Contribution ID: **237** Type: **Presentation**

Research Data Management for Huge Datasets: ”Cloudifying” Data not on the Cloud

Friday 21 March 2025 10:30 (15 minutes)

Data repositories play an essential role in Research Data Management according to the “FAIR principles” (Wilkinson et al. 2016) and leave less and less to be desired. However, they usually cannot accommodate huge datasets towards the PB range, as they e.g. come from supercomputing - for technical, financial or organisational reasons. In fact, for such datasets even movement to an external (paid) cloud storage is often not possible due to the size or costs.

At Munich, the University Libraries and the Leibniz Supercomputing Centre are collaboratively aiming at providing basic FAIR RDM also for such huge datasets. As these datasets are usually produced by user groups with technical excellence, the FAIR solution for them can arguably depend on IT skills. However it should not impose a strong data-lifecycle management or control,as these users often have their own ideas and project (or institutional) policies that shape the data management concept.

We therefore aim at a minimally-invasive approach, where users add metadata as YAML “sidecar files” (with DataCite-compliant contents) to their huge datasets and have those published manually via one of several portals.

At the LRZ, a python-scripted workflow has been prototyped to then push these metadata into the InvenioRDM repository framework, generating a DOI and a landing page. In this usage scenario, InvenioRDM is used purely as a metadata-publication frontend, while the data remain on back-end “Big Data” storage and are linked from the metadata.

We have implemented this concept with two InvenioRDM instances / demonstrators, one in collaboration with LMU University Library and Physics Department (”Open Data LMU - Physics”) and one for LRZ in general (”LRZ FAIR Data Portal”). These two use different approaches for making the actual data available.

Open Data LMU - Physics uses the iRODS ‘’data-grid middleware” and a web-based frontend to make the data openly available, while the LRZ FAIR Data Portal relies on GLOBUS and its data-transfer middleware. These approaches show how data-cloud-like functionality can be implemented on top of classical data storage, which is a central point in making huge datasets FAIR.

Authors: WELLMANN, Alexander (Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities); MUNKE, Johannes (Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities); Dr HACHINGER, Stephan (Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities)

Co-authors: Dr SPENGER, Martin (Ludwig-Maximilians-Universität München, University Library); Dr REDL, Robert (Ludwig-Maximilians-Universität München, Faculty of Physics)

Presenter: WELLMANN, Alexander (Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities)

Session Classification: FAIR Data Management

Track Classification: Main sessions: User Voice: Innovative Applications, Data Science Environments & Open Data