Contribution ID: **82** Type: **Short Talk**

Open Data for DESY and HIFIS

Thursday, 3 October 2024 10:00 (10 minutes)

DESY is one of the largest synchrotron facilities in Europe and as such is involved with a large amount of different scientific fields. Among these are High Energy and Astro particle Physics, Dark matter research, Physics with Photons and Structural Biology which generate huge amounts of data. This data is valuable and mostly handled in accordance with domain and community specific policies which take into account that embargo periods, ownership and license restrictions are respected. Nowadays there is a push towards opening the data up to the public as requested by funding agencies and scientific journals. In order to support this push, DESY IT is implementing and deploying solutions that support and enable the publishing of Open Data sets for the scientific community. These solutions will make the Open Data easily findable, browsable and reusable for further analyses by the long tail of science, especially when it's participants are not supported by large e-infrastructures.

With Open and FAIR data principles in mind, we will provide a metadata catalogue to make the data findable. The accessibility aspect is covered by making use of federated user accounts via eduGAIN, HelmholtzID, NFDI and later EOSC-AAI and will give community members access to the data with their institutional accounts. The interoperability of the data sets is ensured by establishing the use of commonly accepted data formats such as HDF5, specifically NeXuS and openPMD wherever possible. Providing the technical and scientific metadata will finally make the open data sets reusable for subsequent analyses and research. In order to address the spirit of sharing in Open Science, the blueprint for our Open Data solution will be shared with others through HIFIS first and upon successful evaluation also with the wider community.

Our protoype will initially consist of three connected solutions: the metadata catalogue SciCat, the storage system dCache and the VISA (Virtual Infrastructure for Scientific Analysis) portal. Scientific data is placed in a specific directory on dCache together with its metadata which is ingested into SciCat to be availabe for access and download options. Here, it is crucial to ensure that the scientific metadata stored in the catalog is harmonized among similar experiments. In order to achieve this, we are devising a method of creating experiment-specific metadata schemata against which metadata will be validated before ingestion. Simultaneously, a subset of the technical and scientific metadata will be integrated into the VISA portal such that scientists can access the dataset within it. VISA is a portal that allows creating virtual machines with preinstalled analysis tools, the selected data sets already mounted and accessible from a web browser forming a consistent environment allowing easy access to data and tools.

During the talk, we will present the architecture of the system, its individual components as well as their interplay. The focus will be the harmonization of the metadata schemata as well as the roadmap for the development of tooling and processes for ingestion and validation of the ingested metadata.

Topic

EOSC Developments and Open Science: Reproducible Open Science

Primary authors: FUHRMANN, Patrick (DESY); MILLAR, Paul (DESY); Dr WETZEL, Tim

Co-authors: Dr REPPIN, Johannes (Deutsches Elektronen-Synchrotron DESY); Dr PITHAN, Linus (Deutsches Elektronen-Synchrotron DESY); Mr VAN DER REEST, Peter (Deutsches Elektronen-Synchrotron DESY); Dr HINZMANN, Regina (Deutsches Elektronen-Synchrotron DESY); JANDT, Uwe

Presenter: Dr WETZEL, Tim

Session Classification: Reproducible Open Science: making research reliable, transparent and credible