Tools to support climate researchers for long-tail research data in a FAIR context

Christian Pagé[1], Abel Aoun[2], Alessandro Spinuso[3], Klaus Zimmermann[4], and Lars Bärring[5]

[1]CECI, Université de Toulouse, CNRS, Cerfacs, Toulouse, France (christian.page@cerfacs.fr)

[2]CECI, Université de Toulouse, CNRS, Cerfacs, Toulouse, France (aoun.abel@outlook.com)

[3]Royal Netherlands Meteorological Institute (KNMI), R&D Observations and Data Technology, De Bilt, Netherlands (alessandro.spinuso@knmi.nl)

[4]Swedish Meteorological and Hydrological Institute (SMHI), Norrköping, Sweden (klaus.zimmermann@smhi.se)

[5]Swedish Meteorological and Hydrological Institute (SMHI), Norrköping, Sweden (Lars.Barring@smhi.se)

Doing high quality research involves complex workflows and intermediate datasets. An important part is also sharing of those datasets, software tools and workflows among researchers, and tracking provenance and lineage. It also needs to be stored in a citable permanent repository in order to be referenced in papers and reused subsequently by other researchers. Supporting this research data life cycle properly is a very challenging objective for research infrastructures. This is especially true with rapidly evolving technologies, sustainable funding problems and human expertise.

In the climate research infrastructure, many efforts have been made to support end-users and long tail research. There is the basic data distribution, the **ESGF** data nodes, but this is to support mainly specialized researchers in climate science. This basic infrastructure implements quite strict standards to enable proper data sharing in the research community. This is far from **FAIR** compliance, but this has proven to be extremely beneficial for collaborative research. Of course, high level components and services can be built on top. This is not an easy task, and a layered approach is always better to hide the underlying complexity and also to prevent technology locking and too complex codes. One example is the **IS-ENES C4I 2.0 platform** (https://dev.climate4impact.eu/), a front-end that eases very much data access, and is acting like a bridge between the data nodes and computing services. The C4I platform provides a very much enhanced Jupyter-la like interface (**SWIRRL**), with many services to support sharing of data and common workflow for data staging and preprocessing, as well as the development of new analysis methods in a research context. Advanced tools that can calculate end-user products are also made available along with some example notebooks implementing popular workflows. One of these tools is **icclim** (https://github.com/cerfacs-globc/icclim), a python software package. C4I also includes high-level services such as on-the-fly inter-comparisons between climate simulations with **ESMValTool** (https://github.com/ESMValGroup/ESMValTool). All this work is also including large efforts to standardize and to become closer to FAIR for data, workflows and software.

Another way of helping researchers is to pre-compute end-users products like climate indices. This

is extremely useful for users because it can be really complex and time consuming to calculate those products. One example is to provide those users datasets of climate indices pre-computed on **CMIP6** simulations would be very valuable for those users. Of course all specific needs cannot be taken into account but the most general ones can be fulfilled. The **European Open Science Cloud (EOSC)** is providing computing and storage resources through the **EGI-ACE** project, enabling the possibility to compute several climate indices. In this EGI-ACE Use Case, icclim will be used to compute 49 standard climate indices on a large number of **CMIP6** simulations, starting with the most used ones. It could also be extended to **ERA5** reanalysis, **CORDEX** and **CMIP5** datasets.