

Open data at DESY

Tuesday, March 26, 2024 2:00 PM (20 minutes)

The DESY Research Infrastructure, historically supports a large variety of sciences, like High Energy and Astro particle Physics, Dark matter research, Physics with Photons and Structural Biology. Most of those domains generate large amounts of precious data, handled according to domain specific policies and taking into account embargo periods and license restrictions. However, a significant section of this data is supposed to become “Open Data”, often enforced by funding agencies. To support its scientific communities in producing and using open data, DESY-IT is developing and installing central services, making open data sets easily findable, browsable and viewable. In addition, mechanisms will be provided to analyse data for the long tail of science, not covered by large e-Infrastructures.

Following the principles of Open and FAIR data, we will provide a metadata catalogue to make the data findable. The accessibility aspect is approached by making use of federated user accounts via eduGAIN and will enable community members to use their institutional accounts for data access. The interoperability of the data sets is ensured by using community approved 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.

Our proposed setup will initially consist of three components: the metadata catalogue SciCat, the storage system dCache and the VISA (Virtual Infrastructure for Scientific Analysis) portal. Scientific data can then be placed in a specific directory on dCache together with its metadata and will be ingested into SciCat to be available for access and download. 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 has been developed for creating virtual machines that come with analysis tools pre-installed, the selected data already mounted and accessible from a web browser such that anyone can reliably access and explore data without having to install anything themselves.

During the talk at ISGC, we will present the architecture of the system, its individual components as well as their interplay. A public endpoint for the auditorium to test accessing data and the analysis tools by themselves will be given, too.

Primary authors: FUHRMANN, Patrick (DESY/dCache.org); Mr RHEE, Franz (Deutsches Elektronen-Synchrotron DESY); WETZEL, Tim (Deutsches Elektronen-Synchrotron DESY); MILLAR, Paul (DESY); Ms SERVAN, Sophie (Deutsches Elektronen-Synchrotron DESY)

Co-authors: VAN DER REEST, Peter (Deutsches Elektronen-Synchrotron DESY); Dr REPPIN, Johannes (Deutsches Elektronen-Synchrotron DESY); Mrs KWEE-HINZMANN, Regina (Deutsches Elektronen-Synchrotron DESY); Mr BARTH, Noel (Deutsches Elektronen-Synchrotron DESY)

Presenter: FUHRMANN, Patrick (DESY/dCache.org)

Session Classification: VRE

Track Classification: Track 5: Virtual Research Environment (including tools, services, workflows, portals, ... etc.)