepts from other existing ontologies to improve interoperability. The resulting combined ontology serves as input to the generator. To form a knowledge graph with the ontology and existing data sources, we integrate both data and metadata within a conceptual model diagram of the relevant sources (including APIs, files, databases and manual user input). To facilitate this, we use an approach to extract the conceptual model from existing (meta)data. The extracted conceptual model can be fine tuned by the user using a visual editor. It then allows metadata to be added to its respective data with a visual graph-based user interface. The generated code will also allow direct access to the integrated data in the knowledge graph.
As part of the presentation, we will go into the basic concepts for the development of the generator and also present our roadmap for the following steps in this area.

In addition, please add 3 to 5 keywords.

Code Generation, Metadata, Ontology based Engineering, FAIR

Please specify ”other”

For whom will your contribution be of most interest?

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

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

Researchers

Please specify ”other”

Primary authors: SCHMIDT, Andreas (IT4EDM/IAI); Mr KOUBAA, Mohamed Anis (Institute for Automation and applied Informatics); LIU, Nan; SCHMURR, Philipp (KIT IAI); STUCKY, Karl-Uwe (KIT); SUESS, Wolfgang (KIT)

Presenter: SCHMIDT, Andreas (IT4EDM/IAI)

Session Classification: Session A2

Track Classification: Connecting research data: 6. Interoperable semantics at domain and application level

Contribution ID: **72** Type: **POSTER&PITCH**

Presenting a prototype platform for mapping epidemiological cohort metadata with environmental and earth observation metadata: the MetaMap^3 project

Monday 4 November 2024 14:00 (1 hour)

Introduction: The environment plays an important role for human health and efficient linkage of epidemiological cohorts with environmental data is crucial to quantify human exposures. However, there are no harmonized standards for automatic mapping of metadata of our three domains Health (HMGU), Earth & Environment (UFZ), and Aeronautics, Space & Transport (DLR).

Objective: We aimed to ease the search for appropriate exposure data for subsequent epidemiological analyses by generating and enriching interoperable and machine-readable metadata for exemplar cohort and exposure data and by mapping these metadata so that they can be jointly queried and searched.

Methods: Our use case for cohort metadata consisted of the two German prospective, population-based birth cohorts GINIplus and LISAplus which were conducted in four study regions across Germany and included up to seven examination rounds since 1995. As use cases for environmental data, we considered land cover classification available for a single time point (06/2015-09/2017 mean) and daily soil moisture index data for the entire study period. Since both factors influence air temperature, they are of specific interest when investigating the effects of temperature extremes on human health. We reviewed several standards, strategies and tools and developed an approach to align the heterogenous metadata to a common structure and format.

Results: We identified spatial and temporal 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 international standard for geographic information ISO 19115 and to the eXtensible Markup Language (XML). Based on our conceptual work, we identified the catalog application GeoNetwork as the best tool for our application. After setting up a test instance on a local server for our use cases metadata, the catalog can now be accessed at https://envepi.helmholtz-munich.de/geonetwork/.

Discussion: We are continuously populating the mapping platform with further metadata. Also, we are currently testing the full functionality of the tool, especially the filtering and search options of the application to enable the intended mapping.

This work was funded by the Helmholtz Association's Initiative and Networking Fund (INF): ZT-I-PF-3-018

Please specify "other"

In addition, please add 3 to 5 keywords.

health, environment, earth observation, metadata mapping, metadata catalog

Please specify "other"

For whom will your contribution be of most interest?

Researchers

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

Researchers

Primary authors: WOLF, Kathrin (Helmholtz Zentrum München GmbH - German Research Center for Environmental Health, Institute of Epidemiology, Neuherberg, Germany); DALLAVALLE, Marco (Helmholtz Zentrum München GmbH - German Research Center for Environmental Health, Institute of Epidemiology, Neuherberg, Germany); 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); STANDL, Marie (Helmholtz Zentrum München GmbH - German Research Center for Environmental Health, Institute of Epidemiology, Neuherberg, Germany); BUMBERGER, Jan (Helmholtz Centre for Environmental Research –UFZ, Research Data Management (RDM), Smart models and Monitoring, Leipzig); 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)

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

Session Classification: Poster Session A

Track Classification: Connecting research data: 5. Technical solutions for findable and machine-readable metadata

Contribution ID: **74** Type: **TALK**

Building FAIR image analysis pipelines for high-content-screening (HCS) data using Galaxy

Tuesday 5 November 2024 13:00 (20 minutes)

Imaging is crucial across various scientific disciplines, particularly in life sciences, where it plays a key role in studies ranging from single molecules to whole organisms. However, the complexity and sheer volume of image data, especially from high-content screening (HCS) experiments involving cell lines or other organisms, present significant challenges. Managing and analysing this data efficiently requires well-defined image processing tools and analysis pipelines that align with the FAIR principles—ensuring they are findable, accessible, interoperable, and reusable across different domains.
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.
Galaxy, an open-source, web-based platform for data-intensive research, offers a solution by enabling the construction of reproducible pipelines for image analysis. By integrating popular analysis software like CellProfiler and connecting with cloud services such as OMERO and IDR, Galaxy facilitates the seamless access and management of image data. This capability is particularly valuable in bioimaging, where automated pipelines can streamline the handling of complex metadata, ensuring data integrity and fostering interdisciplinary collaboration. This approach not only increases the efficiency of HCS bioimaging but also contributes to the broader scientific community's efforts to embrace FAIR principles, ultimately advancing scientific discovery and innovation.
In the present study, we proposed an automated analysis pipeline for storing, processing, analysing, and sharing HCS bioimaging data. The (semi)automatic workflow was developed by taking as a case study a dataset of zebrafish larvae and cell lines images previously obtained from an automated imaging system generating data in an HCS fashion. In our workflows, 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 a novel OMERO tool suite developed with GALAXY. 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 specify "other"

In addition, please add 3 to 5 keywords.

OMERO, Galaxy, FAIR, workflow, HCS

Please specify ”other”

For whom will your contribution be of most interest?

Researchers

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

Data professionals and stewards

Primary author: MASSEI, Riccardo

Co-authors: Dr NYEFFLER, Jo (UFZ); Dr BERNDT, Matthias (UFZ); SCHOLZ, Stefan (UFZ - ETOX); Dr DUNKER, Susanne (iDiv); BUSCH, Wibke; BUMBERGER, Jan; BOHRING, Hannes

Presenter: MASSEI, Riccardo

Session Classification: Session E2

Track Classification: Connecting research data: 5. Technical solutions for findable and machine-readable metadata

Contribution ID: **75** Type: **TALK**

Tracking the cell –metadata for single-cell genomics in biomedicine [CellTrack]

Monday 4 November 2024 11:45 (20 minutes)

As the volume of omics single-cell data continues to grow, so too must our data management and processing capabilities to ensure its effective secondary use, particularly in research and diagnostics. While single-cell data holds immense potential for AI applications, current documentation standards fall short of being AI-ready. To address these challenges, we organized a Writathon, resulting in a comprehensive white paper.

The outcomes of this project will directly contribute to significant national and international infrastructure initiatives, most notably the German Human Genome-Phenome Archive (GHGA), a national genomics platform funded by the NFDI, and scverse, a community-driven framework dedicated to develop core infrastructure and interoperable software for essential analytics in single-cell genomics.

Please specify ”other”

In addition, please add 3 to 5 keywords.

Single-cell, Metadata, AI-ready, Omics, Big data

Please specify ”other”

For whom will your contribution be of most interest?

Data professionals and stewards

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

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

Primary author: HEYL, Florian (The German Cancer Research Center (DKFZ))

Co-authors: Prof. THEIS, Fabian (Institute of Computational Biology, Department of Computational Health, Helmholtz Munich; School of Computation, Information and Technology, Technical University of Munich;TUM School of Life Sciences Weihenstephan, Technical University of Munich); Prof. STEGLE, Oliver (The German Cancer Research Center (DKFZ), Division of Computational Genomics and Systems Genetics, Heidelberg, Germany; German Human Genome-Phenome Archive (GHGA),)

Presenter: HEYL, Florian (The German Cancer Research Center (DKFZ))

Session Classification: Session B2

Track Classification: Connecting research data: 4. Metadata annotation and management

Contribution ID: **76** Type: **POSTER&PITCH**

Datathons: fostering equitability in data reuse in ecology

Monday 4 November 2024 15:00 (1 hour)

Approaches to rapidly collecting global biodiversity data are increasingly important, but biodiversity blind spots persist. We organized a three-day Datathon event to improve the openness of local biodiversity sequence data and facilitate data reuse by local researchers. The first Datathon, organized among microbial ecologists in Uruguay and Argentina assembled the largest microbiome dataset in the region to date and formed collaborative consortia for microbiome data synthesis.

Please specify ”other”

In addition, please add 3 to 5 keywords.

Sequence data, metadata enrichment, community engement, capacity building

Please specify ”other”

Interesting for researchers, data professionals and stewards who maintain data infrastracture, and stakeholders

For whom will your contribution be of most interest?

other (please specify below)

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

Researchers

Primary author: JURBURG, Stephanie (UFZ)

Presenter: JURBURG, Stephanie (UFZ)

Session Classification: Poster Session B

Track Classification: From recommendations to implementations: 10. Enabling and incentivising community-driven initiatives

Contribution ID: **77** Type: **POSTER&PITCH**

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

Monday 4 November 2024 14:00 (1 hour)

The **N**uclear, **A**stro, and **P**article **M**etadata **I**ntegration for e**X**periments (NAPMIX) project was recently awarded funding within the scope of the OSCARS call on open science and will start in December 2024. The project aims to facilitate data management and data publication under the FAIR principles on the European level by developing a cross-domain metadata schema and generator, tailored for diverse datasets. It focuses on creating a standardised, adaptable framework that enhances FAIR data. By creating a comprehensive and adaptable metadata schema, the project will ensure scalable integration of both machine and human-readable metadata, thereby improving the efficiency of data discovery and utilization.

A pivotal component of the scheme is its nodal, multi-layered schema structure, allowing metadata enrichment from multiple domains while maintaining essential overlaps for enhanced versatility. This comprehensive approach supports the unification of data standards across various research institutions, promoting interoperability and collaboration on a European scale. Our efforts will extend to the development of a user-friendly frontend generator, designed not only to facilitate metadata input but also to allow users to specify field-specific attributes, customize generic names to suit their needs, and export schemas in various formats such as JSON and XML, adhering to different nomenclatures. In addition, API's will be developed to enable automated metadata generation.

The project involves RIs and ESFRIs, and leverages synergies from existing Open Science initiatives like EOSC, ESCAPE, EURO-LABS, and PUNCH4NFDI. In this contribution, we will present an overview of the project, detailing the development steps, key features of the metadata schema, and the functionality of the frontend generator.

As a starting point, a pilot study has been completed at GSI to develop a backend database structure. This poster will describe the development steps of the pilot study and discuss the planned implementation of the NAPMIX project.

Please specify "other"

In addition, please add 3 to 5 keywords.

Nuclear, Astrophysics, Particle Physics, OSCARS, Schema

Please specify "other"

For whom will your contribution be of most interest?

Data professionals and stewards

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

Data professionals and stewards

Primary author: Dr MISTRY, Andrew. K. (GSI Helmholtzzentrum für Schwerionenforschung GmbH(GSI))

Presenter: Dr MISTRY, Andrew. K. (GSI Helmholtzzentrum für Schwerionenforschung GmbH(GSI))

Session Classification: Poster Session A

Track Classification: Connecting research data: 6. Interoperable semantics at domain and application level

Contribution ID: **78** Type: **TALK**

Project MEMAS: Integrated Data Management for Additive Manufacturing enabling High-Fidelity Modeling

Tuesday 5 November 2024 13:40 (20 minutes)

Predicting the performance of aerospace and automotive structures requires detailed reflection of the actual manufacturing process of each produced part. This is especially the case for composite structures produced with additive manufacturing processes in view of their process complexity and its influence on the product reliability. For high-fidelity numerical models to reflect the actual state of the manufactured structures and cover their individual load-bearing capability, it is essential to consider data across pre-production, production, and post-production stages comprehensively. In this study, we established a robust data acquisition and database infrastructure using the shepard integrated data management system (IDMS) tailored for Robotic Screw Extrusion Additive Manufacturing (RSEAM). Shepard IDMS is designed for storing highly heterogeneous research data adhering to the FAIR principles and offers a consistent API for depositing and accessing various types of supported data. Our data acquisition strategy integrates KUKA Robot Sensor Interface (RSI) and OPC Unified Architecture (OPC UA) protocols for collecting high-frequency time-series data during production. By capturing end-to-end manufacturing data along with associated metadata, we ensure a comprehensive overview of RSEAM activities. Additionally, we developed graphical user interfaces (GUI) in Python using Taipy and Streamlit, streamlining data management including metadata integration and facilitating analysis within this infrastructure. The coupling of the IDMS to a multi-field ontology enables the creation of high-quality and well-documented datasets, which can be converted into predictive numerical models. The contribution will present key solutions for live data acquisition, structuring and storage. The benefit of data enhancement will be highlighted on an exemplary structure.

Please specify ”other”

In addition, please add 3 to 5 keywords.

Additive Manufacturing, Integrated Data Management, Numerical Modeling, Ontology, Composite Materials

Please specify ”other”

For whom will your contribution be of most interest?

Researchers

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

Researchers

Primary author: Mr UNGER, Nicolas (German Aerospace Center (DLR e.V.) Institute of Vehicle Concepts)

Co-authors: Mr KAMBLE, Pradnil (German Aerospace Center (DLR e.V.) Institute of Vehicle Concepts); VINOT, Mathieu (German Aerospace Center (DLR e.V.) Institute of Structures and Design); Mr GLÜCK, Roland (German Aerospace Center (DLR e.V.) Center for Lightweight Production Technology)

Presenter: Mr UNGER, Nicolas (German Aerospace Center (DLR e.V.) Institute of Vehicle Concepts)

Session Classification: Session E2

Track Classification: Connecting research data: 7. Infrastructure and common practices for consolidation of (meta)data

Contribution ID: **79** Type: **POSTER&PITCH**

Results and Insights into the HMC Data Professionals Survey 2024

Monday 4 November 2024 14:00 (1 hour)

The objective of the HMC Data Professionals Survey 2024 is to gain insights into the perspectives, workflows and needs of research data professionals in the Helmholtz Association. The survey is part of HMC's mission to enhance the sustainable management of research data and to more closely align its services with the needs of data professionals at various Helmholtz centers. A comprehensive and anonymized questionnaire is used to collect data from participants across a range of Helmholtz research centers.

The objective of this poster is to present the findings of the survey and provide an overview of the current state of research data management within the Helmholtz Association. The objective of our analysis is to identify common themes and areas for improvement, which will provide a basis for the development of targeted support and resources for data professionals. The analysis will entail customizing the HIFIS-analysis framework, mapping free-text responses and iterating data representations in order to ensure an accurate and comprehensive understanding of the survey data.

The responses of the survey are analyzed using both quantitative and qualitative methods to gain a better understanding of the challenges and opportunities facing data professionals. The analysis will assist HMC in identifying gaps and needs related to metadata practices, training, resource allocation, and collaboration tools.

The results of this survey will be made available to the community to ensure transparency and encourage feedback from participants. Dissemination of these results will facilitate an ongoing dialogue with data experts about the future of metadata management, which will contribute to the advancement of data science and research in the Helmholtz Association.

Please specify "other"

Please specify "other"

In addition, please add 3 to 5 keywords.

Survey, HMC, RDM

For whom will your contribution be of most interest?

Data professionals and stewards

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

Researchers

Primary authors: SCHMIDT, Andreas (IT4EDM/IAI); KULLA, Lucas (DKFZ); KUBIN, Markus (HMC, HZB); SHANKAR, Sangeetha (German Aerospace Center)

Presenter: KULLA, Lucas (DKFZ)

Session Classification: Poster Session A

Track Classification: Assessing and monitoring FAIR data: 1. Human actors in the FAIR data landscape

Contribution ID: **80** Type: **POSTER&PITCH**

From Schema to Questionnaire: Humanizing data description

Monday 4 November 2024 16:00 (1 hour)

Enriching data with metadata is a key concept for the data output of scientific research to be FAIR. Data processing software and custom code often do not support the annotation with metadata out-of-the-box or the usage process does not mandate it. This confronts data creators and maintainers with challenges to annotate their data. From a Human Machine Interface (HMI) perspective, metadata forms are a superior way to support this process compared to manually editing the respective data.
Depending on the use case and the user's role, different ways to generate and use forms for metadata may be helpful. A researcher without programming skills does benefit from ready-to-use services when entering relevant metadata. Software developers might instead prefer a software library that provides forms based on the metadata schema. Managers and those responsible for processes are especially interested in solutions that fit into existing workflows.
For each of the above scenarios we will present a list of criteria to categorize available solutions. For example, a developer might be interested in the particular front end technology that is used by a library, while for a data provider the export formats of a given service are much more relevant. Of course several features do apply to all the use cases, like the capability to allow different profiles of metadata or a real time validation of inputs.
We present a review of available services, tools and software libraries that help with form based metadata annotation. The list will be published on Zenodo to summarize the capabilities in regard to a collection of features that will help to find the most suiting solution. We hope this can empower more people to make their data FAIR.

One solution is Research Data Management Organiser (RDMO), a web application to assist in the structured planning and administration of the data in a scientific project. The interaction with RDMO is based on a set of questionnaires, which can be attached to a project and filled in by data providers.
The so gathered information can be cast into textual forms suitable for funding agencies' report guidelines, or it can be used as metadata for generated datasets.
The integration of RDMO in a data providing and curation workflow starts with generating a questionnaire from a given metadata schema (e.g. the Open Energy Platform metadata standard). Data providers willing to publish on this platform have to answer specific questions concerning their dataset. RDMO additionally applies existing super-ordinate metadata (author and affiliation, funding agency, etc.). Those information sets are inherited from a designated project.
After gathering this information in a schema-compliant file, the provider is assisted in publishing them on a generic metadata store with a download link to the actual data.
The main advantage is that the data provider can focus on answering questions and needs not struggle with machine readable files. Additionally, the data provider is guided through the whole process. File system monitoring can also be used to increase the reliability by motivating the provider to publish newly detected data.

Please specify ”other”

In addition, please add 3 to 5 keywords.

FAIR, Metadata, Questionnaire, Process Reliability

Please specify ”other”

For whom will your contribution be of most interest?

Data professionals and stewards

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

Researchers

Primary authors: Mr KOUBAA, Mohamed Anis (Institute for Automation and applied Informatics); SCHMURR, Philipp (KIT IAI)

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

Presenters: Mr KOUBAA, Mohamed Anis (Institute for Automation and applied Informatics); SCHMURR, Philipp (KIT IAI)

Session Classification: Poster Session C

Track Classification: Connecting research data: 4. Metadata annotation and management

Contribution ID: **81** Type: **TALK**

From scientific terms to linked electronic lab notebooks - A holistic approach

Tuesday 5 November 2024 10:05 (20 minutes)

When gathering your analog research data and metadata, including challenging-to-digitize experimental parameters, aiming at creating a knowledge graph, we suggest the following pipeline for achieving high data quality: agreeing on a shared vocabulary, expanding it to an ontology and eventually semantically annotating the recorded data.
To facilitate this pipeline we developed and use the following tools: The first one, VocPopuli, is the entry point for domain experts. VocPopuli enables the collaborative definition and editing of metadata terms and assigns a PID to each term in the vocabulary, as well as the vocabulary itself. After the export as a SKOS vocabulary or OWL classes, these can be completed to a full ontology by adding relationships. Examples are PolyMat and PolyLab, two domain ontologies for membrane research. Extending ontologies with SHACL shapes creates forms in the Electronic Lab Notebook (ELN) Herbie which automatically make the recorded research data globally understandable and meaningful by semantic enrichment, and ensuring local conformance and completeness of the data by running validations. Managing RDF graphs makes Herbie a platform for collaborative editing of a shared knowledge graph.
However, interdisciplinary research institutions often employ various ELNs systems, creating barriers to metadata exchange. Addressing this interoperability gap, we have developed ELNdataBridge, an API-based data exchange, to enhance interoperability between ELNs, in our case Herbie and Chemotion. ELNdataBridge handles the mapping of the ELN entry fields, as well as the synchronization routine, which can be configured via a user-friendly web interface. An extension to other ELNs can be easily achieved.
Completing the pipeline from vocabulary terms to data exchange between ELNs enables semantically linked and FAIR research metadata and data management, ready for future exploration.

Please specify ”other”

In addition, please add 3 to 5 keywords.

Electronic Lab Notebook, ontology, vocabulary, RDF, SHACL

Please specify ”other”

For whom will your contribution be of most interest?

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

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

Researchers

Primary authors: KIRCHNER, Fabian (Helmholtz-Zentrum Hereon); HELD, Martin (Hereon); ESCHKE, Catriona (Helmholtz-Zentrum Hereon)

Presenters: KIRCHNER, Fabian (Helmholtz-Zentrum Hereon); HELD, Martin (Hereon)

Session Classification: Session C2

Track Classification: Connecting research data: 6. Interoperable semantics at domain and application level

Contribution ID: **82** Type: **POSTER&PITCH**

Improving interdisciplinary research with cross-domain metadata for qualitative data objects

Monday 4 November 2024 15:00 (1 hour)

The aim of a cooperation between the DDI Alliance and QualidataNet - a network for qualitative data that is being created as part of the NFDI - is to describe qualitative data in a standardized way so that researchers can find it and use it for their own research, regardless of discipline and thematic location.

Since last year, QualidataNet has been involved in the metadata developments of the DDI Alliance, which is working on an international metadata standard for Cross Domain Integration to provide integration-ready data, and which should play an important role in the context of the CDIF Framework of the WorldFAIR Project.

Together, we are focusing on the development of a model for the description of qualitative data across disciplines. We understand qualitative data in a broad sense referring to the type or nature of the data objects, i.e., data that are not presented in a quantitative, coded or structured manner. By qualitative data we do not refer to the methods of their collection, generation or analysis, nor to their sensitivity or the quality of their content, but to the (semi)unstructured nature of the data objects before they are processed or analyzed (i.e. raw), in other words, to the **resources that are inputs for analysis** - any observation, measurement or fact represented digitally as free text, images, video, sounds, etc. that form the basis for extracting information.

As responsible for qualitative data and in this cross-domain context, QualidataNet is pursuing the goal of enhancing the interoperability of qualitative data beyond our domains. In order to achieve this, we need to find out how non-numerical or qualitative data should be provided in order to make it ready for analysis. To ensure the integration of a broad variety of examples of qualitative data objects, we are searching for further use cases which will give us an overview of the different types of qualitative data and their usage in different research domains. We are interested in how researchers or data curators work with qualitative data objects, how they prepare them for analysis, and how they combine different data types to represent these procedures in the DDI-CDI metadata model. Use cases are examples of studies or datasets in which qualitative data have been collected and have the potential to be reused in other studies and/or disciplines or combined with other (e.g., quantitative) data.

With our poster we want to get in touch with you, talk about the unmet needs in your areas regarding the description of qualitative data such as social media and sensor/tracking data, audiovisual data, visual maps and deep-sea images, etc. and perhaps lay the groundwork for future collaboration to make non-numeric and qualitative data more interoperable.

Please specify ”other”

In addition, please add 3 to 5 keywords.

qualitative data; interoperability; cross-domain integration; metadata model;

Please specify ”other”

For whom will your contribution be of most interest?

Data professionals and stewards

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

Data professionals and stewards

Primary author: BETANCORT, Noemi (SuUB Bremen, Research Data Center Qualiservice)

Co-author: MOZYGEMBA, Kati (Research Data Center Qualiservice, University of Bremen)

Presenter: BETANCORT, Noemi (SuUB Bremen, Research Data Center Qualiservice)

Session Classification: Poster Session B

Track Classification: Connecting research data: 6. Interoperable semantics at domain and application level

Contribution ID: **84** Type: **POSTER&PITCH**

HARMONise - Towards the sustainable management of metadata associated with marine molecular sequencing data

Monday 4 November 2024 14:00 (1 hour)

Biomolecules, such as DNA and RNA, provide a wealth of information about the distribution and function of marine organisms, and molecular sequencing data from the marine realm is generated across several Helmholtz Centers. Biomolecular (meta)data, 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 (Enhancing the interoperability of marine biomolecular (meta)data across Helmholtz Centres), we developed sustainable solutions 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. A web-based portal enables the standardized export and exchange of core metadata, e.g. with the Marine Data Portal (https://marine-data.de/), which will enhance the findability and accessibility of biomolecular (meta)data within and across research areas. The alignment of HARMONise-hosted metadata with domain-specific standards and the provision of data in the relevant exchange formats will facilitate interoperability with the Helmholtz knowledge graph (UNHIDE, https://docs.unhide.helmholtz-metadaten.de/intro.html) and global digital ecosystems (Ocean Info Hub of the UNESCO Ocean Data and Information System, https://oceaninfohub.org/). HARMONise thus specifically targets the advancement of F, A, and I in FAIR for biomolecular (meta)data, and supports Helmholtz researchers in delivering high-quality metadata to international data repositories. HARMONise connects with high-level international projects of the Ocean Biomolecular Observing Network (OBON) Programme of the UN Decade of Ocean Science, to further align our developments with global strategies and ensure Helmholtz-to-global interoperability.
The project HARMONise (ZT-I-PF-3-027) is funded by the Initiative and Networking Fund as part of the Helmholtz Metadata Collaboration Project cohort 2021.

Please specify ”other”

In addition, please add 3 to 5 keywords.

sequencing metadata, eDNA, interoperability, metadata standards

Please specify ”other”

For whom will your contribution be of most interest?

Researchers

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

Researchers

Primary authors: BIENHOLD, Christina (AWI Helmholtz Centre for Polar and Marine Research); BAYER, Till (GEOMAR); HARMS, Lars; KOPPE, Roland; NEUHAUS, Stefan; SIEBERT, Isabell (AWI)

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

Session Classification: Poster Session A

Track Classification: Connecting research data: 4. Metadata annotation and management

Contribution ID: **85** Type: **TALK**

Ontowhat? Journey towards a sensor maintenance ontology

Tuesday 5 November 2024 13:20 (20 minutes)

The collection and use of sensor data is crucial for scientists monitoring and observing the Earth's environment. In particular, it enables the evaluation of real natural phenomena over time and is essential for the validation of experiments and numerical simulations. Assessment of data quality beyond statistics includes knowledge and consideration of sensor state, including operation and maintenance, e.g. calibration parameters and maintenance time windows. Today, maintenance metadata is often collected even digitally but not readily accessible due to a lack of standardization and findability. In the HMC project MOIN4Herbie, digital recording of FAIR sensor maintenance metadata is developed using the electronic lab notebook Herbie.

In information science, ontologies are a formalization of concepts, their relations and properties. Using ontologies allows to collect input which is right away fit for purpose as findable, machine-readable and interoperable (meta)data. Ontologies can ensure the usage of controlled vocabularies and organize the knowledge stored within for assessibility and thus reuse. Herbie relies on ontologies and vocabularies to generate user-friendly web-forms for collecting, validating and then provisioning FAIR (meta)data.

What is the challenge regarding ontologies within the MOIN4Herbie context? No ontology for sensor maintenance metadata has yet been developed that covers all aspects considered relevant for data quality assessment in marine science. Several industrial maintenance ontologies exist, but none is sensor specific. Therefore, a new task-specific ontology needs to be created. Furthermore, the domain experts in the project had heard of ontologies but had never used or developed one.

With this contribution we would like to disseminate the process of learning about ontologies from scratch and describe step by step our process from the idea to the first version of the final ontology. Starting with the basic definition of an ontology, we learned about ontology levels and competency questions. We collected and evaluated controlled vocabularies and ontologies related to sensor description and industrial maintenance. We developed competency questions and evaluated them in collaboration with our sensor experts. We visualised the competency questions in flowcharts and attached ontology terminologies to all features. We structured the reused ontologies and developed our own first draft of a maintenance ontology. We will share our experiences and are open to feedback!

Please specify ”other”

In addition, please add 3 to 5 keywords.

Herbie, ELN, ontologies

Please specify ”other”

For whom will your contribution be of most interest?

Data professionals who provide and maintain data infrastructure

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

Data professionals who provide and maintain data infrastructure

Primary authors: BALDEWEIN, Linda (Helmholtz-Zentrum Hereon); SCHIRNICK, Carsten (GEOMAR Helmholtz Centre for Ocean Research Kiel); FABER, Claas (GEOMAR Helmholtz Centre for Ocean Research Kiel); HEPACH, Helmke (GEOMAR Helmholtz Centre for Ocean Research Kiel); SRINIVASA, Smruthishree (Helmholtz-Zentrum Hereon); ESCHKE, Catriona (Helmholtz-Zentrum Hereon); KIRCHNER, Fabian (Helmholtz-Zentrum Hereon)

Presenter: BALDEWEIN, Linda (Helmholtz-Zentrum Hereon)

Session Classification: Session E1

Track Classification: Connecting research data: 6. Interoperable semantics at domain and application level

Contribution ID: **86** Type: **TALK**

Progress and lessons learned from Data and Metadata management in laser-particle acceleration at HZDR

Monday 4 November 2024 11:45 (20 minutes)

Metadata is a key element in data management when taking account of the F.A.I.R.(findable, accessible, interoperable and reusable) principles, answering the need for better data integration and enrichment. In the field of high-intensity laser-plasma physics, numerical simulations and experiments go hand in hand, complementing each other. While simulation codes are well documented and output files can follow some standards (like openPMD or Smilei/happi), input files were often neglected. Experiment documentation is typically diverse, containing the description of manually executed setup steps (with photos or hand drawings), tabular data of experiment execution (parameters and observations) alongside with actual detector data. Often, data from the driving laser system is better organized but poorly connected to the experiment.
We will report on recent progress of data management in the field of high-intensity laser-plasma physics at HZDR by means of the center's data strategy and external projects like "HELPMI"by HMC, "DAPHNE4DNFI"by NDFI and others.
HELPMI is an HMC project aiming towards a data standard for laser-plasma (LPA) experiment data, making data interoperable (I) and reusable (R). While openPMD is an open standard for simulations in that domain, NeXus is an open standard for experimental data in Photon and Neutron sciences. HELPMI has identified benefits and challenges when adopting NeXus for the LPA domain and extended openPMD for arbitrary data hierachies like NeXus. Alongside, a domain-specific glossary is being developed, where the community must be involved.
DAPHNE4DNFI supports metadata capture and data enrichment activities at HZDR. One major achievement is a web app for manual experiment logging, i.e. taking the above-mentioned tabular data of parameters and observations. This app is highly configurable, following the changes and improvements of experimental setup and techniques. It can connect to other electronic lab documentation resources (like the Mediawiki system deployed at HZDR) in order to directly re-use metadata. Data is stored in a central database instead of multiple spreadsheet files and can be directly plotted, also against historical data.
Another important outcome of DAPHNE4DNFI is metadata capturing and cataloging of simulation input data. This way, tables of simulation input data can be generated, allowing to re-use input files. Of course, output data and analysis scripts are also linked, thus the in-house re-use of data is strongly enhanced.
On top of that, but also as necessary tool, DAPHNE4DNFI has strongly promoted the use of metadata catalogues, in particular SciCAT. Even though daily usage is automated via scripts, a web interface to browse and search for data and metadata is extremely helpful. Such metadata catalogues complement data repositories by enabling the F (findable) but require a lot of effort in data enrichment.

Please specify "other"

In addition, please add 3 to 5 keywords.

metadata capture, enrichment, metadata catalogue

Please specify ”other”

For whom will your contribution be of most interest?

Data professionals who provide and maintain data infrastructure

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

Researchers

Primary author: SCHLENVOIGT, Hans-Peter (HZDR)

Co-authors: DEBUS, Alexander (Helmholtz-Zentrum Dresden-Rossendorf); KESSLER, Alexander; POESCHEL, Franz (CASUS/HZDR); HORNUNG, Johannes (GSI Helmholtzzentrum für Schwerionenforschung); TIPPEY, Kristin Elizabeth (HZDR); KALUZA, Malte Christoph (Helmholtz-Institut Jena); SCHWAB, Matthew Bradley (Friedrich-Schiller-Universität Jena); BUSSMANN, Michael (HZDR); KNODEL, Oliver (Helmholtz-Zentrum Dresden-Rossendorf); KLUGE, Thomas (HZDR); BAGNOUD, Vincent

Presenter: SCHLENVOIGT, Hans-Peter (HZDR)

Session Classification: Session B1

Track Classification: Connecting research data: 6. Interoperable semantics at domain and application level

Contribution ID: **87** Type: **POSTER&PITCH**

Digital sample management and documentation of analytical methods –Development of an electronic lab notebook at the Helmholtz-Institute Freiberg

Monday 4 November 2024 14:00 (1 hour)

At the Helmholtz-Institute Freiberg for Resource Technology (HIF), researchers develop new technologies to improve circular economy. In this context, different types of samples (e.g. rock samples, recycling material) play an important role. The sample passes through different states and labs – starting at the sample preparation, through the analysis of the particular sample to the final storage.
With electronic lab notebooks (ELNs) this entire process is digitized, thus improving findability, accessibility, interoperability and reusability (FAIR) of the samples and their corresponding data. Once the sample is registered in the system, every further work on the sample will be connected to this sample, explicitly. Thus, all important metadata can be recorded digitally in a structured way.
At the HIF, we are developing an ELN based on semantic MediaWiki. For this, we are incorporating all analytical methods existing at the HIF into the system. Thus, scientists and lab personnel need to adjust accustomed processes of data documentation. Therefore, the system needs to be as user friendly and as close to the familiar processes a possible. Where this is not possible, future users need to be motivated and included into the entire development process. For this, personal exchange between scientists/ lab personnel and developers is of great importance.
In this contribution, we will discuss the challenges in the development of an ELN, including technical and personal aspects and present the structure of our ELN.

Please specify ”other”

In addition, please add 3 to 5 keywords.

Electronic lab notebook, Sample management, semantic MediaWiki

Please specify ”other”

For whom will your contribution be of most interest?

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

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

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

Primary author: SCHALLER, Theresa (HMC/ HZDR)

Co-authors: Dr RAU, Florian (HZDR); STEINMEIER, Leon (Helmholtz Institute Freiberg); GRUBER, Thomas (HZDR)

Presenter: SCHALLER, Theresa (HMC/ HZDR)

Session Classification: Poster Session A

Track Classification: Connecting research data: 7. Infrastructure and common practices for consolidation of (meta)data

Contribution ID: **89** Type: **TALK**

NetCDF Metadata Guidelines for the Helmholtz Earth and Environment Community

Monday 4 November 2024 10:50 (20 minutes)

In the pursuit of making data FAIR (Findable, Accessible, Interoperable, Reusable) (Wilkinson et al., 2016: https://doi.org/10.1038/sdata.2016.18), the need for well and comprehensively described datasets is decisive. In order to facilitate interoperability and reusability, it is essential to have self-describing data, which can only be achieved by enriching data with metadata. Within the Earth System science community, the NetCDF data format has become the quasi-standard, supported by general metadata standards such as the CF conventions (Climate and Forecast, https://cfconventions.org/) and more specific ones like AtMoDat (https://www.atmodat.de/) for atmospheric modeling. However, current NetCDF metadata schemas often have limitations that hinder seamless data integration, findability and reuse through metadata portals. This project aims to bridge these gaps and harmonize the diverse schemas to ensure a more robust and unified metadata framework.
From the DataHub, the central data infrastructure of the Research Field Earth and Environment within the Helmholtz Association, an initiative was launched with the objective of developing harmonized guidelines that would unify the diverse sub-communities within both the observational and modeling fields. Under the direction of the centers KIT and Hereon, the initiative's approach entails a comprehensive examination of existing guidelines, followed by the integration and expansion of these guidelines to address the diverse needs within the Earth and Environment disciplines. The outcome will be a set of comprehensive guidelines designed to enhance data interoperability and reusability, and tools to facilitate their adoption.
Key milestones include:
- Reconciling attributes in the draft guidelines.
- Technical implementation of the guidelines document
- Development of machine-readable templates and validation tools.
- Provision of tailored tools (e.g. Jupyter Notebooks) for user-friendly implementation of metadata profiles based on the guidelines.
A notable enhancement would be the implementation of these guidelines in repositories, enabling automatic extraction and mapping of NetCDF metadata to repository-specific metadata fields.

This initiative aims to foster a more unified and efficient data management practice within the Earth and Environment community, ultimately enhancing the utility and accessibility of scientific data.

Please specify "other"

In addition, please add 3 to 5 keywords.

NetCDF, metadata schema, DataHub

Please specify "other"

For whom will your contribution be of most interest?

Researchers

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

Data professionals and stewards

Primary author: FÖSIG, Romy (KIT)

Co-authors: ERTL, Benjamin; SASS, Björn Lukas (Helmholtz-Zentrum Hereon); LORENZ, Christof (Karlsruhe Institute of Technology); REBMANN, Corinna (KIT); LOEWE, Katharina; SOMMER, Philipp Sebastian (Helmholtz-Zentrum Hereon); BARTHLOTT, Sabine (KIT); HASSLER, Sibylle; KERZENMACHER, Tobias (KIT)

Presenter: FÖSIG, Romy (KIT)

Session Classification: Session A1

Track Classification: Connecting research data: 4. Metadata annotation and management

Contribution ID: **91** Type: **TALK**

Towards Standardizing Catalysis and Beamline experiments at HZB

Monday 4 November 2024 10:30 (20 minutes)

Advanced catalysts are key to sustainable energy, reducing emissions, and improving resource efficiency. However, the synthesis of novel catalysts usually involves a unique blend of scientific methods, precise catalyst formulations, and the empirical knowledge of scientists. Additionally, the wide variety of techniques performed at different beamlines in synchrotron radiation facilities, along with continuously changing sample environments, produces highly heterogeneous data.
In this work we present the efforts to develop a common (meta)data schema that combines existing international community-driven standards in order to integrate and make such heterogeneous data FAIR (Findable, Accessible, Interoperable, and Reusable) at HZB catalysis labs and BESSY II. This common schema for managing catalysis and beamline (meta)data is built with LinkML, an internationally adopted framework for building data models that are both human-readable and machine-actionable, particularly useful in linked data and semantic web technologies. We use voc4cat –a SKOS vocabulary for catalysis emerged from NFDI4Cat –and NeXus –a common data exchange format for X-ray, neutron, and muon experiment, being developed as an international standard –to enhance and standardize data related to catalysis experiments and beamline operations. Our ultimate goal is aiming for data that is consistently formatted and integrated across different sources and systems. To ensure consistency and interoperability, we test the schema by integrating it with electronic laboratory notebook (ELN) workflows using NOMAD, an open-source data management platform developed by FAIRmat.

This effort intrinsically supports the data management milestones of the ROCK-IT (Remote, Operando Controlled, Knowledge-driven, and IT-based) project, funded by the Helmholtz Association, which exemplifies the practical application of FAIR principles at BESSY II and demonstrates how state-of-the-art IT can significantly enhance control and insights with a focus on catalysis experiments.

Please specify ”other”

In addition, please add 3 to 5 keywords.

Catalysis, NeXus, LinkML

Please specify ”other”

For whom will your contribution be of most interest?

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

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

Data professionals and stewards

Primary author: VELAZQUEZ SANCHEZ, Ana (HZB)

Co-author: PATEL, Sonal Ramesh (HZB)

Presenters: VELAZQUEZ SANCHEZ, Ana (HZB); PATEL, Sonal Ramesh (HZB)

Session Classification: Session A1

Track Classification: Connecting research data: 4. Metadata annotation and management

Contribution ID: **92** Type: **TALK**

Enabling a Global Research Commons through vertical interoperability of research tools and services

Tuesday 5 November 2024 11:20 (20 minutes)

Interoperability is an ongoing challenge given the diverse nature of research and the tools and services researchers use. Addressing interoperability challenges and FAIRification of research at scale is therefore only possible with solid knowledge about the tools and services used in each stage of the research cycle and a forward-facing vision of how they might work together.

For the latter, the RDA Global Open Research Commons Working Group published in October, 2023 the Global Open Research Commons International Model (GORC Model) and made available a well-researched and fully featured template for a Research Commons. To borrow the definition by Scott Yockel, University Research Computing Officer at Harvard, a research commons "brings together data with cloud computing infrastructure and commonly used software, services and applications for managing, analyzing and sharing data to create an interoperable resource for a research community".

For the former, the RDA Mapping the landscape of digital research tools Working Group has mapped existing tools across the research cycle in the so called MaLDreTH (Mapping the Landscape of Digital Research Tools Harmonised) model which may facilitate initiatives for improving the interoperability of these tools to better support researchers with FAIR workflows.

In this presentation, we will provide an overview about the GORC model, and how it has been applied in designing the proposed Research Commons for Norway (REASON), that was submitted to the Norwegian Infrastructure Fund in November, 2023. Additionally, we will share the latest version of the MaLDreTH model to stimulate and inform discussions on how to build highly interoperable research ecosystems.

Please specify "other"

Tool provider

In addition, please add 3 to 5 keywords.

Research Commons, Vertical Interoperability, Research Cloud, FAIR data

Please specify "other"

For whom will your contribution be of most interest?

Expert panels, strategists and administrative stakeholders

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

other (please specify)

Primary author: MACNEIL, Rory (Research Space)

Presenter: MACNEIL, Rory (Research Space)

Session Classification: Session D1

Track Classification: From recommendations to implementations: 9. Consulting concepts

Contribution ID: **93** Type: **TALK**

Metadata considerations in research tool design to achieve vertical interoperability

Tuesday 5 November 2024 09:45 (20 minutes)

RSpace is an open-source platform that supports researchers in the active research phase to plan, conduct, and document their work, and thereby make their research more robust and FAIR (Findable, Accessible, Interoperable, Reproducible). Interoperability with tools and services used by researchers throughout the research lifecycle is a fundamental element of RSpace's development philosophy. This interoperability is crucial for enabling efficient workflows and seamless connections between the different phases of the research process. Here, metadata plays a key role in facilitating the flow of information and data between diverse systems and applications used by researchers.

In this presentation, we will explore the opportunities and challenges of metadata in achieving vertical interoperability between tools and services in the design and development of research tools like RSpace. We will report on RSpace's metadata vision and opportunities to address interoperability challenges for better integration with complementary tools and services. The presentation will delve into the specifics of RSpace's metadata approach, highlighting challenges faced and strategies being employed to implement this vision in collaboration with the open-source and research data management community. This includes discussions around metadata standards, typing approaches, controlled vocabularies, the development of APIs and other integration mechanisms, such as RO-Crates, to facilitate seamless data exchange and tool interoperability.

In addition, please add 3 to 5 keywords.

Vertical Interoperability, Research Tools, FAIR data, Metadata management

Please specify "other"

For whom will your contribution be of most interest?

Data professionals and stewards

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

other (please specify)

Please specify "other"

Tool provider

Primary author: MATHES, Tilo (Research Space)

Presenter: MATHES, Tilo (Research Space)

Session Classification: Session C2

Track Classification: Connecting research data: 4. Metadata annotation and management

Contribution ID: **94** Type: **TALK**

Streamlined Submission of Human Omics Data via the GHGA Metadata Model

Monday 4 November 2024 12:25 (20 minutes)

The German Human Genome Phenome Archive (GHGA) is a national infrastructure that promotes the secure storage, exchange, and management of access-controlled human omics data. To facilitate user-friendly and comprehensive data submissions, we developed the GHGA metadata model. The standardized model aims at maximizing the amount of collected metadata on the submitter side, enabling reusable submissions of different types of -omics data into GHGA. This metadata model is embedded in a robust ethico-legal framework addressing sensitive data and can satisfy the heterogeneous needs of submitters while maintaining FAIR principles, interoperability with European Genome Archive (EGA) and facilitating streamlined user journeys.

The GHGA metadata schema models a bottom-up experimental approach from sample collection via omics experiment procedure to bioinformatic analysis. The model allows capturing information about submitter-specified datasets, access restrictions, and inherent studies. To appropriately model this approach, the schema consists of classes that comprise both research and administrative metadata. The research metadata resembles the skeleton of the metadata model, based on the central EGA (cEGA) model, which are: Experiment, Analysis, Sample and Individual. Other classes such as Experiment Method and Analysis Method capture specifications that are tailored to perform different experiment or analysis types based on the submission type. Furthermore, the metadata model controls submitted Files through three different classes, namely Research Data File, Process Data File, and Supporting File. These file types differ with regard to the information they contain. A Research Data File holds the raw data that is the basis for further processing and analyses, which will result in a Process Data File. A Supporting File can be submitted for Experiment Method, Analysis Method and the Individual and may contain additional (un-) structured details, such as protocols or phenopackets. The administrative metadata captures information related to governance, access controls, and data use policies (Data Access Committee, Data Access Policy, Study, Publication).

The GHGA metadata schema is equipped to enable metadata exchange between GHGA and central EGA, as well as between GHGA and NFDI4Health, a partner consortium in the national research data infrastructure project, and the model project genome sequencing (MV GenomSeq, §64e SGB V). Information about the type of data collected, the methodology used, the purpose, and the governing entities are required for GHGA functionality. Details regarding downstream analysis are only required when submitting processed files. Classes in the metadata schema are further explained using properties, such as ‘sex’or ‘phenotypic feature’for Individual or ‘instrument model’ and ‘library type’for Experiment Method. These properties can either be restricted, recommended, or optional, highlighting their importance in FAIRifying omics metadata. Further, we make use of community-accepted ontologies to control the content of submitted properties, promoting the significance of standardizing metadata collections and limiting the number of free-text fields in our model to an absolute minimum.

To summarize, we have developed the GHGA metadata model to be an openly accessible resource with an easy-to-use and streamlined data submission process. The model is designed to be FAIR, flexible, and interoperable to address diverse community needs, while maintaining data subject anonymity.

Please specify ”other”

In addition, please add 3 to 5 keywords.

FAIR data, metadata model, metadata quality, human omics metadata, biomedicine

Please specify ”other”

For whom will your contribution be of most interest?

Researchers

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

Data professionals and stewards

Primary authors: Dr IYAPPAN, Anandhi (EMBL Heidelberg); Ms MAUER, Karoline (Systems Medicine, German Center for Neurodegenerative Diseases (DZNE) e.V, German Center for Neurodegenerative Diseases (DZNE), PRECISE Platform for Genomics and Epigenomics at DZNE, and University of Bonn, Bonn, Germany)

Co-authors: Mr MENGES, Paul (German Center for Cancer Research (DKFZ), Heidelberg, Germany); Ms SÜRÜN, Bilge (Quantitative Biology Center (QBiC), University of Tübingen, Tübingen, Germany); Ms TREMPER, Galina (German Center for Cancer Research (DKFZ), Heidelberg, Germany); Dr KIRLI, Koray (German Center for Cancer Research (DKFZ), Heidelberg, Germany); Dr ULAS, Thomas (Systems Medicine, German Center for Neurodegenerative Diseases (DZNE) e.V., Germany, German Center for Neurodegenerative Diseases (DZNE), PRECISE Platform for Genomics and Epigenomics at DZNE, and University of Bonn, Bonn, Germany); Dr NAHNSEN, Sven (Quantitative Biology Center (QBiC), University of Tübingen, Tübingen, Germany); Prof. SCHULTZE, Joachim L. (Systems Medicine, German Center for Neurodegenerative Diseases (DZNE) e.V., Bonn, Germany, German Center for Neurodegenerative Diseases (DZNE), PRECISE Platform for Genomics and Epigenomics at DZNE, and University of Bonn, Bonn, Germany, Life and Medical Sciences (LIMES) Institute, University of Bonn, Bonn, Germany); Prof. BORK, Peer (European Molecular Biology Laboratory (EMBL), Heidelberg, Germany); CONSORTIUM, GHGA (German Human Genome-Phenome Archive (GHGA, W620), Deutsches Krebsforschungszentrum, Heidelberg, Baden-Württemberg, Germany)

Presenters: Dr IYAPPAN, Anandhi (EMBL Heidelberg); Ms MAUER, Karoline (Systems Medicine, German Center for Neurodegenerative Diseases (DZNE) e.V, German Center for Neurodegenerative Diseases (DZNE), PRECISE Platform for Genomics and Epigenomics at DZNE, and University of Bonn, Bonn, Germany)

Session Classification: Session B1

Track Classification: Connecting research data: 6. Interoperable semantics at domain and application level

Contribution ID: **95** Type: **POSTER&PITCH**

How to make Biomedical Imaging Datasets AI-ready?

Monday 4 November 2024 15:00 (1 hour)

The vast amount of observations needed to train new generation AI models (Foundation Models) necessitates a strategy of combining data from multiple repositories in a semi-automatic way to minimize human involvement. However, many public data sources present challenges such as inhomogeneity, lack of machine-actionable data, and manual access barriers. These issues can be mitigated through the consequent adherence to the FAIR (Findable, Accessible, Interoperable, Reusable) data principles, as well as state-of-the-art data standards and tools. In the poster, we highlight the inhomogeneity of the schema definitions in the field, provide helpful tips on what could improve the AI-readiness of data and inspect example data sources which implement the most novel concepts in working with data and metadata in the machine-actionable fashion.

Please specify ”other”

In addition, please add 3 to 5 keywords.

Artificial Intelligence, Fair Data Point, Bioimaging, Data harmonization

Please specify ”other”

For whom will your contribution be of most interest?

Data professionals who provide and maintain data infrastructure

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

Researchers

Primary author: DVORETSKII, Stefan (HMC Hub Health, DKFZ)

Co-authors: Mr MOORE, Josh (German BioImaging e.V.); KULLA, Lucas (DKFZ); NOLDEN, Marco (DKFZ); Mr SCHADER, Philipp (HMC Hub Health, DKFZ)

Presenter: DVORETSKII, Stefan (HMC Hub Health, DKFZ)

Session Classification: Poster Session B

Track Classification: Connecting research data: 4. Metadata annotation and management

Contribution ID: **96** Type: **TALK**

The Reusability of Scientific Data Course

Tuesday 5 November 2024 11:40 (20 minutes)

The "Reusability of Scientific Data" course is designed to equip researchers with the knowledge and practical skills necessary to ensure their data adheres to the principles of FAIR (Findable, Accessible, Interoperable, Reusable), with a specific focus on reusability. This 4-hour online course provides a detailed exploration of the critical role that data reusability plays in enhancing the impact and longevity of research.

The course aims to foster a deep understanding of the methods required to achieve data reusability, offering participants hands-on experience in key steps to make their research data accessible and valuable to the broader scientific community.

With an interactive format designed for 20 participants, the course encourages discussion and practical application of data reusability techniques. By the end of the course, participants will have a comprehensive understanding of the essential processes involved in making research data reusable. They will leave equipped with the tools and insights necessary to implement these practices in their work, thereby contributing to the advancement of open science and the reproducibility of research.

Please specify "other"

In addition, please add 3 to 5 keywords.

Data Reusability, FAIR Principles, Research Data, Open Science, Reproducibility,

Please specify "other"

For whom will your contribution be of most interest?

Researchers

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

Data professionals and stewards

Primary author: Dr OEZKAN, OEzlem (HMC)

Co-author: MANNIX, Oonagh (HMC matter/HZB)

Presenter: Dr OEZKAN, OEzlem (HMC)

Session Classification: Session D1

Track Classification: From recommendations to implementations: 8. Development of training programmes

Contribution ID: **97** Type: **TALK**

Towards FAIR digital objects for Heritage science data

Tuesday 5 November 2024 13:20 (20 minutes)

Heritage science is an interdisciplinary field that involves the scientific study of cultural and natural heritage. It entails collecting and producing a wide variety of data, including descriptions of objects and sites, samples, sampling locations, scientific instrumentation, analytical methods, conservation and restoration records, environmental monitoring data, documentation, and digital representations. E-RIHS (European Research Infrastructure for Heritage Science) is a collaborative effort among key players in Europe and beyond, on the verge of becoming an operational European Research Infrastructure Consortium (ERIC). The focus is on cooperation, for which interoperability of processes and the need for FAIR and linked data are indispensable prerequisites.

This presentation will cover ongoing efforts to advance the realisation of digital objects for encapsulating and describing data collected in E-RIHS and may serve as a framework for heritage science data in general. The work was carried out within the scope of IPERION HS, the last of the past projects building towards E-RIHS. The focus is on minimising documentation burden while enhancing interoperability through the creation of reusable, linkable digital objects.

A systematic, bottom-up methodology is employed to break down the entire metadata environment of heritage science into accessible individual digital objects. These objects are semantically consistent and correspond to separate entities that are linked together to form the data flow involved in a research project. The models are meticulously modelled, taking into account all essential metadata associated with them. Existing data and metadata gathered during previous and current E-RIHS-related projects, as well as similar models and standards defined in other research domains, were studied wherever possible.

The Simple Dynamic Modelling tool, developed by the National Gallery, was extensively used to facilitate the development of the models and their visualisation. The models can be found in the E-RIHS organisation on GitHub. They are revised frequently as related models evolve, ensuring efficient linking between the models and avoiding duplication of metadata.

The models are then converted into JSON Schema, which after validation, can be visualized and evaluated using commonly used software libraries that generate web-forms from JSON schema documents. An instance of Cordra, open source software for managing digital objects instance and implementing the Digital Object Interface Protocol (DOIP), was created for E-RIHS to serve as a searchable database and metadata repository for heritage science digital objects. The ultimate goal is to pave the way for a future where information is not only preserved but also effortlessly navigable and actionable within the broader heritage scientific landscape.

Please specify ”other”

In addition, please add 3 to 5 keywords.

Heritage Science, Data Modelling, JSON Schema, Digital Objects

Please specify ”other”

For whom will your contribution be of most interest?

Researchers

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

Researchers

Primary authors: FREMOUT, Wim (Royal Institute for Cultural Heritage (KIK-IRPA)); Mr PADFIELD, Joseph (The National Gallery); Prof. SOTIROPOULOU, Sophia (FORTH –Institute of Electronic Structure and Laser (IESL)); Dr SCHMIDLE, Wolfgang (Free University of Berlin (FU Berlin))

Presenter: FREMOUT, Wim (Royal Institute for Cultural Heritage (KIK-IRPA))

Session Classification: Session E2

Track Classification: Connecting research data: 5. Technical solutions for findable and machine-readable metadata

Contribution ID: **98** Type: **TALK**

Using SHACL Shapes to create semantic (meta)data

Monday 4 November 2024 10:30 (20 minutes)

Using RDF is a natural choice for modelling semantically linked metadata for FAIR research data. However, the learning curve for RDF is steep, and even for data stewards, becoming familiar with all the relevant technicalities can be a major barrier. Therefore, ULB Darmstadt is heavily involved in developing and providing services that facilitate the creation and use of semantic metadata, making the technology accessible to scientists without extensive knowledge of linked data.

We will present several services for the description of research data with structured, semantic metadata based on SHACL shapes and RDF. The services include:

- The NFDI4Ing Metadata Profile Service (https://profiles.nfdi4ing.de) allows creation, sharing, curation and reuse of SHACL-based application profiles. The profiles are created in a graphical user interface by combining suitable terms that can be selected from existing ontologies.
- The NFDI4Ing Data Ingest Service (https://ingest.nfdi4ing.de) is the gateway for enhanced research data publication in architecture and civil engineering. With flexible metadata profiles based on SHACL and automation processes, ing.est visualizes various 3D data formats, assigns persistent identifiers, and ensures long-term archiving.
- The CSV-RDF Mapper (demo at https://ulb-darmstadt.github.io/csv-rdf-mapper) allows conversion of tables to RDF data by mapping the table to a target format defined by a SHACL shape.
- The SHACL Search Engine (under active development, demo available in October 2024) indexes RDF data that conform to SHACL shapes using Apache SOLR and, based on the given SHACL shapes, generates a user interface with search facets that enable filtering the indexed data.

We will give an outlook on future plans and on possibilities to cooperate in this field.

Please specify ”other”

In addition, please add 3 to 5 keywords.

RDF, SHACL, application profiles, web-service

Please specify ”other”

For whom will your contribution be of most interest?

Researchers

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

Data professionals who provide and maintain data infrastructure

Primary authors: Dr FUHRMANS, Marc (TU Darmstadt); TITTEL, Stephan (TU Darmstadt)

Presenters: Dr FUHRMANS, Marc (TU Darmstadt); TITTEL, Stephan (TU Darmstadt)

Session Classification: Session A2

Track Classification: Connecting research data: 5. Technical solutions for findable and machine-readable metadata

Contribution ID: **99** Type: **POSTER&PITCH**

Development of recommendations for the implementation of semantic artifacts in HMC Earth and Environment

Monday 4 November 2024 16:00 (1 hour)

Embedding semantics within research metadata serves to standardize, refine and contextualize it, thereby improving interoperability between data sources and promoting the FAIR principles. Within the Helmholtz Association, we are committed to evaluating existing semantic resources and established practices and to developing guidelines for their handling and use in the field of earth and environment. In order to develop appropriate recommendations, we have set up a community working group to discuss similarities and differences between institutions in dealing with 'semantics', and to relate these to practices in other organizations and the literature. The working group is divided into the work packages "Observable properties" and "Measurement instruments/methodology" and will be expanded to include further work packages if required. Discussion points and derived recommendations are made available transparently on our HMC Earth and Environment community portal and are explicitly open to changes, ensuring that they are continuously adapted to current developments. The aim of the working group is to develop and agree upon concrete recommendations for the implementation of semantics in the field of Earth and Environment, to support others in their decisions and to improve data interoperability within Helmholtz Earth and Environment.

Please specify "other"

In addition, please add 3 to 5 keywords.

Semantics, Controlled vocabulary, Earth and Enviroment

Please specify "other"

For whom will your contribution be of most interest?

Data professionals and stewards

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

Data professionals and stewards

Primary author: KOTTMEIER, Dorothee (HMC E&E @PANGAEA/AWI)

Co-authors: PÖRSCH, Andrea (HMC Hub EE at GFZ); MALINOVSCHII, Stanislav (HMC); LORENZ, Sören; SÖDING, Emanuel (GEOMAR); RAZEGHI, Yousef

Presenter: KOTTMEIER, Dorothee (HMC E&E @PANGAEA/AWI)

Session Classification: Poster Session C

Track Classification: Connecting research data: 6. Interoperable semantics at domain and application level

Contribution ID: **100** Type: **TALK**

DatAasee - A Metadata-Lake for Libraries

Tuesday 5 November 2024 11:20 (20 minutes)

A library is a super repository of digital and physical data archives, which is organized by metadata. This metadata however, maybe distributed across various databases due to, for example, topical or typical grouping. To provide a unified view or overview of all resources, the metadata needs to be aggregated, normalized, and potentially interconnected. **DatAasee** is such a metadata aggregator based on a metadata-lake architecture: Metadata records are ingested from upstream (meta)data sources specified by protocol-format combinations, partially transformed, indexed, and connected to related metadata while their original formatting remains available. The resulting normalized metadata can then be queried by downstream services or clients. With this self-hostable open-source service, we aim to contribute an overarching metadata-layer that makes repositories and thus libraries more FAIR.

Please specify ”other”

In addition, please add 3 to 5 keywords.

Metadata Catalog, Metadata Management, Metadata Aggregation

Please specify ”other”

Mainly data professionals, yet also all of the above.

For whom will your contribution be of most interest?

other (please specify below)

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

Data professionals who provide and maintain data infrastructure

Primary author: Dr HIMPE, Christian (University of Münster)

Presenter: Dr HIMPE, Christian (University of Münster)

Session Classification: Session D2

Track Classification: Connecting research data: 5. Technical solutions for findable and machine-readable metadata

Contribution ID: **101** Type: **TALK**

Fundamentals of Scientific Metadata –A Hands-on Training Course on FAIR Data Handling for Researchers and Data Stewards

Tuesday 5 November 2024 11:00 (20 minutes)

Scientific research has been subject to the fast-progressing digitalization, which impacts how research is conducted today. Generation and sharing of data according to best practices, that support the digital change, bears numerous challenges for the scientists: Implementation of data documentation recommendations like the FAIR Principles 1 require profound knowledge and technical skills and thus well-trained, data-competent, and technically skilled researchers. The typical scientific curricula however, often don't include these aspects: More than 45 % of scientific staff state to have little to know prior knowledge about the FAIR principles and metadata handling, while general interest in training formats on these topics is high (91.7 %)2.
With our training course „Fundamentals of Scientific Metadata", we established training material that covers the fundamental elements of (meta)data generation and addresses early-career researchers of any scientific domain. The didactic concept of the course encourages and motivates the participants to begin and sustainably proceed with the structured, schema-conform documentation of their scientific (meta)data. The material covers the fundamentals of (semi-)structured metadata, schemas and standards, as well as persistent identifiers (PIDs). (Meta)data generation is a predominantly practical skill that should be acquired and consolidated in a hands-on manner. Therefore, our course makes use of familiar problems and interrelated exercises to encourage the participants to practically test and consolidate the newly acquired skills and concepts. An initially unfamiliar data object is annotated with increasingly structured metadata throughout the course, complying with the FAIR principles. We set our focus on the confident handling on JSON and the development and understanding of JSON schemas.The training material was created in a modular manner that effortlessly enables the adaption of the material to various target groups in skill level as well as scientific domains: we have realized adaptions in the domains of Materials Science and Engineering and Particle Physics as well as for the target groups Researchers and Data Stewards. The domain-agnostic version of the training material has been published comprehensively via The Carpentries Incubator 3. Publications of the adaptions are in preparation.
To date, 11 instances of the domain-agnostic course and its individual adaptions have been conducted. Each instance was met with overwhelming interest in participation. We conclude every course instance with a comprehensive participants'evaluation. The evaluation results confirm the target group-oriented accuracy of the course contents as well as the high quality of our material.

Acknowledgements
This work was supported by (1) the Helmholtz Metadata Collaboration (HMC), an incubator-platform of the Helmholtz Association within the framework of the Information and Data Science strategic initiative and (2) the NFDI-MatWerk consortium funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the National Research Data Infrastructure –NFDI 38/1 –project number 460247524.

Please specify "other"

In addition, please add 3 to 5 keywords.

training course

JSON
schemas
education
HMC

Please specify ”other”

This contribution can be equally interesting for multiple groups:
- Data professionals and stewards
- Data professionals who provide and maintain data infrastructure
- Expert panels, strategists and administrative stakeholders

For whom will your contribution be of most interest?

other (please specify below)

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

Data professionals and stewards

Primary author: Dr GERLICH, Silke (HMC)

Co-authors: Dr AZÓCAR GUZMÁN, Abril (Institute for Advanced Simulations –Materials Data Science and Informatics (IAS-9), Forschungszentrum Jülich GmbH, Aachen, Germany); Dr OEZKAN, OEzlem (HMC); HOFMANN, Volker; SANDFELD, Stefan

Presenter: Dr GERLICH, Silke (HMC)

Session Classification: Session D1

Track Classification: From recommendations to implementations: 8. Development of training programmes

Contribution ID: **102** Type: **TALK**

Sample metadata at GEOMAR using LinkAhead: From registration to publication

Tuesday 5 November 2024 10:25 (20 minutes)

At GEOMAR, a multidisciplinary research centre, a large number of heterogeneous biological and geological samples need to be managed: Among other requirements, their metadata and data need to be stored in a FAIR way, their provenance information as well as their physical location in the sample storage need to be available, and scientists need to be supported in organizing their sample management and providing sample metadata in a standardized format to data portals like Marine Data where samples from different sources are searchable .

We present how GEOMAR's sample metadata schema is implemented in two instances of the research data management system LinkAhead. The workflow from (pre-)registration of samples prior to a sampling over updates of samples and storage information to export of metadata for publication is realized in the form of server-side scripts that can be accessed by scientists without any programming background from LinkAhead's web user interface. We show how exporting and updating information of multiple samples is handled by down- and uploading csv files, respectively. We also present how extensions of the datamodel that can be done by curators within the web user interface are automatically reflected in the LinkAhead crawler responsible for the upload of sample information.

Please specify "other"

In addition, please add 3 to 5 keywords.

Sample Management, Open Source,

Please specify "other"

For whom will your contribution be of most interest?

Researchers

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

Data professionals who provide and maintain data infrastructure

Primary authors: MAICHER, Doris (GEOMAR Helmholtz Centre for Ocean Research Kiel); MITTERMAYER, Felix (GEOMAR Helmholtz Centre for Ocean Research Kiel); SPRECKELSEN, Florian (IndiScale GmbH, Göttingen); MEHRTENS, Hela (GEOMAR Helmholtz Centre for Ocean Research Kiel); TOM WÖRDEN, Henrik (IndiScale GmbH, Göttingen)

Presenters: MITTERMAYER, Felix (GEOMAR Helmholtz Centre for Ocean Research Kiel); SPRECKELSEN, Florian (IndiScale GmbH, Göttingen)

Session Classification: Session C2

Track Classification: Connecting research data: 7. Infrastructure and common practices for consolidation of (meta)data

Contribution ID: **103** Type: **TALK**

Metadata for Ionospheric and Space Weather Observations (MISO)

Monday 4 November 2024 11:10 (20 minutes)

The focus of this project is the development of a standardized metadata vocabulary, essential for creating interoperable and easily discoverable data products across various research groups. By examining space weather-specific data products and formats, the project addresses the need for consistent metadata standards that will enhance collaboration and data sharing on an international scale. This can lead to seamless integration and combination of data in space physics and space weather research.

Significant progress has been made in standardizing data formats related to space weather, including the magnetosphere, atmosphere, and ionosphere. Recent efforts include extensive collaboration with the COSPAR Panel on Radiation Belt Environment Modeling (PRBEM) to refine and update existing standards. These updates involve incorporating new variables, improvements to data uncertainty estimation, and enhanced naming conventions. These advancements are crucial for ensuring uniformity and accuracy in data representation across different domains.

Future work will focus on engaging international groups to further discuss and refine space weather metadata standards. We are conducting consultations within Committee on Space Research and WG4 for the ISO committee 4. The project will also explore compatibility with existing data standards such as SPASE and HAPI, with the ultimate goal of contributing to the development of future ISO standards that support a more integrated and accessible data ecosystem in space weather research.

Please specify ”other”

In addition, please add 3 to 5 keywords.

space weather, space physics, radiation belt

Please specify ”other”

For whom will your contribution be of most interest?

Researchers

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

Researchers

Primary authors: CASTILLO, Angelica (GeoForschungs Zentrum, Potsdam (GFZ)); Dr KHAWAJA, Asim (GeoForschungs Zentrum, Potsdam (GFZ)); Dr BERDERMANN, Jens (Deutsches Zentrum für Luft und Raumfahrt (DLR), Neustrelitz); SZABO-ROBERTS, Matyas (GeoForschungs Zentrum, Potsdam (GFZ)); Dr SINNHUBER, Miriam (Karlsruhe Institute of Technology (KIT)); Prof. SHPRITS, Yuri (GeoForschungs Zentrum, Potsdam (GFZ))

Presenters: CASTILLO, Angelica (GeoForschungs Zentrum, Potsdam (GFZ)); Dr KHAWAJA, Asim (GeoForschungs Zentrum, Potsdam (GFZ))

Session Classification: Session A1

Track Classification: Connecting research data: 4. Metadata annotation and management

Contribution ID: **104** Type: **POSTER&PITCH**

Software Curation and Reporting Dashboard

Monday 4 November 2024 16:00 (1 hour)

Software is important research output. Therefore, funding agencies are interested in the value that a software contributes to the overall results of a funded project. The Helmholtz Association is working towards a system to evaluate data and software publications. The "Task Group Helmholtz Quality Indicators for Data and Software Publications" has already published a vision paper about how such Key Perfomance Indicators (KPIs) could look like (https://doi.org/10.48550/arXiv.2401.08804).

The goal of the Software Curation and Reporting Dashboard (Software CaRD; ZT-I-PF-3-080) is to support the publication and evaluation of software outputs of funded projects. We therefore regard KPIs as additional metadata (or a quality of certain metadata).

From the HERMES project (ZT-I-PF-3-006) comes a solution that can automatically extract and collate metadata from source code repositories. However, the HERMES process comes short when it comes to curation of the collected metadata, which is important as we need validated inputs for evaluation of KPIs. We also took a deeper look at the current state of the software KPI definitions by the Task Group and concluded that most of the criteria could be informed but not directly measured by the hermes software.

This gap should be bridged by Software CaRD by offering different views and processing options on the metadata produced by the HERMES workflow. On the one hand, we aim to provide an interactive interface that allows the browsing of the enriched CodeMeta graph provided by hermes. This will also include traces to the original source for each meta date and if possible hints or tools to fix inconsistencies. By checking the graph against constraints, we also intend to identify and highlight problems and do automated validation where possible.

Using constraints alongside additional hermes plugins that harvest relevant metadata, we also want to implement as many of the Helmholtz KPIs as possible.

The results of both aspectes are then presented in an explorative and interactive Dashboard that does also present the trail that led to a certain evaluation result. The poster explains the overall envisioned architecture and used technologies.

The HMC Conference is a good opportunity to collect further input about the required features and also to collect further ideas about automatic assessment of metrics.

Please specify "other"

In addition, please add 3 to 5 keywords.

automation kpi curation assessment publication

Please specify "other"

For whom will your contribution be of most interest?

Data professionals and stewards

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

Researchers

Primary author: MEINEL, Michael (Deutsches Zentrum für Luft- und Raumfahrt e.V.)

Co-authors: Dr MEESSEN, Christian (GFZ Potsdam); PAPE, David (Helmholtz-Zentrum Dresden-Rossendorf (HZDR)); BERTUCH, Oliver (Forschungszentrum Jülich); KERNCHEN, Sophie

Presenter: MEINEL, Michael (Deutsches Zentrum für Luft- und Raumfahrt e.V.)

Session Classification: Poster Session C

Track Classification: Assessing and monitoring FAIR data: 2. Metrics, tools and workflows for metadata assessment

Contribution ID: **105** Type: **POSTER&PITCH**

The Helmholtz Digitization Ontology: Harmonized semantics for the Helmholtz digital ecosystem

Monday 4 November 2024 14:00 (1 hour)

Abstract
Research in the Helmholtz Association undergoes continuous digitization. The heterogeneity of scientific contexts within Helmholtz leads to ambiguity and conflicts regarding metadata semantics. To ensure semantic interoperability of this decentralized data ecosystem, metadata should be aligned and harmonized with European and global initiatives to ensure an open and interoperable flow of data and information. Thus, HMC provides the Helmholtz Digitization Ontology (HDO), a mid-level ontology that contains concepts and relationships representing digital assets and processes that exist in the Helmholtz digital ecosystem. HDO is developed by contributors from all Helmholtz research fields. The main goal for developing HDO is to serve as a harmonized and machine-actionable institutional reference to represent digital assets and procedures pertinent to their handling and maintenance within Helmholtz.
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).
HDO development is carried out in three phases: 1) Initialization phase: an internal GitLab repository was created to gather a set of core classes and their definitions in per-term YAML files, 2) Implementation phase: YAML files were converted and merged into one OWL file. For this, keys of the template were mapped onto existing and imported annotation properties, and 3) Adoption and Adaption: a phase of continuous iterative development in which the ontology will be used in use cases across the different Helmholtz research fields. This will test the developed ontology against use case-specific requirements and allow further adoption based on iterative exchange. One example we are currently pursuing is the semantic representation of FAIR digital objects.
The current development and the first release can be found in our public git repository 1. The ontology is made accessible via a persistent identifier 2 and terms are derefrenced via their PIDs. An HTML documentation of HDO is available online 3.

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

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

Please specify ”other”

In addition, please add 3 to 5 keywords.

Ontology
Semantics
Metadata Management
OWL
FAIR

Please specify ”other”

For whom will your contribution be of most interest?

Data professionals who provide and maintain data infrastructure

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

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

Primary authors: FATHALLA, Said; HOFMANN, Volker

Co-authors: GUENTHER, Gerrit (Helmholtz-Zentrum Berlin); STEINMEIER, Leon (Helmholtz Institute Freiberg); LEMSTER, Christine (Geomar); KOTTMEIER, Dorothee (HMC E&E @PANGAEA/AWI); SIVAPATHAM, Lakxmi; SANDFELD, Stefan

Presenter: FATHALLA, Said

Session Classification: Poster Session A

Track Classification: Connecting research data: 6. Interoperable semantics at domain and application level

Contribution ID: **106** Type: **TALK**

Roles in FAIR data and their needs

Tuesday 5 November 2024 09:45 (20 minutes)

Different roles interact with research data in very different ways: Technicians, experimental scientists, data analysts, modellers, supervisor, infrastructure providers, data stewards, toolchain providers, project managers, administrative personelle, liberians, publishers, NFDI contact persons, indexing service providers, external data user, programmers,…

Non of them can establish an effective research data management all on their own. Currently only very few of them have the training and the tools they need. In order to make FAIR data a widespread reality we will need to educate people, establish toolchains, provide long term services, and adopt standards.

FAIR research requires an organisation wide innovation and cultural transition. Large research organisations like the Helmholtz Centers and their institutes are particularly well suited to lead this transition.

This talk tries to give an overview of what is required: What are the tasks of the different roles, we need to think of? Where is money required? Who needs in-service-training on what? What toolchains need to be established? Who might provide these tools? What can HMC do? What does the center need to do? What are the potential benefits for the center, and how can they be realized?

The main aim of this contribution is to bring the different groups together, to establish an understanding of what is required from them and what they can expect from others.

Please specify ”other”

In addition, please add 3 to 5 keywords.

Responsibility of the centers in science organisation, Data Management, Training

Please specify ”other”

For whom will your contribution be of most interest?

Expert panels, strategists and administrative stakeholders

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

Expert panels, strategists and administrative stakeholders

Primary authors: VAN DEN BOOGAART, Karl Gerald (HZDR/FWGB); RAU, Florian (HZDR); SCHALLER, Theresa (HMC/ HZDR); STEINMEIER, Leon (Helmholtz Institute Freiberg)

Presenter: VAN DEN BOOGAART, Karl Gerald (HZDR/FWGB)

Session Classification: Session C1

Track Classification: Assessing and monitoring FAIR data: 1. Human actors in the FAIR data landscape

Contribution ID: **107** Type: **TALK**

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

Tuesday 5 November 2024 10:25 (20 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). In PATOF we would like to build a "FAIR Metadata Factory"(see Figure 1) 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 specify "other"

In addition, please add 3 to 5 keywords.

Metadata, DataCite, Software

Please specify ”other”

For whom will your contribution be of most interest?

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

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

Data professionals and stewards

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

Co-author: KOEHLER, Martin (Deutsches Elektronen-Synchrotron DESY)

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

Session Classification: Session C1

Track Classification: Assessing and monitoring FAIR data: 1. Human actors in the FAIR data landscape

Contribution ID: **108** Type: **POSTER&PITCH**

Supporting Polymer Membrane Research: Enabling Semantics with PolyMat Ontology

Monday 4 November 2024 15:00 (1 hour)

The laboratory of the future necessitates innovative solutions for efficient digital (meta)data capture. Electronic laboratory notebooks (ELNs) are progressively replacing traditional documentation methods, significantly improving research data management and laboratory processes. However, free-text data entry presents challenges for automation and data quality. Ontologies address these issues by formalising scientific terminology and procedures, creating a semantic model that enhances interoperability and aligns with FAIR principles. These ontologies facilitate structured descriptions of experiments, capturing relationships between steps, instruments, and other components, thereby optimising processes. In the field of polymer membranes, we introduce PolyMat, a domain ontology that bridges material science and chemistry. PolyMat enables standardised, FAIR-compliant data collection and fosters cross-domain discoveries. Designed for integration with ELNs, PolyMat standardises terminology from the outset, supporting advanced features such as consistency checks and cross-experiment analysis.

Please specify ”other”

In addition, please add 3 to 5 keywords.

Ontology, Polymer Membrane, Electronic Lab Notebook, Research Data Management, Interoperability

Please specify ”other”

For whom will your contribution be of most interest?

Researchers

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

Researchers

Primary authors: Dr DEMBSKA, Marta; Dr HELD, Martin (Hereon); Dr SCHINDLER, Sirko (DLR Institute of Data Science)

Presenter: Dr DEMBSKA, Marta

Session Classification: Poster Session B

Track Classification: Connecting research data: 6. Interoperable semantics at domain and application level

Contribution ID: **109** Type: **TALK**

Harmonizing Marine Data for Real-Time Ingestion into Digital Twins of the Ocean (DTOs): An Open Data Infrastructure Approach

Monday 4 November 2024 12:05 (20 minutes)

The study of climate change and its impact on marine environments requires large-scale, multidisciplinary data that are often collected by various national and marine institutes, fishery associations, as well as by research groups. With the proliferation of underwater observatories, profilers, and autonomous underwater vehicles (AUVs), significant progress has been made in collecting continuous, high-resolution data for in-situ ecological monitoring. However, much of this data remains static and stored in formats such as NetCDF or CSV, making it difficult to integrate into dynamic DTO systems. Furthermore, distribution shifts—variations in the data due to differing collection methods or environmental conditions—pose significant challenges for AI-based systems, which rely on consistent and harmonized data for training and prediction.
Archived and ecological monitoring network data from in-situ robotic and other scientific and societal sources, while essential, are highly heterogeneous and encoded in different formats, posing significant challenges for harmonization and integration. In this context, the Digi4Eco Project (https://digi4eco.eu/the-project/) focuses on addressing the lack of tools to effectively harmonize this vast amount of marine data, ensuring a suitable format for ingestion into DTOs. To address these challenges, our work focuses on developing a comprehensive Open Data Infrastructure (ODI) that adheres to FAIR principles: Findable, Accessible, Interoperable, and Reusable. The ODI will harmonize data typologies, procedures, and instrument specifications, making it easier to process and feed into DTO systems. This pipeline will include automated data validation and quality control mechanisms, following best practices. Special attention will be given to ensuring the data is suitable for AI applications, particularly in solving distribution shift problems.
The proposed ODI integrates existing open-source data services into a modular architecture that covers the entire data and metadata lifecycle. Key components include the SensorThings API for data/metadata storage, ERDDAP for data delivery, Zabbix for system monitoring and alerting, and Grafana for visualizations. Additionally, to ensure long-term impact and community adoption, all procedures and tools developed within this framework will be made open-source and publicly accessible, fostering standardized practices across the marine research community.

Please specify ”other”

In addition, please add 3 to 5 keywords.

DTOs, marine data harmonization, data interoperability, quality control

Please specify ”other”

For whom will your contribution be of most interest?

Data professionals who provide and maintain data infrastructure

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

Researchers

Primary authors: Dr MARTÍNEZ, Enoc (SARTI-UPC); Dr MIHAI TOMA, Daniel (SARTI-UPC); Dr CARANDELL, Matías (SARTI-UPC); Prof. DEL RÍO, Joaquín (SARTI-UPC); AGUZZI, Jacopo (ICM-CSIC)

Presenter: Dr MARTÍNEZ, Enoc (SARTI-UPC)

Session Classification: Session B1

Track Classification: Connecting research data: 6. Interoperable semantics at domain and application level

Contribution ID: **110** Type: **TALK**

Bringing samples to the digital data curation world - the FAIR WISH Project

Tuesday 5 November 2024 10:05 (20 minutes)

Persistent identifiers (PID) are an essential component of digital research data infrastructure. They are used to unambiguously identify, locate, and cite digital representations of a growing range of entities like publications, data, and others. Physical samples represent the basis for many research results and data and are at the same time deeply in the "long tail"of research data. The HMC project "FAIR Workflows to establish IGSN for Samples in the Helmholtz Association"(FAIR WISH) established standardised workflows for sample description and registration of International Generic Sample Numbers (IGSN) in the Earth Science community within the Helmholtz Association. The IGSN is a globally unique, citable PID for physical samples with discovery functionality in the internet.

FAIR WISH has developed (1) standardised and discipline-specific IGSN metadata schemes for different sample types within the Earth and Environment research, (2) workflows to generate machine-actionable IGSN metadata from different states of digitisation, (3) workflows to automatically register IGSNs, and (4) registered more than 35,000 curated metadata sets with IGSNs for the use cases of the three project partners GFZ, AWI and Hereon. We further fully documented the IGSN metadata schema of GFZ representing the current status quo (Brauser et al., 2024) and made the project results available at the website of GFZ Data Services (https://dataservices.gfz-potsdam.de) and the dedicated Zenodo FAIR WISH Community (https://zenodo.org/communities/fair_wish/).

The FAIR SAMPLES Template (Wieczorek et al, 2023) - the main project output - includes the new, extended (IGSN) metadata schema and the controlled vocabularies identified during FAIR WISH. The Excel-built FAIR SAMPLES enables metadata collection and batch upload of sample descriptions at various sample hierarchies (parent, children at different hierarchy levels) at once. The ability to fill the FAIR SAMPLES Template by individual researchers for a wide range of sample types makes the template flexible and widely applicable. The structured metadata, captured with the FAIR SAMPLES Template and converted into XML files, already represents an important step for the standardisation of rich sample descriptions and their provision in machine-actionable form. The new Software Tool SAMIRA: FAIR SAMPLES Template Processing (Frenzel, 2023), enables semi-automated workflows for IGSN registration.

This presentation will introduce the project and its outcomes and describe lessons learned. It will also look forward to possible solutions for fully automated and quality assured metadata generation and collection.

References:

Brauser, A.; Frenzel, S.; Mohammed, A.; Elger, K. (2024): GFZ Metadata Schema for International Generic Sample Numbers (IGSN) and documentation. V. 1.3. GFZ Data Services. https://doi.org/10.5880/GFZ.LIS.2024.001

Frenzel, S. (2024). FAIR WISH Software Tool: SAMIRA: FAIR SAMPLES Template Processing [Computer software]. GFZ Data Services. https://doi.org/10.5880/GFZ.LIS.2023.001

Wieczorek, M., Brauser, A., Frenzel, S., Heim, B., Baldewein, L., Kleeberg, U., & Elger, K. (2023b). FAIR WISH "FAIR SAMPLES Template. Zenodo, https://doi.org/10.5281/zenodo.10436276

Please specify "other"

In addition, please add 3 to 5 keywords.

FAIR samples, IGSN, metadata collection

Please specify ”other”

Researchers and Scientits and technicians who maintain and operate research infrastructure for data generation

For whom will your contribution be of most interest?

other (please specify below)

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

Data professionals and stewards

Primary author: ELGER, Kirsten

Co-authors: BRAUSER, Alexander (Deutsches GeoForschungsZentrum (GFZ) Potsdam); HEIM, Birgit (Alfred-Wegener-Institut Helmholtz Zentrum für Polar- und Meeresforschung); BALDEWEIN, Linda (Helmholtz-Zentrum Hereon); WIECZOREK, Mareike (Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung); FRENZEL, Simone Christina (GFZ); KLEEBERG, Ulrike

Presenter: ELGER, Kirsten

Session Classification: Session C1

Track Classification: Assessing and monitoring FAIR data: 1. Human actors in the FAIR data landscape

Contribution ID: **111** Type: **POSTER&PITCH**

Empowering community-driven change and developments towards a FAIR data future in agrosystem science - First Evidence from the NFDI initiative FAIRagro

Monday 4 November 2024 16:00 (1 hour)

In agrosystem science, the transition to a FAIR (Findable, Accessible, Interoperable, Reusable) data future is essential for fostering innovation and collaboration. While technical developments provide the necessary infrastructure, the true challenge lies in changing ingrained habits and cultural practices. To address this, the FAIRagro initiative has developed a participation concept aimed at empowering community-driven change, incorporating various measures at both institutional and individual levels.

With this contribution, we would like to share insights from an intensive stakeholder and target group analysis, which forms the groundwork for FAIRagro's community engagement and participatory efforts. We also reflect on the effectiveness and impact of different measures and explore what is needed to make a meaningful contribution towards a FAIR data future in agrosystem science.

The measures implemented by FAIRagro include, among others, governance instruments such as the FAIRagro Community Advisory Boards (CAB), institutional participation through Use Cases (UCs), and individual interaction via FAIRagro data stewards from the Data Steward Service Centre (DSSC). Insights from these and other community measures and activities are presented, along with relevant indicators to evaluate their effectiveness and impact.

Preliminary evidence indicate that trust and ownership are fundamental for effectively advancing community-driven initiatives. The importance of being FAIR role models, or lighthouses, at the institutional level, and the influence of individual FAIR ambassadors, is often underestimated. These insights underscore the need for engagement that spans from individual interactions to institutional commitments and vice versa, with a focus on making FAIR principles tangible and experiential. Nevertheless, identifying meaningful key performance indicators continues to be a challenge and demands further collaborative efforts across various initiatives and the broader community.

Please specify "other"

Sci project manager

In addition, please add 3 to 5 keywords.

community-driven, community participation, meaningful KPIs

Please specify "other"

For whom will your contribution be of most interest?

Expert panels, strategists and administrative stakeholders

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

Expert panels, strategists and administrative stakeholders

Primary author: SENNHENN, Anne (Leibniz Institute for Agricultural Engineering and Bioeconomy (ATB))

Presenter: SENNHENN, Anne (Leibniz Institute for Agricultural Engineering and Bioeconomy (ATB))

Session Classification: Poster Session C

Track Classification: From recommendations to implementations: 10. Enabling and incentivising community-driven initiatives

Contribution ID: **112** Type: **TALK**

Enhancing Research Data Annotation: The SEPIA Sample Database for Metadata Storage and Exchange

Tuesday 5 November 2024 11:00 (20 minutes)

The SEPIA project aims to improve the management and annotation of research data by providing a comprehensive sample database integrated with an open API. This initiative facilitates the capture and exchange of sample metadata, thereby enriching the research data collected at the Helmholtz-Zentrum Berlin (HZB). This presentation will explore the architecture and functionalities of the SEPIA project, highlighting its role in improving data accessibility and interoperability among researchers. We will discuss the potential impact of SEPIA on collaborative research efforts, data sharing practices, and the overall improvement of scientific research through better metadata management.

Please specify ”other”

Web Developer

In addition, please add 3 to 5 keywords.

Sample Information, Sample PID, Sample Database

Please specify ”other”

For whom will your contribution be of most interest?

Data professionals who provide and maintain data infrastructure

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

other (please specify)

Primary author: KRAHL, Rolf (Helmholtz-Zentrum Berlin für Materialien und Energie)

Co-authors: Ms RIAL, Katherine (Helmholtz-Zentrum Berlin); SEDEQI, Mojeeb Rahman (Helmholtz-Zentrum Berlin für Materialien und Energie GmbH (HZB), Helmholtz Metadata Collaboration (HMC))

Presenters: Ms RIAL, Katherine (Helmholtz-Zentrum Berlin); SEDEQI, Mojeeb Rahman (Helmholtz-Zentrum Berlin für Materialien und Energie GmbH (HZB), Helmholtz Metadata Collaboration (HMC)); KRAHL, Rolf (Helmholtz-Zentrum Berlin für Materialien und Energie)

Session Classification: Session D2

Track Classification: Connecting research data: 5. Technical solutions for findable and machine-readable metadata

Contribution ID: **113** Type: **TALK**

M3eta: An extensible metadata scheme for advanced momentum microscopy in the age of big data

Monday 4 November 2024 12:25 (20 minutes)

The electronic structure determines many of the macroscopic physical properties of a material. Photoelectron momentum microscopy (MM) has matured into a powerful tool for the detailed characterization of the exciting electronic properties of novel quantum materials. By applying the principles of high-resolution imaging modern instruments simultaneously capture hundreds of tomographic slices of an electronic structure in a high-dimensional parameter space. Despite the rapid worldwide adoption as a universal tool for material characterization, there is currently no common scheme to describe the highly diverse parameters that define a MM experiment. M^3eta aims to establish an extensible and sustainable metadata scheme for momentum microscopy, which will be stored in a structured file together with the measured data voxels. This will be the basis for a standardized work-flow that interprets the stored metadata and to reconstruct views of the multidimensional electronic structure of a material.

As a test bed for the handling of multidimensional electronic structure information, we have created a set of tomographic slices through the Fermi surface of the noble metal palladium. By use of linearly polarized synchrotron radiation with photon energies between 34eV and 200eV the perpendicular crystal momentum coordinate (k_z) is scanned over the entire first Brillouin zone, while the transverse momentum (k_x,k_y) is simultaneously recorded by the momentum microscope setup at different binding energies (E_B). The resulting 4D data volume (k_x,k_y,k_z,E_B) serves as a prototypical example of multidimensional Photoemission data and allows us to test data/metadata structures for an analysis and visualization workflow.

Representation of such rich data sets and connection with experiment specific meta data will be discussed don the example of the NeXus format. The format will allow to enrich the measured data voxels with information for their physical interpretation and allow work flows to extract and visualize electronic structure information.

Please specify ”other”

In addition, please add 3 to 5 keywords.

momentum microscopy, photoemission, multidimensional, visualization, physical representation

Please specify ”other”

For whom will your contribution be of most interest?

Researchers

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

Researchers

Primary authors: TUSCHE, Christian (Forschungszentrum Jülich); Dr SCHLUETER, Christoph (Deutsches Elektronen Synchrotron DESY); Prof. SCHNEIDER, Claus (Forschungszentrum Jülich); Dr HOESCH, Moritz (Deutsches Elektronen Synchrotron DESY)

Presenter: TUSCHE, Christian (Forschungszentrum Jülich)

Session Classification: Session B2

Track Classification: Connecting research data: 4. Metadata annotation and management

Contribution ID: **114** Type: **TALK**

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

Tuesday 5 November 2024 11:40 (20 minutes)

Research in the Helmholtz Association is carried out in inter- and multidisciplinary collaborations that span between its 18 independently operating research centers across Germany. Helmholtz digital infrastructure is institutional, and thus Helmholtz's research data and other digital assets are stored and maintained in independent data infrastructures, lacking visibility and accessibility. As a consequence their full value remaines unavailable to scientists, managers, strategists, and policymakers.
The Helmholtz Metadata Collaboration (HMC) is taking on this challenge by establishing a Helmholtz FAIR Data Space. As part of this, we develop the Helmholtz Knowledge Graph (Helmholtz KG) 1 as a lightweight interoperability layer that connects Metadata Helmholtz digital assets, which are stored in a decentralized manner. With this KG, we envision (1) providing better cross-organizational access to Helmholtz's (meta)data and information assets on an upper semantic level, (2) harmonizing and optimizing the related metadata across the association, and (3) forming a basis from which the semantic quality and the depths of metadata descriptions is improved and extended into domain and application levels.
In the initial phase, we focused on establishing a working system that (1) contains harvesting pipelines 2 as demonstrators, (2) a User Interface 3 to explore the data, and (3) a SPARQL endpoint 4 to query the graph. Currently, the system harvests and aggregates data from more than 30 data providers. We are further developing the code base to reach a higher maturity level and increase the scalability of the infrastructure in order to accommodate further resources in the future –namely all open data repositories and infrastructures. At the same time, we work on the data-level in order to harmonize metadata representation across Helmholtz. This establishes common standards for Helmholtz data providers and harmonizes metadata where it is stored. For this, HMC established unHIDE: the Unified Helmholtz Information and Data Exchange as a network between Helmholtz data and infrastructure providers.
In the presentation, we will show the status quo of the Helmholtz KG as well as future development avenues and potentials to join forces with infrastructure providers and users.

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

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

Please specify "other"

In addition, please add 3 to 5 keywords.

Knowledge Graph, Semantic Interoperability, Infrastructure,

Please specify ”other”

For whom will your contribution be of most interest?

Data professionals and stewards

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

Data professionals who provide and maintain data infrastructure

Primary authors: D’MELLO, Fiona (Forschungszentrum Jülich); PREUSS, Gabriel (Helmholtz Zentrum Berlin für Materialien und Energie, Berlin, Germany); KULLA, Lucas (DKFZ); SOYLU, Mustafa (Forschungszentrum Jülich); FATHALLA, Said; SANDFELD, Stefan; HOFMANN, Volker

Presenter: HOFMANN, Volker

Session Classification: Session D2

Track Classification: Connecting research data: 7. Infrastructure and common practices for consolidation of (meta)data

Contribution ID: **115** Type: **TALK**

The EM Glossary: a community effort towards a harmonised terminology in electron microscopy

Monday 4 November 2024 11:10 (20 minutes)

For data to be fully exploitable and re-usable in different contexts it needs to be annotated with rich metadata that uses commonly understood vocabularies and semantics 1. Using terminology that is standardized and agreed upon within a community ensures unambiguous understanding of metadata.
In the field of EM, a number of application-level initiatives independently started developing metadata schemas [e.g. 2,3], to describe experimental equipment, workflows, and analysis procedures. Domain-level semantic harmonisation of such efforts is required to ensure data interoperability down the line. As a first step, misalignment of terminology has to be addressed by mapping concepts of different scientific contexts, and user groups.
The Helmholtz Metadata Collaboration (HMC) 4 is currently coordinating an effort, the EM glossary group 5, to establish a documented terminology for electron and ion microscopy (IM). This community involves scientists from more than 23 institutions across Switzerland, Austria, and Germany and representatives of the FAIRmat and the MatWerk NFDI consortia.

In a remote collaborative workflow and bi-weekly online meetings, we work towards formulating consensus on terms that are commonly used in the EM and IM communities. Here we produce concise, unpacked definitions with rich annotations in accordance with semantic best practices. By now, we provide harmonized definitions for more than 60 terms which can be explored via a web interface 6.

For implementation and machine readability the glossary is further implemented as an OWL ontology [7]. For this we use a fully automated workflow in separate, but coupled gitlab repository which translates novel terms from a community repository into OWL. Then releases are triggered depending on the progress of terminology extension by the community. Both these representations are intended as a central resource to map and align application-level semantics, thereby acting as semantic glue within the field.

Interested to get involved? See 5 and send an email to hmc@fz-juelich.de to get in touch!

Acknowledgements
This work was supported by (1) the Helmholtz Metadata Collaboration (HMC), an incubator-platform of the Helmholtz Association within the framework of the Information and Data Science strategic initiative and the NFDI consortia “MatWerk”and “FAIRmat”, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the National Research Data Infrastructure –NFDI 38/1 –project number 460247524 (MatWerk) & 460197019 (FAIRmat)

References:
1 Wilkinson, M.D. et al. Scientific Data. https://dx.doi.org/10.1038/sdata.2016.18
2 Könnecke, M, et al. Journal of applied crystallography https://doi.org/10.1107/S16005767140275753
Joseph, R, et al doi.org/10.5445/IR/10001416044 Helmholtz Metadata Collaboration: https://helmholtz-metadaten.de/en5 EM Glossary Group https://codebase.helmholtz.cloud/em_glossary/em_glossary (development repository)6 EM Glossary User Interface: https://emglossary.helmholtz-metadaten.de/ (temporary demonstrator)[7] EM Glossary OWL: https://owl.emglossary.helmholtz-metadaten.de/

Please specify ”other”

In addition, please add 3 to 5 keywords.

Metadata Harmonization, Ontology development, Community Project,

Please specify ”other”

For whom will your contribution be of most interest?

Researchers

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

Data professionals who provide and maintain data infrastructure

Primary authors: MANNIX, Oonagh (HMC matter/HZB); PAULY, Christoph; KÜHBACH, Markus; WOLLGARTEN, Markus; KONIJNENBERG, Peter (FZ Juelich IAS-9); RASMUS, Schroeder; AVERSA, Rossella (Karlsruhe Institute of Technology); BROCKHAUSER, Sandor; ÖZKAN, Özlem; SEDEQI, Mojeeb Rahman (Helmholtz-Zentrum Berlin für Materialien und Energie GmbH (HZB), Helmholtz Metadata Collaboration (HMC)); AZOCAR-GUZMAN, Abril; HOFMANN, Volker; SANDFELD, Stefan

Presenter: MANNIX, Oonagh (HMC matter/HZB)

Session Classification: Session A2

Track Classification: Connecting research data: 6. Interoperable semantics at domain and application level

Contribution ID: **116** Type: **POSTER&PITCH**

The HMC FAIR Data Dashboard: A Data-Driven Approach to Monitor and Improve FAIR Data in the Helmholtz Association

Monday 4 November 2024 14:00 (1 hour)

Here we present the latest updates of our data-driven approach to monitoring and assessing the state of open and FAIR data in the Helmholtz Association. The approach consists of two parts: a modular data harvesting-, validation- and assessment pipeline, and a dashboard with interactive statistics about the Helmholtz-data publications identified. The dashboard provides insight into which data repositories research communities use to publish research data and allows for assessing systematic gaps of this data with respect to selected FAIR data principles. We illustrate how the approach can be used to engage communities and infrastructure towards an improved FAIR data landscape.

Please specify ”other”

In addition, please add 3 to 5 keywords.

dashboard, data-mining, data publications, Scholix, F-UJI

Please specify ”other”

For whom will your contribution be of most interest?

Data professionals who provide and maintain data infrastructure

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

Data professionals and stewards

Primary author: SEDEQI, Mojeeb Rahman (Helmholtz-Zentrum Berlin für Materialien und Energie GmbH (HZB), Helmholtz Metadata Collaboration (HMC))

Co-authors: Mr EHLERS, Pascal (German Aerospace Center (DLR)); Mr SCHMIDT, Alexander (Helmholtz-Zentrum Berlin für Materialien und Energie GmbH (HZB)); Ms SERVE, Vivien (Helmholtz-Zentrum Berlin für Materialien und Energie GmbH (HZB), Helmholtz Metadata Collaboration (HMC)); GILEIN, Astrid (Helmholtz-Zentrum Berlin für Materialien und Energie GmbH (HZB)); Mrs GLODOWSKI, Tempest (Helmholtz-Zentrum Berlin für Materialien und Energie GmbH (HZB)); PREUSS, Gabriel

(Helmholtz-Zentrum Berlin für Materialien und Energie GmbH (HZB), Helmholtz Metadata Collaboration (HMC)); MANNIX, Oonagh (Helmholtz-Zentrum Berlin für Materialien und Energie GmbH (HZB), Helmholtz Metadata Collaboration (HMC)); KUBIN, Markus (Helmholtz-Zentrum Berlin für Materialien und Energie GmbH (HZB), Helmholtz Metadata Collaboration (HMC))

Presenter: SEDEQI, Mojeeb Rahman (Helmholtz-Zentrum Berlin für Materialien und Energie GmbH (HZB), Helmholtz Metadata Collaboration (HMC))

Session Classification: Poster Session A

Track Classification: Assessing and monitoring FAIR data: 2. Metrics, tools and workflows for metadata assessment

Contribution ID: **117** Type: **POSTER&PITCH**

Harmonizing the Implementations of PIDs across Repositories

Monday 4 November 2024 16:00 (1 hour)

In our increasingly digital and interconnected world, the integration of Persistent Identifiers (PIDs) in metadata are essential for machine-readable and -understandable metadata as also described in the FAIR Guiding Principles for research data management. PIDs provide unique, permanent and machine-readable references to various types of digital objects, including publications, datasets, scientific software, individuals, organizations, samples that together represent the broad range of research outcomes.
Within the AK Metadata-PIDs working group (a joint initative between the HMC Hub Earth and Environment and the Helmholtz DataHub Earth and Environment), we discussed several PID systems and reached a consensus on recommending specific systems for different purposes: “ORCID” for identifying individuals, “ROR” for organizations, and the “PIDINST” PID for instruments.
For the full integration and sustainability of individual PIDs, different players are involved: these range from the consortium developing a PID to the research institution supporting (and ideally enforcing) its implementation for their employees (e.g., ORCID) to the individual research infrastructures and repositories where the required information is collected and the provision of the PID within the metadata is ensured.
Our working group has focused on supporting the ongoing PID implementation in research infrastructures by conserving existing, well-established PID implementations (best practices) and promoting their integration in future systems. We further aim to provide support and guidance for new implementations.
We observe differences in the metadata content, even between two DOI-registering data repositories that use the DataCite Schema. What is the reason for this? Would we need specific mapping tables to harmonise cross-repository metadata? Could a stronger guidance and definition of metadata properties (note: metadata schemas are intended to be very generic) achieve the envisioned higher grade of harmonization?
This poster highlights and discusses the differences in PID implementations across various exchange formats, such as DataCite, ISO 19115/19139, and schema.org. The goal is to encourage and support scientists and IT specialists who maintain and develop research infrastructure to foster metadata harmonization for ensuring interoperable metadata exchange across repositories.

Please specify ”other”

HMC staff

In addition, please add 3 to 5 keywords.

PID, DataCite, ISO, Schema, ORCID, ROR

Please specify ”other”

For whom will your contribution be of most interest?

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

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

other (please specify)

Primary authors: PÖRSCH, Andrea (HMC Hub EE at GFZ); SÖDING, Emanuel (GEOMAR)

Co-authors: KOTTMEIER, Dorothee (HMC E&E @PANGAEA/AWI); MALINOVSCHII, Stanislav (HMC); LORENZ, Sören; RAZEGHI, Yousef

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

Session Classification: Poster Session C

Track Classification: Connecting research data: 6. Interoperable semantics at domain and application level

Contribution ID: **118** Type: **POSTER&PITCH**

Graduate - An intuitive user interface for modeling semantic graph data

Monday 4 November 2024 16:00 (1 hour)

Research data management (RDM) is an important aspect of modern scientific research, which is heavily relying on interconnected data sets and corresponding metadata. For modeling and integrating these interconnections and metadata, the Resource Description Framework (RDF) has often been proposed as a standard, since it has been in use by search engines and knowledge management systems for decades by now.

The RDF provides a graph data structure for enriched vocabularies, so-called ontologies, and therein expressed information. However, the complexity of RDF and ontologies can be a barrier to adoption, especially for those without extensive training. To overcome this challenge, we are developing Graduate, a software tool that allows users to create RDF graph data in a user-friendly, graphical interface.

Graduate provides an intuitive visual representation of RDF graph data as an editable diagram, allowing users to create and modify RDF triples with minimal training. The software supports the use of terms from ontologies and enables users to create rich, structured data that conforms to established standards.

In addition it allows for versioning, sharing and collaborative work on graph datasets via GitLab. This can facilitate interdisciplinary research and collaboration, allowing researchers to work together on datasets and share their findings with a wider audience. The software's integration with GitLab also allows researchers to track changes via version control, improving the reproducibility and transparency of their research.

In summary, Graduate represents a significant advance in RDM technology and provides a user-friendly interface for the creation and management of RDF graph data. The visual representation of RDF data in the software provides an intuitive way to understand and engage with complex data structures, supporting knowledge transfer in teaching and research.

Please specify "other"

In addition, please add 3 to 5 keywords.

semantic
graph data
RDF
user interface

Please specify "other"

For whom will your contribution be of most interest?

Data professionals and stewards

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

Data professionals and stewards

Primary author: STEINMEIER, Leon (Helmholtz Institute Freiberg)

Presenter: STEINMEIER, Leon (Helmholtz Institute Freiberg)

Session Classification: Poster Session C

Track Classification: Connecting research data: 4. Metadata annotation and management

Contribution ID: **119** Type: **POSTER&PITCH**

Sharing the load - defining responsibilities for common data elements to the appropriate stakeholders in data management

Monday 4 November 2024 15:00 (1 hour)

At the Helmholtz Association, we strive to establish a well-formed harmonized data space, connecting information across distributed data infrastructures. This requires standardizing the description of data sets with suitable metadata to achieve interoperability and machine actionability.
One way to make connections between datasets and to avoid redundancy in metadata is the consistent use of Persistent Identifiers (PIDs). A lot of information within the metadata such as people, organizations, projects, laboratories, repositories, publications, vocabularies, samples, instruments, licenses, and methods should be commonly referenced by PIDs, but not for all of these agreed identifiers exist yet.
Typically, researchers who are publishing datasets are also tasked with compiling the metadata for those datasets. However, researchers are usually not in charge of a lot of information that should be part of the documentation of a dataset. They often have to rely on information they receive from other sources, e.g. technicians, responsible for the measuring devices or librarians, who are experts in assigning licenses. Starting from PID Systems ROR, ORCID, IGSN, PIDInst, DataCite-DOI and CrossRef DOI we suggest to share the load, and assign certain expert stakeholder groups responsibility to maintain specific information and to conduct certain tasks within the research data management (RDM) workflow.
The conclusions from this process do not only affect the implementation of PID metadata, but may also be used for the harmonization of vocabularies, digital objects, interfaces, licenses, quality flags and others, in order to connect our global data systems, to redefine stakeholder responsibility and to ultimately reach the data space.

Please specify ”other”

In addition, please add 3 to 5 keywords.

PID, Role, Stakeholders, Data Management

Please specify ”other”

For whom will your contribution be of most interest?

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

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

Data professionals and stewards

Primary authors: PÖRSCH, Andrea (HMC Hub EE at GFZ); KOTTMEIER, Dorothee (HMC E&E @PANGAEA/AWI); SÖDING, Emanuel (GEOMAR); MALINOVSCHII, Stanislav (HMC); LORENZ, Sören; RAZEGHI, Yousef

Presenter: SÖDING, Emanuel (GEOMAR)

Session Classification: Poster Session B

Track Classification: Assessing and monitoring FAIR data: 1. Human actors in the FAIR data landscape

Contribution ID: **120** Type: **TALK**

Managing research data in plant sciences through the DataPLANT ontology service landscape

Tuesday 5 November 2024 13:00 (20 minutes)

The DataPLANT consortium, a German National Research Data Infrastructure (NFDI), aims to provide plant researchers a robust and sustainable infrastructure for managing research data. Since the complexity of research data continues to grow, effective methods for managing, annotating, and sharing this data becomes increasingly important. DataPLANT integrates different established concepts for FAIR research data management and ontologies to provide tools and services to aid plant researchers in their research data management (RDM).

At the core of the DataPLANT infrastructure is the Annotated Research Context (ARC), a data-centric approach to capturing and structuring the entire research cycle. By leveraging the ISA (Investigation-Study-Assay) standard, Research Object Crate, and Common Workflow Language, the ARC serves as a standardized and comprehensive method for researchers to document their experimental designs, protocols, workflows, and data in a structured format. By utilizing Git services, data provenance is tracked, facilitating collaboration between multiple researchers involved in a common project.

To assist researchers withthe ARC creation and data annotation, theSwate tool, a spreadsheet-based softwarewas developed,whichallows researchers to annotate their data with standardized metadata. This process leverages selected ontologies relevant in plant research, which are stored in a database (SwateDB) and linked to the Swate tool via an API, allowing users to search for specific terms that fit their needs. Inaddition, DataPLANT manages the curation of the DataPLANT biology ontology (DPBO), a broker ontology that fills in gaps by providing missing terms not yet availabe in existing ontologies. SwateDB updates occur through the Swate OBO Updater (Swobup) via Git repository changes, ensuring that researchers have access to the most up-to-date ontologies. Making further use of Git's capabilities, users can easily request new terms during their annotation process and contribute to the SwateDB, either through opening new issues, or through direct contributions via pull requests. The request for the addition of a new term will then be reviewed by the DataPLANT team and incorporated into the DPBO to immediately provide the user with the option to add their term in their metadata spreadsheets. Each newly added term immediately gets a new persistent identifier to serve as an immutable link to this term. As a long-term solution-for maintaining the new terms, each new addition will be evaluated individually and pushed to existing ontologies, which have a defined scope that should include this term. If a term is accepted by an external ontology, the original DPBO term will be deprecated and linked to the new term in the external ontology. In the future, this process will be improved by automating the term reading from the spreadsheets and creating new terms in DPBO for every metadata term that was not already taken from the SwateDB. Furthermore, ontologies from other research areas can be easily integrated into the current framework, making it a flexible resource for guiding scientist through their RDM processes.

With our approach, we show that standards such as ISA in combination with ontologies can be efficiently used across all life science domains for (meta)data annotation.

Please specify "other"

In addition, please add 3 to 5 keywords.

ontologies, RDM, DataPLANT, ARC

Please specify ”other”

researchers and technicians in their day-to-day lab work, data professionals who provide and maintain infrastructure, data professionals and stewards

For whom will your contribution be of most interest?

other (please specify below)

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

Researchers

Primary author: DOERPHOLZ, Hannah (Forschungszentrum Jülich)

Presenter: DOERPHOLZ, Hannah (Forschungszentrum Jülich)

Session Classification: Session E1

Track Classification: Connecting research data: 4. Metadata annotation and management

Contribution ID: **121** Type: **POSTER&PITCH**

Interactively exploring metadata with Beaverdam

Monday 4 November 2024 15:00 (1 hour)

Scientists frequently need to get an overview of their experiments by summarizing information spread over multiple files and storage locations. This metadata may include items such as experimental conditions, subject details, and characteristics of the experimental data. It is common for researchers to spend time developing their own solutions tailored to their specific use case. However, overviews of metadata have similar requirements across research fields. We leveraged these similarities to develop generic software for efficiently exploring collections of metadata, which scientists can quickly customize for their own work.

Our software Beaverdam (Build, Explore, And Visualize ExpeRimental DAtabases of Metadata) combines metadata from multiple experiments into a central database, then builds an interactive dashboard to explore the contents of the database. Graphs show a high-level overview of multiple experiments, a table shows details of each experiment, and interactive filters help researchers identify experiments meeting specific criteria. Users customize the dashboard using a single configuration file. We developed Beaverdam in Python and have released it as a Python package which users can run from the command line or incorporate into their own code.

Although we designed Beaverdam for all sizes of datasets, its automated approach is particularly useful for datasets with many experiments and/or extensive metadata. We tested Beaverdam with metadata from a neuroscience dataset in which each of the hundreds of experimental sessions contains thousands of items of metadata. Using Beaverdam, researchers were able to efficiently identify experimental sessions meeting their criteria for further analysis – a task that would have been impossible by hand.

We expect that Beaverdam will help scientists efficiently explore their metadata, identify gaps, and inform further analyses.

Beaverdam on GitHub (open source): https://github.com/INM-6/beaverdam

Funding: This work is supported by the Helmholtz Metadata Collaboration (HMC), EU Grant 945539 (HBP SGA3), and the NRW-network iBehave (NW21-049).

Please specify ”other”

In addition, please add 3 to 5 keywords.

database, metadata, software, visualization, Python

Please specify ”other”

For whom will your contribution be of most interest?

Researchers

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

Data professionals and stewards

Primary authors: MORE, Heather (Institute for Advanced Simulation (IAS-6 and IAS-9), Forschungszentrum Jülich); DENKER, Michael (INM-10, Forschungszentrum Jülich); Prof. GRUEN, Sonja (FZJ); SANDFELD, Stefan; HOFMANN, Volker

Presenter: MORE, Heather (Institute for Advanced Simulation (IAS-6 and IAS-9), Forschungszentrum Jülich)

Session Classification: Poster Session B

Track Classification: Connecting research data: 7. Infrastructure and common practices for consolidation of (meta)data

Contribution ID: **122** Type: **POSTER&PITCH**

Connecting information across repositories –a keyword-based approach

Monday 4 November 2024 15:00 (1 hour)

Knowledge Graphs help to connect and organize information from different sources and entities. They can be used to apply advanced search and filtering techniques on very large datasets and reveal connections and dependencies across the data. To be useful, however, they require highly uniform and harmonized data sets. So far, most knowledge graphs on scientific data have used bibliographic data to build a network of information. These data are of limited use for scientific purposes because they contain little scientifically relevant information. In order to enhance the scientific usability in the Helmholtz research area Earth and Environment, we aim to identify seven parameters in data sets and build a knowledge graph from it:

```
Measuring Instrument (type, manufacturer, model)
Methodology
Measured Attribute (e.g. sulfur content)
Measured Parameter (e.g. MS Spectrum)
Measured Unit (e.g. velocity)
Measured Object / Medium (e.g. rock sample)
Sample ID (e.g. as IGSN)
```

DataCite, ISO191XX and schema.org are among the most common standards currently implemented by repositories, to retrieve and exchange metadata. However, most of the mentioned parameters are not yet well documented in the common metadata standards used to export data from repositories. Repositories thus apply very different approaches to include this information within their metadata.
In this poster we discuss our approaches, challenges and successes to harvest this information from several repositories from the Helmholtz Earth and Environment research field. We also discuss the potential to create knowledge graphs from this data, and how the quality of these graphs can be improved. Finally, we present some statistics on the harvested data and make suggestions on how the data can be improved.

Please specify ”other”

In addition, please add 3 to 5 keywords.

Knowledge graph, metadata harmonization, PID, Metadata schema, semantics

Please specify ”other”

For whom will your contribution be of most interest?

Researchers

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

Data professionals and stewards

Primary authors: PÖRSCH, Andrea (HMC Hub EE at GFZ); KOTTMEIER, Dorothee (HMC E&E @PANGAEA/AWI); SÖDING, Emanuel (GEOMAR); MALINOVSCHII, Stanislav (HMC); LORENZ, Sören; RAZEGHI, Yousef

Presenters: SÖDING, Emanuel (GEOMAR); MALINOVSCHII, Stanislav (HMC)

Session Classification: Poster Session B

Track Classification: Connecting research data: 5. Technical solutions for findable and machine-readable metadata

Contribution ID: **123** Type: **POSTER&PITCH**

Standardised Metadata Provision in the Communication Protocol SECoP - SECoP@HMC

Monday 4 November 2024 16:00 (1 hour)

The Sample Environment Communication Protocol (SECoP) provides a generalized way for controlling measurement equipment –with a special focus on sample environment (SE) equipment [1,2]. 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 and
- self-explaining providing a machine readable description of all available data and metadata.

The project SECoP@HMC focuses on the standardised provision of metadata for typical SE equipment at large scale facilities (photons, neutrons, high magnetic fields) and on standardized metadata storage. The fact that SECoP is self-explaining and machine-readable favours the automated interpretation of data and metadata. With the latest definition of SECoP, we were able to integrate the use of vocabularies or glossaries.

With the ongoing development of SECoP and the provision of several tools for its easy implementation, a complete standardized system for controlling SE equipment and collecting and saving SE metadata is available and usable for experimental control systems. This approach can be applied to other research areas as well.

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

1 K. Kiefer, et al. (2020). An introduction to SECoP –the sample environment communication protocol. Journal of Neutron Research, 21(3-4), pp.181–195
2 https://github.com/sampleenvironment/secop

Please specify ”other”

In addition, please add 3 to 5 keywords.

sample environment, communication protocol, machine readable

Please specify ”other”

For whom will your contribution be of most interest?

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

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

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

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

Co-authors: ZAFT, Alexander (Forschungzentrum Jülich); PETTERSSON, Anders (European Spallation Source); KLEMKE, Bastian (Helmholtz-Zentrum Berlin); FAULHABER, Enrico (Forschungs-Neutronenquelle Heinz Maier-Leibnitz); BRANDL, Georg (Forschungzentrum Jülich); GUENTHER, Gerrit (Helmholtz-Zentrum Berlin); KOTANSKI, Jan (Deutsches Elektronen Synchrotron DESY); ROSSA, Lutz (Helmholtz-Zentrum Berlin); UHLARZ, Marc (Helmholtz-Zentrum Dresden Rossendorf); ZOLLIKER, Markus (Paul Scherrer Institut); EKSTRÖM, Niklas (European Spallation Source); BRAUN, Peter (Helmholtz-Zentrum Berlin); KRACHT, Thorsten (Deutsches Elektronen Synchrotron DESY)

Presenter: KIEFER, Klaus (Helmholtz-Zentrum Berlin)

Session Classification: Poster Session C

Track Classification: Connecting research data: 5. Technical solutions for findable and machine-readable metadata

Contribution ID: **124** Type: **TALK**

Using application-level ontologies in materials simulation workflows

Tuesday 5 November 2024 13:40 (20 minutes)

The microstructure of materials is characterized by crystallographic defects, which ultimately determine the material properties. In computational materials science, methods and tools are used to predict and analyze defect structures. The increase of computational power has led to the generation of large amounts of complex and heterogeneous data, increasing the need for the implementation of data-driven approaches. In the field of atomistic simulations, we currently face several challenges that impair data reusability: (1) to facilitate the understanding and use of computational samples (or atomic structures), well-described and harmonized metadata and data are crucial. However, most existing approaches focus on perfect crystal structures only, i.e. neglecting defects. (2) Calculations often involve a combination of different software tools and various file formats. This results in heterogeneous metadata which leads to a lack of semantic interoperability. (3) The workflow provenance that was used to set up a digital sample is frequently lacking.
To address these problems and facilitate data re-use in the field, we have developed the Computational Materials Sample Ontology (CMSO), an application-level ontology for material science computational samples. CMSO initially focuses on describing structures at the atomistic level 1. The use of the CMSO ontology is complemented by the development of domain-level ontologies describing crystallographic defects 2 and atomistic simulation concepts.
Importantly, to aid domain scientists in implementing ontologies in their everyday research, we developed software tools for the automated annotation and identification of structural features. AtomRDF 3 provides a way for users to automatically annotate their data with ontologies and create application-level knowledge graphs. This improves the querying and findability of their research data. In addition, atomID 4 showcases the use of ontologies in identification processes frequently performed in materials simulations.
The here shown combination of controlled vocabularies and software tools for generating linked open data ensures interoperability between different file formats and software, while also offering the potential for data to be findable and reusable 5.

References
1 https://purls.helmholtz-metadaten.de/cmso/
2 https://github.com/OCDO/
3 https://github.com/pyscal/atomRDF
4 https://github.com/Materials-Data-Science-and-Informatics/atomID
5 Wilkinson, M., Dumontier, M., Aalbersberg, I. et al., Sci Data, 2016, 3, 160018.

Please specify ”other”

In addition, please add 3 to 5 keywords.

domain ontology, materials science, semantic interoperability

Please specify ”other”

For whom will your contribution be of most interest?

Researchers

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

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

Primary author: AZOCAR GUZMAN, Abril (IAS-9, FZJ)

Co-authors: MENON, Sarath (Max-Planck-Institut für Eisenforschung GmbH); HOFMANN, Volker; Dr HICKEL, Tilmann (BAM); SANDFELD, Stefan

Presenter: AZOCAR GUZMAN, Abril (IAS-9, FZJ)

Session Classification: Session E1

Track Classification: Connecting research data: 6. Interoperable semantics at domain and application level

Contribution ID: **125** Type: **POSTER&PITCH**

Enhancing Metadata Handling in Research Software

Monday 4 November 2024 16:00 (1 hour)

The rapid evolution of research software necessitates efficient and accurate metadata management to ensure software discoverability, reproducibility, and overall project quality. However, manually curating metadata can be time-consuming and prone to errors. This poster presents two innovative tools designed to streamline and improve metadata management: *fair-python-cookiecutter* and *somesy*.

fair-python-cookiecutter is a GitHub repository template that provides a structured foundation for Python projects. It aids researchers and software developers in meeting the growing demands for comprehensive software metadata during Python tool and library development. By adopting this template, developers gain access to best practices, software development recommendations, and essential metadata standards that enhance software quality and ensure proper citation. The template aligns with standards such as the DLR Software Engineering Guidelines, OpenSSF Best Practices, REUSE, CITATION.cff, and CodeMeta. Additionally, it incorporates somesy to enhance the FAIRness (Findability, Accessibility, Interoperability, and Reusability) of software metadata. *fair-python-cookiecutter* is regularly maintained and updated in response to new library versions and user feedback, ensuring it remains a robust and up-to-date resource.

somesy (**So**ftware **Me**tadata **Sy**nchronization) is a user-friendly command-line tool that simplifies the synchronization of software project metadata. Supporting key metadata standards like CITATION.cff and CodeMeta, somesy ensures the consistency and integrity of crucial project information, such as names, versions, authors, and licenses, across multiple files. This allows developers to focus on their core work without the burden of manual metadata upkeep. somesy is compatible with Linux, Windows, and macOS, providing cross-platform support for a wide range of development environments. It is regularly updated based on user feedback and the latest developments in software libraries.

Links:
1. https://pypi.org/project/fair-python-cookiecutter/
2. https://pypi.org/project/somesy/

Please specify ”other”

software developer

In addition, please add 3 to 5 keywords.

python template metadata

Please specify ”other”

For whom will your contribution be of most interest?

Researchers

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

other (please specify)

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

Co-authors: SANDFELD, Stefan; HOFMANN, Volker

Presenter: SOYLU, Mustafa (Forschungszentrum Jülich)

Session Classification: Poster Session C

Track Classification: Connecting research data: 4. Metadata annotation and management

Contribution ID: **126** Type: **POSTER&PITCH**

SmartER: Synergizing Metadata from Scholarly Repositories to Support Research Data Management

Monday 4 November 2024 14:00 (1 hour)

There has been substantial increase in number of scientific publications across diverse disciplines. These publications often generate metadata, scholarly content and scientific models/source code etc. Though such information is made available to research communities under open science initiative, numerous scholarly repositories have emerged over the years to harvest metadata in various exciting aspects. For example, DBLP, ORCID, ROR and arXiv systematically provide access to scholarly metadata features and content respectively.
The Digital Bibliography and Library Project (DBLP) offers free access to metadata features within the field of informatics and interdisciplinary research. The Research Organization Registry (ROR) collects information on research organizations globally. Additionally, the Open Researcher and Contributor ID (ORCID) allows authors to contribute their information. The arXiv allows users to share scholarly content and maintains the metadata. This scholarly metadata could provide invaluable support to domain scientists across disciplines thus supporting research data management.
These repositories collect and maintain wide range of scholarly metadata features. However, to have a comprehensive scholarly metadata overview, it is essential to focus on these repositories together which could bring potential challenges that need to be addressed carefully while maintaining the data integrity. These challenges encompass aspects like Author and Affiliation Disambiguation. To optimize scholarly metadata coverage, strategic integration of open-access initiatives is crucial. This is where the innovative mechanism of SmartER steps in.
SmartER emerges as an innovative framework designed to address these challenges by synergizing metadata from diverse scholarly repositories, incorporating metadata features from sources such as ORCID, arXiv, ROR, Google Scholar etc. SmartER aims to create a unified and enriched metadata ecosystem that enhances the discoverability, interoperability, and reusability of research data thus facilitating seamless incorporation of available metadata into research metadata cycle. This initiative cultivates the repositories harmonization by structuring metadata features. It considers open-access repositories to address challenges related to Author and Affiliation Disambiguation. The SmartER data acquisition approach is proficient in extracting and accessing metadata from scientific publications and scholarly repositories. The repository harmonization enhances, validates, and associates extracted metadata with diverse repositories. Hence elevating the visibility and discoverability of research outputs across scientific communities. Moreover, it ensures a more comprehensive understanding of scholarly contributions. SmartER acts as a catalyst for improved scientific metadata in the research metadata cycle. It promises an era of enhanced collaboration, discoverability, and knowledge extraction within research communities, including interdisciplinary research endeavors that bridge diverse scientific domains.

Please specify ”other”

In addition, please add 3 to 5 keywords.

Research Metadata Cycle, SmartER, Repositories Harmonization

Please specify ”other”

For whom will your contribution be of most interest?

Researchers

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

Researchers

Primary author: Dr SURYANI, Muhammad Asif (GESIS - Leibniz-Institut für Sozialwissenschaften in Köln)

Co-author: Dr MATHIAK, Brigitte (GESIS - Leibniz-Institut für Sozialwissenschaften in Köln)

Presenter: Dr SURYANI, Muhammad Asif (GESIS - Leibniz-Institut für Sozialwissenschaften in Köln)

Session Classification: Poster Session A

Track Classification: Connecting research data: 4. Metadata annotation and management

Contribution ID: **128** Type: **not specified**

Welcome

Monday 4 November 2024 09:00 (15 minutes)

Presenter: LORENZ, Sören

Contribution ID: **129** Type: **KEYNOTE**

Meta data - An industry point of view

Monday 4 November 2024 09:15 (1 hour)

The needs from the industry and how to address them in a collaborative data ecosystem

Over the last years, the industry's needs have gone from simple delivery of materials and parts to a growing need for reliable and easily accessible meta-data. On the basis of concrete examples out of the Aerospace supply chain, the presentation will show some examples and will underline the importance of transparent and trustworthy data exchange and harmonized standards. Finally, a collaborative project will be presented which aims at building a data ecosystem for the Aerospace industry.

In addition, please add 3 to 5 keywords.

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

Please specify "other"

For whom will your contribution be of most interest?

Please specify "other"

Presenter: HOF, Tobias (Airbus)

Session Classification: Keynote

Contribution ID: **130** Type: **not specified**

Introduction Poster sessions

Monday 4 November 2024 13:45 (15 minutes)

In addition, please add 3 to 5 keywords.

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

Please specify ”other”

For whom will your contribution be of most interest?

Please specify ”other”

Contribution ID: **134** Type: **not specified**

Closing

Monday 4 November 2024 17:00 (5 minutes)

Contribution ID: **135** Type: **not specified**

Wake-up call (’Keynote’) from NFDI4Energy

In addition, please add 3 to 5 keywords.

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

Please specify ”other”

For whom will your contribution be of most interest?

Please specify ”other”

Presenter: HÜLK, Ludwig (Reiner Lemoine Institut (RLI))

Contribution ID: **136** Type: **KEYNOTE**

FAIRly Intelligent - What LLMs Bring to the Research Data Management Table

Tuesday 5 November 2024 14:15 (1 hour)

FAIR Research Data Management in interdisciplinary large-scale projects is very challenging. Data formats, acquisition processes, and infrastructure are highly heterogeneous. Furthermore, many tasks in FAIR RDM are tedious and complex for the researchers. In this keynote, we will discuss the potentials of generative AI to support FAIR RDM on examples from a large-scale project.

In addition, please add 3 to 5 keywords.

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

Please specify ”other”

For whom will your contribution be of most interest?

Please specify ”other”

Presenter: GEISLER, Sandra (RWTH Aachen University; Fraunhofer Institute for Applied Information Technology FIT)

Session Classification: Keynote

Contribution ID: **137** Type: **not specified**

Closing

Tuesday 5 November 2024 15:15 (45 minutes)

Highlights
Intro workshops
Thanks

Contribution ID: **142** Type: **INTERACTION**

1. Workshop - Advancing Interoperable Semantics: A Practical Workshop on DMPonline's Contributions and Collaborative Future

Wednesday 6 November 2024 09:00 (4 hours)

We will discuss the world of interoperable semantics at both domain-specific and application-wide levels, focusing on how DMPonline has pioneered enhancements and integrations that promote seamless data exchange and usage across diverse research contexts.
Join us in understanding how DMPonline's developments in interoperable semantics improve data management and use across various domains. We invite you to contribute your experiences and ideas to what promises to be a highly interactive and outcome-focused workshop that will drive the agenda on semantic interoperability and machine-actionable capabilities in data management. This session is crucial for anyone involved in or affected by metadata management at an operational or strategic level, offering practical insights and forward-looking discussions.

In addition, please add 3 to 5 keywords.

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

Please specify "other"

For whom will your contribution be of most interest?

Please specify "other"

Primary author: DRAFIOVA, Magdalena (University of Edinburgh)

Presenter: DRAFIOVA, Magdalena (University of Edinburgh)

Session Classification: Interactive session

Contribution ID: **143** Type: **INTERACTION**

2. Workshop - Comparative Analysis of Automated FAIR Assessment Tools' results: F-UJI, FAIR Enough, and FAIR Checker

Wednesday 6 November 2024 09:00 (4 hours)

After a brief introduction to the FAIR principles and the significance of automated assessments, participants will engage in hands-on sessions where they will compare the outputs of these tools on a curated list of datasets. The list represents datasets from various repositories that are typical within the biomedical context. Both a generalized overview of FAIR screening results at the repository level, and results for individual datasets will be prepared for the workshop. The workshop will introduce the participants to the FAIR principles, and how they can be translated into executable tests. Thus, it will showcase the different methodologies used by each tool, and how metadata is interpreted and scored, and, more generally, discuss the broader application of FAIR assessments for monitoring purposes.
This workshop provides participants with hands-on experience in evaluating automated tools designed for FAIR data assessment. It contributes to the conference by fostering a deeper understanding of how different tools measure up against the FAIR Principles and by developing an in-depth understanding of such automated tools and their limitations.

In addition, please add 3 to 5 keywords.

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

Please specify ”other”

For whom will your contribution be of most interest?

Please specify ”other”

Primary author: IARKAEVA, Anastasiia (BIH@Charité (QUEST Center))

Co-authors: STRECKER, Dorothea (Humboldt Universität zu Berlin); ROTHFRITZ, Laura (Humboldt Universität zu Berlin)

Presenter: IARKAEVA, Anastasiia (BIH@Charité (QUEST Center))

Session Classification: Interactive session

Contribution ID: **144** Type: **INTERACTION**

3. Workshop - The Two Motors of the DataHub Initiative for Environmental Sciences: a Powerful FAIR and Open Research Data Infrastructure together with Joint Semantic Metadata Schemas.

Wednesday 6 November 2024 09:00 (4 hours)

In environmental sciences, time-series data is crucial for monitoring environmental processes, validating earth system models and remote sensing products, training of data driven methods and better understanding of climate processes. However, even today, there is no uniform standard and interface for making such data consistently available according to the FAIR principles. Therefore, within the DataHub initiative, seven research centers from the Helmholtz research field Earth & Environment initiated the HMC project STAMPLATE. The aim of STAMPLATE is to adopt the SensorThings API (STA) from the Open Geospatial Consortium as the main framework and interface through which such data is made accessible.
Since project start in 2023, there have been numerous side activities and initiatives, which led to the establishment of a full digital ecosystem for time-series data, built around the STA. This ecosystem includes tools for the management of sensor metadata, quality-control of observational data, the consolidation and visualization via an overarching (meta)data portal and fully automated data pipelines connecting all these tools for a simple and user-friendly publication of data according to the FAIR principles.
The challenging task of the STAMPLATE committee was to harmonize the extremely heterogeneous metadata formats stemming from the different observation domains such as the earth, atmosphere and ocean. Moreover, within the domains different metadata formats developed historically due to diverging system architectures and missing guidelines.

Main content:
• Presentation of the architecture of our ecosystem
• Introduction to the STA as generic and modern interface for time-series data
• Presentation of the work on metadata homogenization
• Presentation and hands-on-tutorials of integrated tools and sub-systems

Presenters: HANISCH, Marc (GeoForschungsZentrum Potsdam GFZ); LORENZ, Christof (Karlsruhe Institute of Technology); LOUP, Ulrich

Session Classification: Interactive session

Contribution ID: **145** Type: **INTERACTION**

4. Workshop - EOC EO Products Service - Accessing and utilising geodata with STAC-API: Efficient access, intelligent search and seamless processing

Wednesday 6 November 2024 10:00 (1 hour)

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 dynamically and efficiently. As a data and service 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 contains 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 users so that they can visualize the data or analyze it further with the right tool.
In the following interaction we present tutorials for our STAC catalog. We show how to efficiently access the catalog and its available content. In addition, examples of various geoscientific analyses are shown to demonstrate the various possibilities of accessing and processing geodata using the STAC catalog.

Presenter: HAUG, Jan-Karl

Session Classification: Interactive session

Contribution ID: **146** Type: **INTERACTION**

5. Workshop - EOC EO Products Service - Accessing and utilising geodata with STAC-API: Efficient access, intelligent search and seamless processing

Wednesday 6 November 2024 14:00 (1 hour)

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 dynamically and efficiently. As a data and service 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 contains 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 users so that they can visualize the data or analyze it further with the right tool.
In the following interaction we present tutorials for our STAC catalog. We show how to efficiently access the catalog and its available content. In addition, examples of various geoscientific analyses are shown to demonstrate the various possibilities of accessing and processing geodata using the STAC catalog.

Presenter: HAUG, Jan-Karl

Session Classification: Interactive session

Contribution ID: **147** Type: **INTERACTION**

6. Workshop - From scientific terms to linked electronic lab notebooks –The workflow

Wednesday 6 November 2024 13:30 (4 hours)

In this interaction session, we will consolidate our talk on creating FAIR, rich and shared experimental (meta)data with a knowledge graph in mind. We will present the individual tools of the software workflow live and interactively, starting from vocabulary terms via ontologies to entering research (meta)data and sending it to another Electronic Lab Notebook (ELN).

A prerequisite for FAIR data publication is using FAIR vocabularies. Currently, tools for collaborative vocabulary composition are lacking. Our software tool, VocPopuli (developed in the HMC-funded MetaCook project), a Python-based web application, enables the collaborative definition and editing of metadata terms. Additionally, it annotates each term and the entire vocabulary using the PROV Data Model (PROV-DM), a schema for describing an object's provenance. Finally, it assigns a PID to each term and the vocabulary itself. The generated vocabularies are provided with a PID, contain data (defined terms) annotated with metadata (e.g., authors), and can be exported to SKOS and OWL. We present two exemplary ontologies (PolyMat and PolyLab, developed by Hereon and DLR Inst. f. Data Science) that can be expanded in VocPopuli.

Languages like OWL and SHACL within the RDF ecosystem support semantic enrichment and local conformance via validations, but can be tedious to apply. The ELN Herbie simplifies this process. It's a client-server web application that wraps an RDF triplestore, utilizing frameworks like RDFS, OWL, SHACL, and Schema.org. Herbie uses SHACL to create reusable web forms, turning a lab journal into a semantically rich knowledge base. It also supports RDF graph versioning and access management, allowing collaborative editing through a web interface or REST API.

ELNs are crucial for gathering analog metadata but face interoperability issues. With ELNdataBridge (from the HMC-funded ELN-DIY-Meta project), we have developed an API-based data exchange between the ELNs Herbie and Chemotion to enhance interoperability. The communication tool, a Python-based server application, uses API packages from both ELNs and an adapter for RESTful APIs. An administrator defines a synchronization key and maps data fields via a user-friendly web interface.

All presented software tools are open-source and can be applied to various research fields, as their development included experts from multiple scientific disciplines.

In addition, please add 3 to 5 keywords.

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

Please specify "other"

For whom will your contribution be of most interest?

Please specify ”other”

Presenters: KIRCHNER, Fabian (Helmholtz-Zentrum Hereon); HELD, Martin (Hereon)

Session Classification: Interactive session

Contribution ID: **148** Type: **KEYNOTE**

A Glimpse into the Future of Metadata - Practical Challenges in Development and Application of a Metadata Standard

Tuesday 5 November 2024 09:00 (30 minutes)

The complexity and diverse data requirements of energy system research demands a robust and adaptable metadata standard. The OEMetadata Standard, with its recent update to version 2.0, is designed to meet the needs of this transdisciplinary field. Illustrated through practical examples, the key features and enhancements of the standard are presented. Followed by the introduction of an innovative Open Peer Review Process for Metadata aimed at fostering community engagement and ensuring the ongoing improvement of metadata practices.

In addition, please add 3 to 5 keywords.

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

Please specify ”other”

For whom will your contribution be of most interest?

Please specify ”other”

Presenter: HÜLK, Ludwig (Reiner Lemoine Institut (RLI))

Session Classification: Wake up - Keynote