FEDERATING SCIENTIFIC INFRASTRUCTURES AND SERVICES FOR CROSS-DOMAIN APPLICATIONS OF EARTH OBSERVATION AND CLIMATE DATA

Wolfgang Wagner[1,2], *Matthias Schramm*[1,3], *Bernd Logar*[3,4], *Gergely Sipos*[5], *Christian Briese*[2], *Tom Clark*[2], *Christoph Reimer*[2], *Gottfried Kirchengast*[6], *Mathias W. Rotach*[7], *Leopold Haimberger*[8], *Dirk Tiede*[9], *Harald Rieder*[10], *Gerhard Wotawa*[11], *Matthias Schwarz*[12], *Steffen Fritz*[13]

[1] Department of Geodesy and Geoinformation, Technische Universität Wien, Austria
[2] Earth Observation Data Centre for Water Resources Monitoring, Austria
[3] Information Technology Solutions, Technische Universität Wien, Austria
[4] ACONET Association, Austria
[5] Department for Services, Solutions, and Support, EGI Foundation, The Netherlands
[6] Wegener Center for Climate and Global Change & Institute of Physics, University of Graz, Austria
[7] Department of Atmospheric and Cryospheric Sciences, University of Innsbruck, Austria
[8] Department of Meteorology and Geophysics, University of Vienna, Austria
[9] Department of Geoinformatics, Paris Lodron University Salzburg, Austria
[10] Institute of Meteorology and Climatology, University of Natural Resources and Life Sciences, Austria
[11] GeoSphere Austria, Austria
[12] CCCA Data Centre, GeoSphere Austria, Austria
[13] Strategic Initiatives Program, International Institute for Applied Systems Analysis, Austria

ABSTRACT

In this contribution we describe a collaborative scientific infrastructure for the interdisciplinary use of Earth observation and climate data in research and teaching in Austria. The integration in European infrastructures such as the European Open Science Cloud or established by the Copernicus and Destination Earth programmes is a key requirement in the design of infrastructure and software components. Within Austria the focus will be on strengthening the integration of the Petabyte-scale storage and processing capabilities of the Earth Observation Data Centre, the Vienna Scientific Cluster, GeoSphere Austria with the computer systems and scientific services of the participating universities and research organisations. Access to data and their use will be simplified by a single sign-on procedure and new open source software solutions that enable multi-site workflows. This will enable scientists and students to work in an interdisciplinary manner with large amounts of data that are managed and stored at multiple geographic locations.

Index Terms— Infrastructure federation, scientific services, multi-site workflows, interdisciplinary collaboration, open science

Funding from the BMBWF (GEOCLIM and Cloud4GEO projects), FFG (FAIR2Earth) and Horizon Europe (C-Scale, EGI-ACE, interTwin) is gratefully acknowledged.

1. INTRODUCTION

Earth observation and climate science are two scientific disciplines that have seen massive increases in the volume, diversity and quality of data over the past decade. This has incentivised major infrastructure investments and the adoption of Big Data and cloud technologies for managing, accessing, processing, and visualisation of satellite measurements and model simulations [1]. These developments have been spearheaded by the US tech giants, most notably Google, Amazon and Microsoft. In Earth sciences, the most popular platform is the Google Earth Engine (GEE) that allows its users to access and analyse publicly available satellite and climate data for free [2]. It is particular powerful for users who are content to work with predefined algorithms and workflows. However, more complex and less common tasks may be difficult to implement. This is one of the reasons why Zhao et al. [3] conclude in their recent review of applications of GEE that significant improvements in the supporting technology are still needed. They even go as far as to write that "At present, the development of GEE is still in its infancy." [3].

This is a bold statement by itself, but even more so when considering how other initiatives to establish monolithic Big Data platforms have struggled to attract users. Prominent examples are the four Copernicus Data and Information Access Services (DIAS) platforms that were selected by the European Space Agency (ESA) and the European Commission

doi:10.2760/46796

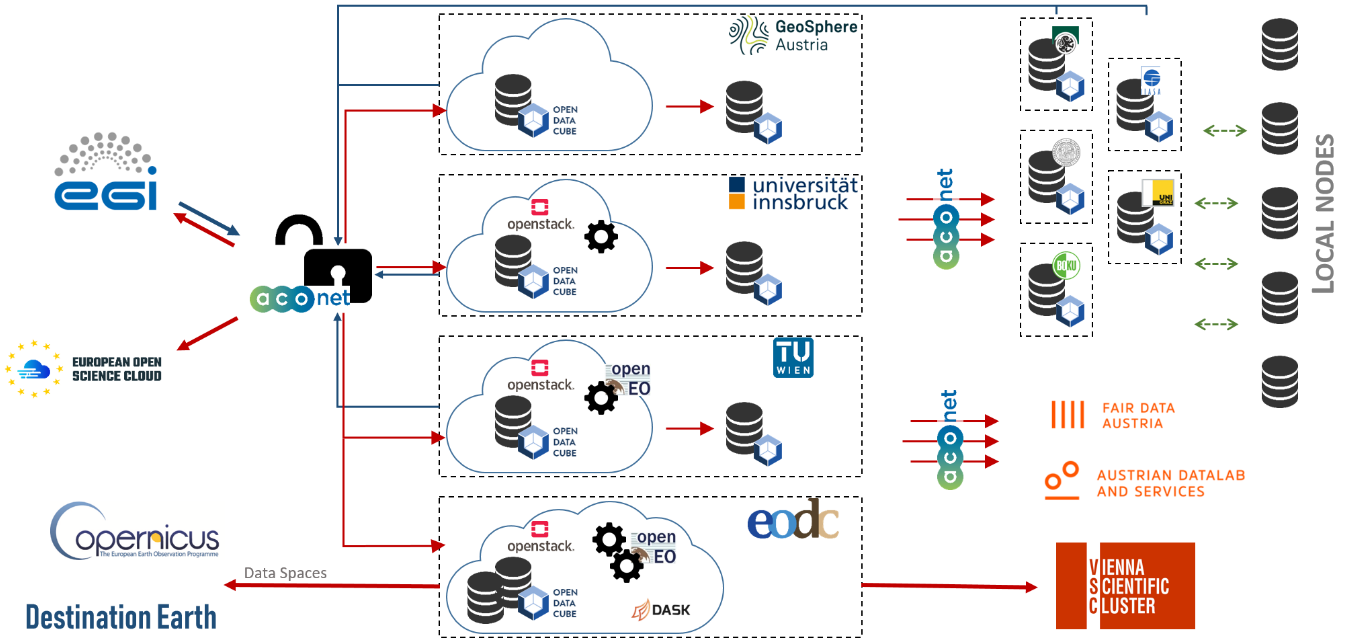

Fig. 1. Schematic view of the infrastructure federation of Austrian core services (ACOnet, VSC, EODC, GeoSphere Austria, Austrian DataLAB, FAIR Data Austria) with European services (EGI, EOSC, Copernicus, Destination Earth) and partner organisations (universities and research organisations) for improving cross-domain applications of Earth observation and climate data. Blue arrows indicate access to the ACOnet identity federation, red ones access to services for authenticated users.

(EC) to develop commercially self-sustainable cloud services that allow users to access, manage, process and download Copernicus data. As none of the commercial DIAS consortia has succeeded in developing a market [4], ESA and the EC re-scoped the programme and replaced the DIAS platforms by only one new service called the Copernicus Data Space Ecosystem (CDSE). This illustrates that developments in this field are still very dynamic and that it needs more than a large data collection and cloud computing resources to convince users to migrate their workflows – or parts thereof – to external platforms.

In parallel to the efforts to scale up data centre capacities there have been many initiatives to federate existing scientific infrastructures and services. An example from the climate domain is the Earth System Grid Federation which is an international collaboration to empower climate research such as the Coupled Model Intercomparison Project. This is realised through a decentralised database and a common software architecture that is deployed on a system of geographically distributed peer nodes [5]. In Europe, the European Open Science Cloud (EOSC) as become the main framework to promote the federation of national and European infrastructures as the basis for creating a web of 'Findable, Accessible, Interoperable, and Re-usable' (FAIR) data [6, 7]. As recounted by Jean-Claude Burgelmann [8] this was not necessarily the intention in the early days when the EOSC was invented to foster Open Science in Europe. However, eventually, federation was seen as the only practical way to realise the EOSC.

Along these lines, in our efforts to strengthen Austria's infrastructure capabilities to serve Earth observation and climate science, a key guiding principle is the concept of federation, seeking an optimal balance between supranational, national and local infrastructure components. Furthermore, the design of the infrastructure is guided by Open Science principles, i.e. the quest to make research data, methods, and findings openly available, enabling broader participation, scrutiny, and the potential for accelerated progress in scientific discovery. In this article, we provide an overview of existing national capacities and present approaches for federation, drawing on concepts developed by the EGI community and more domain specific efforts such as the development of the openEO API [9].

2. AUSTRIAN INFRASTRUCTURES AND SERVICES

Many of the infrastructure services that are crucial for the Austrian science community have their roots in the activities of a smaller group of scientists and technicians from one or few universities [10], and were later scaled up through cooperation agreements of partner organisations requiring these services for their daily work.

Already in 1986 the "Association for the Promotion of an Austrian National Research and Education Network" (ACONET Association) was founded, which paved the way to

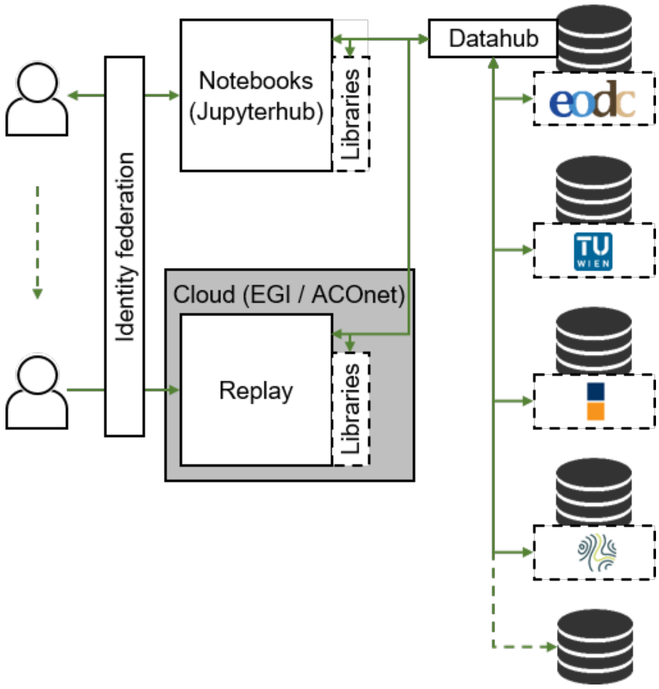

Fig. 2. Services for international users to jointly access and process decentralised data in the federated scientific e-Infrastructure.

develop the 'ACOnet' NREN (national research and education network) into the failsafe high-performance network with excellent connectivity to other scientific networks and the global Internet it is today. The Vienna Scientific Cluster (VSC) started as a project of three universities in 2009 and is now a collaboration of several universities and research organisations that provides supercomputer resources and corresponding services to their users. The Earth Observation Data Centre (EODC) was founded in 2014 as a public-private partnership to enable work with Petabyte-scale satellite data sets. As described in [11] the idea of the EODC has always been to establish a collaborative/federated cloud connecting several data centres throughout Europe rather than pursuing a centralised approach for the provision of the IT infrastructure. As such, the EODC data centre capabilities have been designed to integrate seamlessly with existing Austrian infrastructures, most importantly those of the VSC and the former Austrian met service ZAMG. In 2023 ZAMG was merged with the Federal Geological Institute to form the new Federal Institute for Geology, Geophysics, Climatology and Meteorology (GeoSphere Austria). Among its many responsibilities, GeoSphere Austria has been tasked to establish and operate "a central data infrastructure as a service for science, business, public administration and society with automated access".

Due to the powerful integration of the VSC, EODC and GeoSphere Austria in ACOnet, users have already a wide range of options for using and exchanging Earth observation and climate data, from downloading the data, processing the data on large cloud stacks (Austrian DataLAB and Services) or uploading them on one of the FAIR data repositories (FAIR Data Austria). Within a new infrastructure project called Cloud4GEO, our plan is to strengthen the integration of all described core services with European infrastructures (EGI, EOSC, Copernicus Data Space Ecosystem, Destination Earth, etc.) and the computer systems and scientific services of the participating universities and research organisations (see Figure 1). Access to data and their use will be simplified by a single sign-on procedure and new open source software solutions that enable multi-site workflows as described in the next section.

3. TECHNICAL SOLUTIONS TO FEDERATION

In 2018, under the Austrian Presidency, the European Commission launched the EOSC in Vienna [7, 8]. This symbolic act highlighted the importance of this European programme for Austria, which – as a small country – depends heavily on international collaboration for performing world-class science. Accordingly, our partner network has already been engaged in developing interfaces to EOSC, EGI, EUMETSAT, ESA and DIAS [12], finding practical solutions for Big Data challenges such as the development of global satellite data cubes systems [13], platform-agnostic access to these data cubes [9], and establishing operational multi-site workflows and services [14].

To federate the infrastructures shown in Figure 1 we will make use of existing services from the ACONET Association and from the EGI Foundation. All infrastructures are already connected via encapsulated ACOnet links. A current *identity federation* from ACONET, based on eduGAIN is already connected to the EGI Authentication and Authorization Infrastructure (AAI), called Check-in. This framework will grant access for selected national and international users. Furthermore, a *DataHub service*, provided by EGI and based on the Onedata technology will be implemented at one of the participating cloud platforms (see Figure 2). This service allows catching and moving data from the decentralised storage systems and making them available to distinct user communities. Users (primarily data scientists) will be able to interact with these data files from a central graphical-scripting environment, thus be able to access data as if it was in a single virtual platform - with the various decentral data structures represented as distinct folders. The graphical-scripting environment will be based on the EGI's *Notebooks service* and facilitates development on a JupyterHub, while accessing the collected data provided by the DataHub service. It also mounts locally provided programming libraries. Users will be able to choose programming languages in the JupyterHub environment, write scripts that use / process / visualise the data coming from the DataHub system, and share their note-

doi:10.2760/46796

books together with its references to data e.g. via GitHub to enable reuse by third-party users. These third-party users can reproduce such scripted simulations in the EGI *Replay environment*, which can be setup on any server connected e.g. to the ACOnet NREN using EGI technology. The overall setup enables reproducible open science with the integration of distributed data from Austrian and other repositories.

With the use of the underlying ACONET and EGI identity federations, additional cloud or high performance computing (HPC) platforms within the EGI federation or connected by the ACOnet can be embedded into this processing network in the future, provided that they have access to the needed data and processing libraries, stored in the here presented infrastructures. This will enable the data scientists to run their simulations from the JupyterHub and Replay environments on external cloud or HPC systems.

4. OUTLOOK

From a historical perspective, the Big Data challenge faced by Earth observation and climate science is a fairly recent phenomenon. While some politicians and other decision makers have sought to counterweight the dominance of the US tech giants by promoting monolithic cloud solutions, science has moved on and developed more agile solutions that allow users to work on multiple platforms. Thanks to powerful Open Source software and emerging standards that can regulate the exchange of data and distribute the processing [15, 16, 17], the 'web of FAIR data and services' (as encapsulated by the EOSC) is starting to become a reality. In this 'data web' the role of intermediaries that redistribute the data on their service platforms will be diminished, giving back the recognition to those organisations that produce and maintain the data (the 'data stewards').

REFERENCES

[1] Wagner, W. "Big data infrastructures for processing Sentinel data." Photogrammetric Week 2015, Wichmann, 15, 93-104, 2015, https://phowo.ifp.uni-stuttgart.de/publications/phowo15/110Wagner.pdf

[2] Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R. "Google Earth Engine: Planetary-scale geospatial analysis for everyone." Remote Sensing of Environment, 202, 18-27, 2017, doi: 10.1016/j.rse.2017.06.031

[3] Zhao, Q., Yu, L., Li, X., Peng, D., Zhang, Y., Gong, P. "Progress and trends in the application of Google Earth and Google Earth Engine." Remote Sensing 13(18), 3778, 2021, doi: 10.3390/rs13183778

[4] Timmermans, R. "Comparing Copernicus DIAS Services", 2022, https://www.groundstation.space/comparing-copernicus-dias-services, accessed on 5 July 2023

[5] Cinquini, L., et al. "The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data." Future Generation Computer Systems, 36, 400-417, 2014, doi: 10.1016/j.future.2013.07.002

[6] Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L.O.B., Wilkinson, M.D. "Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud" Information Services & Use, 37, 49-56, 2017, doi 10.3233/ISU-170824

[7] Budroni, P., Burgelman, J.-C., Schouppe, M. "Architectures of knowledge: the European open science cloud" ABI Technik 39, 130-141, 2019, doi: 10.1515/abitech-2019-2006

[8] Burgelman, J.-C. "Politics and open science: How the European Open Science Cloud became reality (the untold story)" Data Intelligence, 3, 5-19, 2021, doi: 10.1162/dint_a_00069

[9] Schramm, M., et al. "The openEO API–Harmonising the Use of Earth Observation Cloud Services Using Virtual Data Cube Functionalities." Remote Sensing, 13, 1125, 2021, doi: 10.3390/rs13061125

[10] Wöber, W. "40 Jahre Netz-werken im Rückspiegel", ACOnet Annual Report 2016, 18-29, https://www.aco.net/jahresberichte, access 11 July 2021.

[11] Wagner, W., et al. "Addressing grand challenges in earth observation science: The earth observation data centre for water resources monitoring" ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 2, 81-88, 2014, doi: 10.5194/isprsannals-II-7-81-2014

[12] Backeberg, B., et al. "An open compute and data federation as an alternative to monolithic infrastructures for big Earth data analytics" Big Earth Data, 2022, doi: 10.1080/20964471.2022.2094953

[13] Wagner, W., et al. "A Sentinel-1 backscatter datacube for global land monitoring applications" Remote Sensing, 13, 4622, 2021, doi: 10.3390/rs13224622

[14] Salamon, P., et al. "The new, systematic global flood monitoring product of the copernicus emergency management service." IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 1053-1056, 2021., doi: 10.1109/IGARSS47720.2021.9554214

[15] Ma, X. "Linked Geoscience Data in practice: Where W3C standards meet domain knowledge, data visualization and OGC standards" Earth Science Informatics, 10, 429-441, 2017, doi: 10.1007/s12145-017-0304-8

[16] Kotsev, A., Minghini, M., Tomas, R., Cetl, V., Lutz, M. "From spatial data infrastructures to data spaces—A technological perspective on the evolution of European SDIs" ISPRS International Journal of Geo-Information, 9, 176, 2020, doi: 10.3390/ijgi9030176

[17] Sudmanns, M., et al. "Think global, cube local: an Earth Observation Data Cube's contribution to the Digital Earth vision." Big Earth Data, (2022): 1-29., 10.1080/20964471.2022.2099236