EGU22-11103
https://doi.org/10.5194/egusphere-egu22-11103
EGU General Assembly 2022

Data amounts and reproducibility: How FAIR Digital Objects can revolutionise Research Workflows

Ivonne Anders, Karsten Peters-von Gehlen, Hannes Thiemann, Martin Bergemann, Merret Buurman, Andrej Fast, Christopher Kadow, Marco Kulüke, and Fabian Wachsmann
German Climate Computing Center (DKRZ), Hamburg, Germany

Some disciplines, especially those that look at the Earth system, work with large to very large amounts of data. Storing this data, but also processing it, places completely new demands on scientific work itself.

Let's take the example of climate research and specifically climate modelling. In addition to long-term meteorological measurements in the recent past, results from climate models form the main basis for research and statements on past and possible future global, regional and local climate. Climate models are very complex numerical models that require high-performance computing. However, with the current and future increasing spatial and temporal resolution of the models, the demand for computing resources and storage space is also increasing. Previous working methods and processes no longer hold up and need to be rethought.

Taking the German Climate Computing Centre (DKRZ) as an example, we analysed the users, their goals and working methods. DKRZ provides the climate science community with resources such as high-performance computing (HPC), data storage and specialised services and hosts the World Data Center for Climate (WDCC). In analysing users, we distinguish between two groups: those who need the HPC system to run resource-intensive simulations and then analyse them, and those who reuse, build on and analyse existing data. Each group subdivides into subgroups. We have analysed the workflows for each identified user and found identical parts in an abstracted form and derived Canonical Workflow Modules.

In the process, we critically examined the possible use of so-called FAIR Digital Objects (FDOs) and checked to what extent the derived workflows and workflow modules are actually future-proof.

The vision is that the global integrated data space is formed by standardised, independent and persistent entities that contain all information about diverse data objects (data, documents, metadata, software, etc.) so that human and, above all, machine agents can find, access, interpret and reuse (FAIR) them in an efficient and cost-saving way. At the same time, these units become independent of technologies and heterogeneous organisation of data, and will contain a built-in mechanism that supports data sovereignty. This will make the handling of data sustainable and secure.

So, each step in a research workflow can be a FDO. In this case, the research is fully reproducible, but parts can also be exchanged and, e.g. experiments can be varied transparently. FDOs can easily be linked to others. The redundancy of data is minimised and thus also the susceptibility to errors is reduced. FDOs open up the possibility of combining data, software or whole parts of workflows in a new and simple but at all times comprehensible way. FDOs will make an important contribution to the reproducibility of research results, but they are also crucial for saving storage space. There are already data that are FDOs, but also self-contained frameworks that store data via tracking workflows. Similar to the TCP/IP standard, DO interface protocols are already developed. However, there are still some open points that are currently being worked on and defined with regard to FDOs in order to make them a globally functioning system.