EarthCODE – ESA's Earth Science Collaborative Open Development Environment

Anca Anghelea [1], Garin Smith [2], Stephan Meissl [3], Stefan Achtsnit[3]

[1] Department of Climate, Sustainability and Science, European Space Agency, 00044 Frascati, Italy – anca.anghelea@esa.int
[2] Telespazio UK, 350 Capability Grn, Luton LU1 3LU, UK, – garin.smith@telespazio.com
[3] EOX IT Services GMbH, Thurngasse 8, 1090 Vienna, Austria – stephan.meissl@eox.at

Keywords: FAIR, Open Science, Earth Observation, Earth System Science, Cloud Computing, Reproducibility.

Abstract

Earth Observations (EO) have become crucial for advancing climate and Earth System research, enabling significant scientific discoveries. Advanced EO missions, such as ESA's Earth Explorers and the EU's Copernicus Programme, provide vast data volumes and continuous and global observations with cutting-edge technology, a vital resource for understanding processes and interactions within Earth's sub-systems, offering critical evidence of climate change impacts on society and ecosystems. Open access to such data products has been instrumental in facilitating global scientific collaboration and innovation. Reproducibility in Earth Science is essential for validating discoveries, enabling the scientific community to trust and build upon each other's work. Open data, coupled with open-source software, is key to achieving this reproducibility, ensuring transparency, and facilitating peer review. Addressing scientific challenges requires collaborative efforts supported by fit-for-purpose technology to enable new insights and ensuring they are FAIR (findable, accessible, interoperable, and reusable), trustable and up to date, supporting long-term climate studies and action-driven science.
Several initiatives within ESA's Earth Observation Programme are fostering FAIR and Open Science. EarthCODE is one such initiative, focusing on implementing ESA's vision of EO Open Science and Innovation by adopting FAIR and Open principles in Earth Science activities funded by ESA's FutureEO Programme. EarthCODE leverages existing European platform solutions for effective cloud-based analyses of EO data and higher-level geospatial products. It aims to ensure the persistence of end-to-end scientific workflows and provide means for managing open research data, including data, code, and documentation. EarthCODE aims to make Earth Science research more reproducible, transparent, and collaborative, driving progress in understanding and addressing global environmental challenges.

1. Introduction

Earth Observations play an essential role in supporting the development of novel methods for climate and Earth System scientific research, enabling new discoveries. Novel EO scientific missions such as ESA's Earth Explorers, complemented by operational missions – for example from the Copernicus Programme of the EU – are providing exponentially increasing data volumes and synoptic observations of our planet using state of the art instrumentation. In the context of the climate crisis, the continuous provision of long-term global data records is essential to convey the spatio-temporal scale of processes, interactions between Earth's sub-systems and fundamental pieces of evidence of the impacts of climate change on society and ecosystems. This has led to an improved understanding of the Earth by the scientific community and has opened the door to an increased adoption of novel products in science-based actionable information systems supporting policy and decision makers.

Notably, the availability of Earth Observation data with wide open access promoted by the EU, ESA, and other space agencies, is a key enabler of discovery by ensuring that researchers, scientists, and organizations worldwide can utilize data without restrictions. It amplifies collaboration among scientific teams and innovation, driving progress in addressing the global challenges of this century. However, the road to science-driven action for the Earth is long, and there are still numerous gaps in our capacity to measure, understand, predict, decide and take action for our planet, such as: quantifying the human pressures on the environment and its global impacts, describing and predicting the changes in the global cycles of water, carbon and energy, or defining reliable metrics for effective adaptation and mitigation policies, that account for the effect of feedbacks and interactions between society and ecosystems.

The scientific challenges require a concerted community effort, sustained collaboration across distributed scientific research groups addressing complementary problems with diverse methods and at various scales, effective exchanges of data and knowledge, iterations and numerous trials integrating approaches based on EO data with models and other data sources, using fit for purpose tools. Additionally, the measure-understand-predict-decide-act cycle is inherently incremental and iterative and requires that relevant results are made FAIR (findable, accessible, interoperable, and reusable) and persisted in the long-term. This is important from many perspectives including to increase trust in the discoveries, enable peer-review and scientific scrutiny, as well as to enable frequent updates and upgrades of methods and datasets required for long-term climate studies, among others. How can we implement this sustainably, what are the technologies required (available today, and in the future), are there changes required in the scientific practice, and what new processes and partnerships need to be put in place?

Acknowledging these needs to support the implementation of the science-driven approach that lies at the heart of the ESA's Earth Observation Programme, several initiatives have been initiated in recent years – along the various parts of the EO value chain, from innovative missions to excellent science to societal benefits and applications – that created a solid foundation for a systematic approach to FAIR and Open Science.

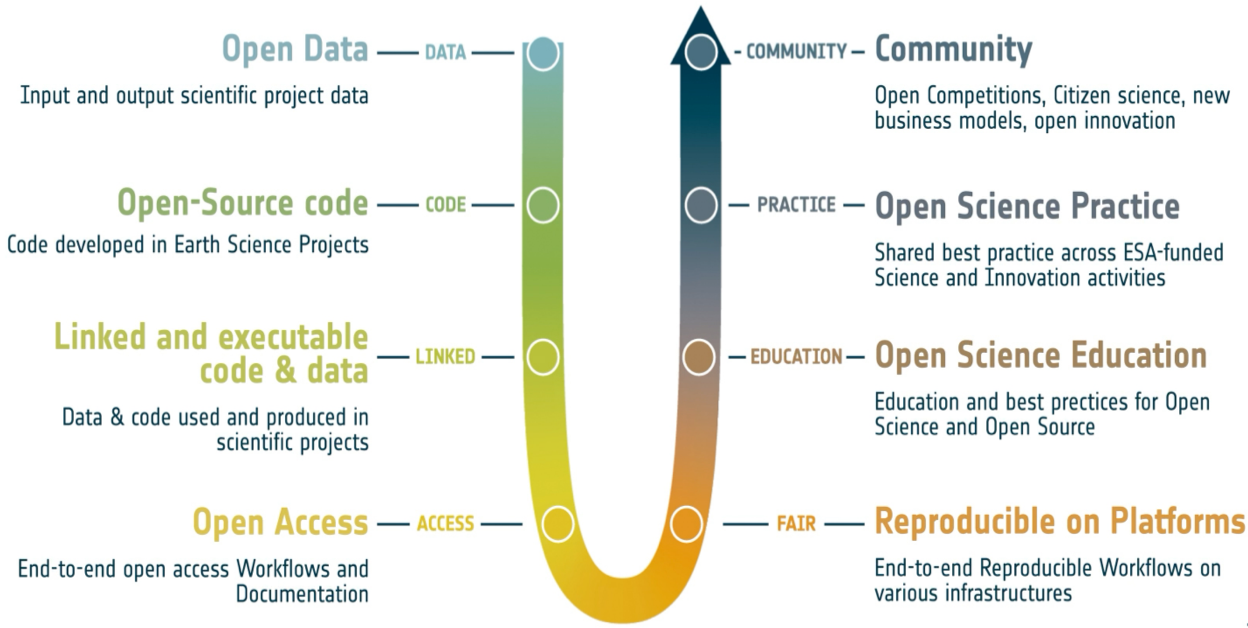

Figure 1. EO Open Science and Innovation elements, ESA

EarthCODE is part of this larger panorama of strategic initiatives. Addressed to the community of researchers engaged in ESA-funded activities for fundamental scientific research on the Earth System, EarthCODE is a technology endeavour as much as a community one. It and attempts to implement ESA’s vision on EO Open Science and Innovation (Figure 1) at the foundational level of the scientific exploitation of EO data.

In a nutshell, EarthCODE aims to facilitate adoption of FAIR and Open principles at the core of the Earth Science activities funded by ESA under the block 4 (Science for Society element) of the FutureEO Programme, by providing the means, the tools, and the ability to do so.

In this paper, we present the concept of EarthCODE – an Earth Science Collaborative Open Development Environment. EarthCODE leverages existing European platform solutions to enable effective analyses of EO data and higher-level geospatial products in the cloud. It ensures the persistence of end-to-end scientific workflows and data, and provides means for the management of open research data, including data, code, and documentation. By integrating these tools, EarthCODE aims to make Earth Science research more reproducible, transparent, and collaborative, ultimately driving progress in understanding and addressing global environmental challenges. This paper introduces EarthCODE – an Earth Science Collaborative Open Development Environment, emphasizing the critical role of open data and open-source tools in enhancing reproducibility and scientific integrity in Earth Science research.

2. Implementing Open Science

Open science principles are increasingly being embraced in Earth sciences to promote transparency, collaboration, and accessibility of research. Commonly, these principles are being applied by promoting open access publications, preprints and an open review process, sharing data and methodologies openly for verification, reproducibility and reuse, embracing open-source principles in software development to allow others to inspect, modify, and contribute to the code, encouraging collaboration among researchers through various platforms like GitHub, GitLab and other collaborative tools, sharing educational resources openly to allow for a global audience, and by promoting citizen science.

Open Science is also creating global impact, as shown by the increasing number of resources on open science, and by the dedicated programmes and initiatives to promote open science adoption in the community (AGU, 2024), (Murphy, 2021), (P. J. Zellner et al., 2024). Policies and recommendations from international bodies (EU, 2020), (European Commission, 2021), (UNESCO, 2022), further aim to make the scientific process more transparent, accessible, and inclusive.

This global trend comes in the age of the cloud revolution, making advanced and resource hungry processing increasingly accessible, as researchers can discover, access and process huge amounts of Earth data from EO platforms, couple their analyses with models, and run complex workflows on powerful infrastructures that can scale and are accessed on demand, from the convenience of their desk.

Together, Open Science and EO platforms create huge opportunities for Earth System science. Still, there is significant complexity to consider. One good example of where common principles promoted by open science are insufficient for a cloud computing scenario is making available the dataset package for a publication, in which researchers should aim to make their data accessible for download in “1-click”. Consider the case of high-resolution global datasets produced by workflows executed on platforms that accesses cloud-optimized Earth Observation and other Earth data. Not only it is cost and resource ineffective to deliver this dataset for download, but it also hinders reproducibility and use, as its sole delivery, even with the accompanying code, is insufficient without the access to the infrastructure. Sustainable Open Science must account for the new complexities and requirements of the cloud-computing research era.

In this context, EarthCODE has the ambition to deliver a model for a Collaborative Open Development Environment for Earth system science, where researchers can leverage the power of the wide range of EO platform services available on the global EO market to conduct their science, while also making use of FAIR Open Science tools to manage their data, code and documentation, create workflows that are end-to-end reproducible on various infrastructures, and have the opportunity to discover, use, reuse, modify and build upon the research of others in a fair and safe way.

EarthCODE aims to make possible the eight enabling elements of the EO Open Science and Innovation vision (Figure 1):

1. **Open data** – understood as all the necessary data to reproduce an experiment.
2. **Open-source code** – understood as an implementation of scientific methods
3. **Linked and executable code and data**
4. **Open access documentation** – all the documentation associated to the data and the code
5. **End-to-end workflows reproducible** on platforms – understood as full scientific experiments
6. **Open Science Education** – openly available educational resources, guidelines and tutorials for Open Science principles in the Earth Sciences
7. **Open Science Practice** – accessible resources and tools to put in practice the Open Science principles
8. **Community** – understood as a healthy and vibrant community of practice that spans from citizen science to new business models and open innovation.

3. EarthCODE Concept

The main goal of EarthCODE is to develop an advanced open science environment, to help project teams, as well as individual researchers in Earth Science and EO to efficiently conduct science in a FAIR and open way. In the scope of EarthCODE, we employ the term "research data" as an umbrella term, to encompass data in its strict sense, as well as to metadata, software, methods, algorithms, and documentation related to that data, inspired by the Beijing Declaration on Research Data (CODATA, 2019).

Through EarthCODE, scientists would be able to:

- Discover, access and use with Research Data;
- Access powerful cloud computing resources provided by existing commercial and institutional platforms;
- Develop scientific workflows;
- Store Research Data persistently on the ESA Cloud;
- Publish FAIR Research Data (+DOI);
- Disseminate through dashboards, advanced visualisations and storytelling;
- Interact and collaborate with Research Data owners and the wider community.

A key stakeholder group are the ESA Science Clusters which aim at promoting networking, collaborative research, and fostering international collaboration in various Earth science domains, including: atmosphere, ocean, carbon, water cycle, polar, extremes and natural disasters, biosphere, land and agriculture, solid and magnetic earth. Clusters involve different ESA funded projects and activities bringing together expertise, data and resources in a synergistic manner.

Figure 2. The EarthCODE key visual. Portal (beta) available at: earthcode.esa.int

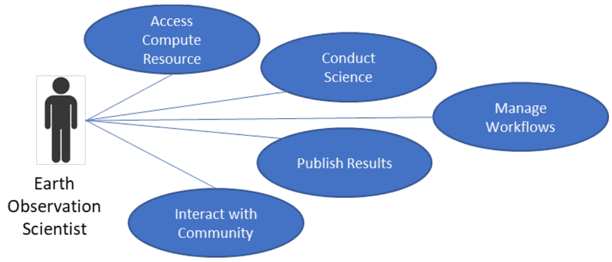

Figure 3. The EarthCODE key user persona is the EO scientist and earth science researcher. There are two types of roles envisaged: the contributing scientist and the user scientist, with the key difference being the access to computational resources as well as the right to publish content in EarthCODE which are reserved to the contributing scientists (typically participating in an ESA-funded activity or an associated one).

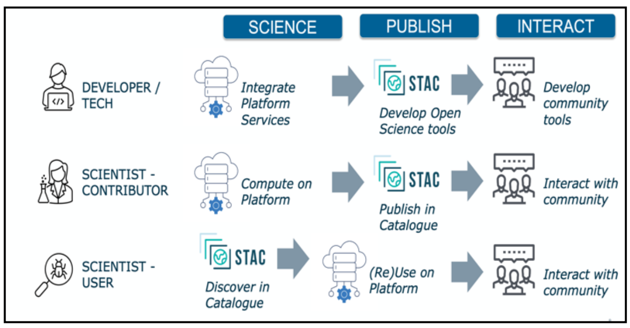

Figure 4: The above figure shows some example user stories that we envisage being supported by EarthCODE, to support the roles of developer/tech, scientist as a contributor and scientist as a user.

Apart from the access to cloud computing capabilities and comprehensive data, EarthCODE will provide workflow and data management tools, as well as resources and information that guide researchers how to organize their data and code to make all research implemented in the Science Hub and Clusters be sustainable, FAIR and reproducible, enabling efficiency and transparency. The high-level conceptual architecture is depicted in Figure 5.

The key user persona (Figure 3) for EarthCODE is the earth observation scientist, for whom EarthCODE looks to provide capabilities such as: development and execution of workflows to capture large scale processes over climatic temporal scales; conducting complex multi-variate spatio-temporal analyses; running various analyses and models, including ML; setting up automated execution of algorithms, potentially recurrent; performing dataset and feature engineering; performing pipeline building and automation; accessing and processing heterogeneous data sources including online collections, data cubes and data at native resolution; having access to complete metadata for the datasets including data lineage and provenance, source and (pre-)processing; extracting, reading, writing and downloading data from APIs; storing, describing, publishing and documenting new data; using own data such as field measurements within scientific workflows; accessing the newest (EO) data available, programmatically; packaging, publishing and preserving the scientific outputs (data and code) with persistent identifiers (e.g., DOIs) in long-term storage repositories; setting up and using software environments; importing and running environments on EO Platforms; exporting environments (e.g. Docker, application package, openEO graph, etc) and sharing them; using Git, including provided by own organisation; managing versions of data, code and documentation; making projects/repositories citable; contributing to open source projects; publishing scientific results and other research outcomes on web, including scientific visualisations; discovering and exploring scientific results, data, code and documentation of other contributing researchers.

3.1 Open-Source Development

EarthCODE will rely on the services provided by already mature EO Platforms and cloud computing infrastructures currently on offer on the global EO market. It will therefore look to integrate services provided by such platforms and infrastructures typically leveraging ESA’s Network of Resources mechanism, while supporting the development of necessary additional interoperability through collaborations and coordination with initiatives such as the “Open Science Persistent Demonstrator” (OGC OSPD, 2023) and the “Interoperable Building Block Evolution Framework” (EOEPCA+) (EOEPCA 2021, EOEPCA 2023a, EOEPCA 2023b)

The Open Science Persistent Demonstrator (OSPD) is a long-term inter-agency initiative aiming to enable and communicate reproducible Earth Science across global communities of users and amplify inter-agency Earth Observation mission data, tools, and infrastructures. A major goal for this activity is to test and demonstrate the current level of interoperability among still maturing Earth Observation and Earth Science cloud technologies and infrastructures developed by ESA, NASA and other space agencies and international organizations and contributed in kind to the initiative in a coordinated approach, building on existing investments where appropriate. Through various Pilots, the activity will demonstrate interoperable, collaborative research that allows reuse of existing components. These other resources are either offered as part of emerging Open Science Environments or in the form of either directly accessible “cloud-native” data/functions or by means of Web APIs.

EarthCODE is one of the key stakeholders of the OSPD and contributes with requirements and practical science cases to the pilots, looking to potentially integrate outcomes of the OSPD into its architecture.

The Interoperable Building Block Evolution Framework ensures continuity of the Common Architecture (EOEPCA) - which has defined a reference architecture for cloud EO platforms and has delivered a number of Building Blocks (BB) at different level of maturity and supporting community. They represent the state-of-the-art in terms of OGC standards. EOEPCA+ also ensures international engagement in interoperability, setting the framework for generic building block development. The Common Architecture Building Blocks provide interoperable open-source elements for the EarthCODE Architecture.

One such component developed on top of EOEPCA open-source building blocks and which will be integrated in EarthCODE is the Open Science Catalogue (OSC).

The Open Science Data Catalogue (ESA OSC, 2024) is an ESA Open Science activity aiming to enhance the discoverability and use of the various scientific and value-added results (i.e. data, code, documentation) achieved in Earth System Science research activities funded by ESA EO. The OSC provides open access for the scientific community to geoscience products (based on EO data from ESA and non-ESA missions and other geospatial information and models) across the whole spectrum of Earth Science domains. The OSC adheres to FAIR principles and promotes reproducibility of scientific studies. The OSC makes use of various Open-Source geospatial technologies such as pycsw, PySTAC, and OpenLayers and tries to contribute back to these projects in terms of software and standardisation (Schindler, 2023).

EarthCODE will have three main platform types based on three Work Streams (WS): these include platforms for FAIR Open Science (WS2), Infrastructure (WS1), and Community (WS3). Driving the needs of the platforms are the science stakeholders, who offer continuous feedback to all the platforms. There is a workflow development lifecycle, where a user can develop code in their preferred language, and execute it on the cloud. Workflows can be verified and published into GitHub and the Open Science Catalog to allow re-use. This will be supported by variety of WS1 and WS2 platforms, as detailed in Figure 5.

At the time of writing, the list of solutions on the right hand side in Figure 5 is only illustrative, as the process of platform procurement for the first year of implementation is still ongoing. Selection of platforms is done in open competition via a Best Practice procurement. This process is intended to have a yearly frequency, to enable incremental development of capabilities and to promote the principle of open innovation, allowing for new participants to integrate with EarthCODE as it develops into a mature solution.

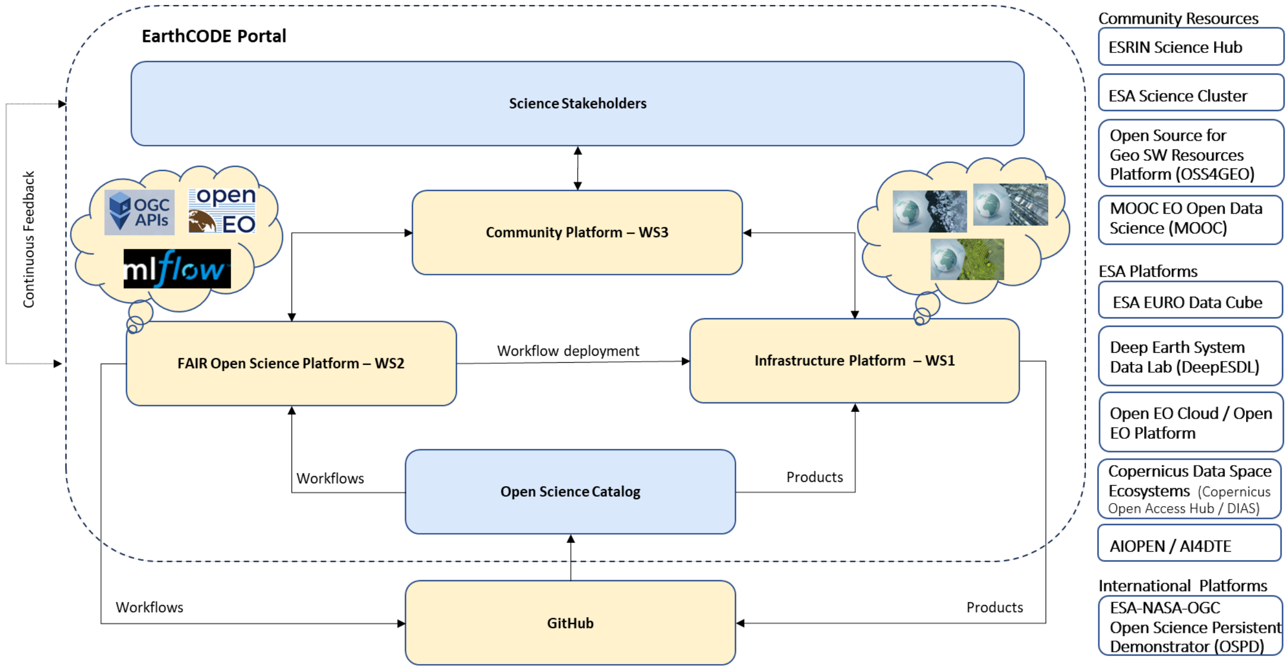

Figure 5. A high-level view of the functionality of the EarthCODE portal, detailing the interaction with the different work streams. The list of solutions on the right hand side is only illustrative and is intended to provide examples as to the types of capabilities and services that EarthCODE is looking to integrate.

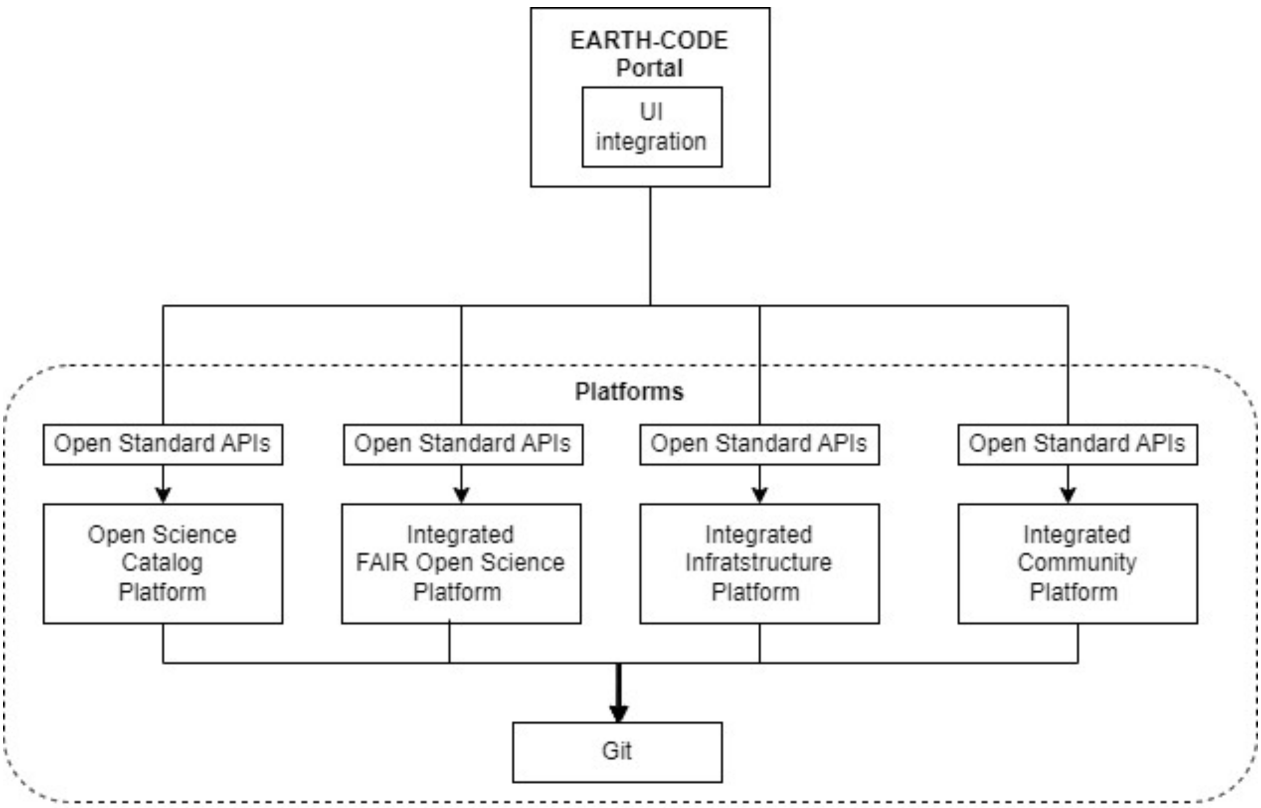

Figure 6: Platform approach to Work Stream Implementation

3.2 Platform Functionalities

EarthCODE will facilitate the use of data, tools and services from a wide variety of platforms to manage workflows. Open standards will facilitate and promote multiple platform interoperability. STAC and OGC API - Records will be used by the Science Catalogue to help describe resources. Platform integration will also be supported using DevOps best practise (Figure 6).

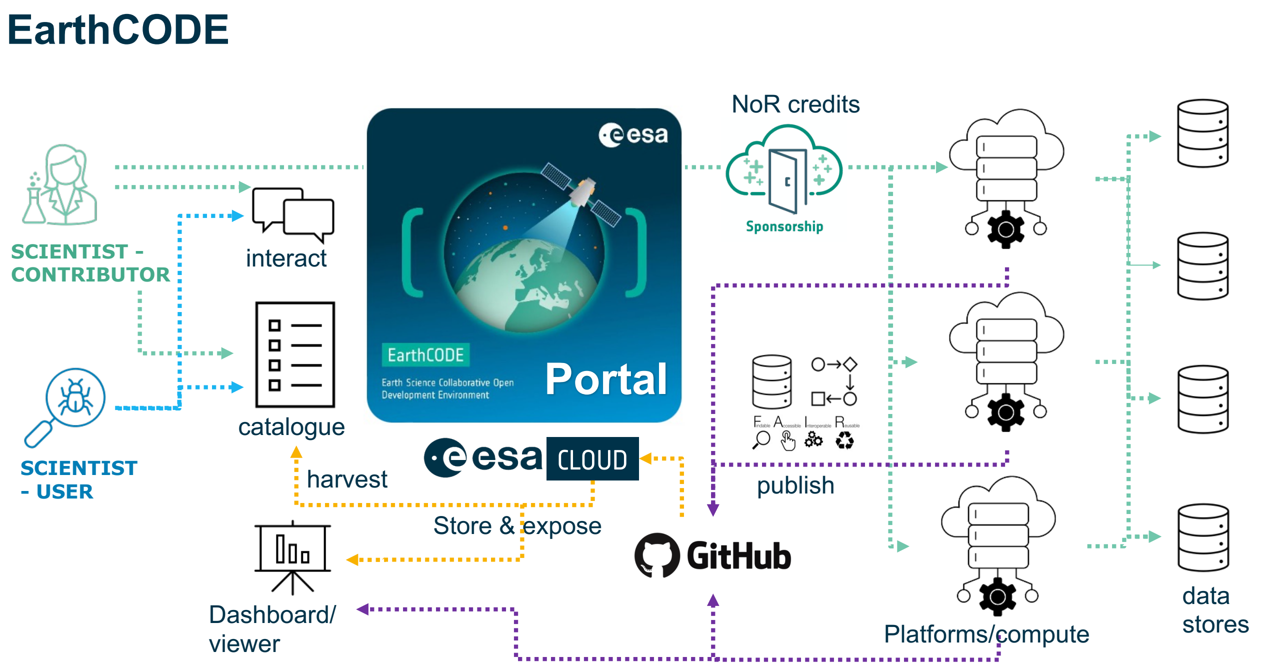

Figure 7: Summary of EarthCODE Vision

4. Conclusions

ESA's EarthCODE aims to provide a technology solution, aligned with the portfolio of related open science initiatives within ESA Earth Observation (such as, e.g., APEx - Application Propagation Environment(s) (ESA APEx, 2024)) integrate with existing platforms, to provide scientists with an open-access solution for ESA Scientific Activities. It will implement an open access cloud-based development environment using federated EO platforms. Phase 1 of this project aims to achieve integration of existing platforms within the EarthCODE portal, publish experiments and experimental output to Open Science Catalog and disseminate science. Phase 2 will focus on further integration, use of ESA EOEPCA+ building blocks, and achieve greater interoperability, providing an Earth Science collaborative development environment.

References

AGU Open Science, 2024 https://www.agu.org/learn-about-agu/about-agu/open-science (01 July 2024)

Murphy K., 2021, Open-Source Science: The NASA Earth Science Perspective, *The Earth Observer,* September – October 2021 Volume 33, Issue 5

Zellner P. J., Dolezalova T, Claus M, et al., "Cubes & Clouds - Cloud Native Open Data Sciences for Earth Observation", Software, v1.0.0, *Zenodo*, 2024, doi:10.5281/zenodo.10513914

EU, 2020. EU Open Science Policy, https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science (14 July 2023).

European Commission, 2021, DG-RTD, Horizon Europe, open science – Early knowledge and data sharing, and open collaboration, *Publications Office of the EU*, 2021, https://data.europa.eu/doi/10.2777/18252 (01 July 2024)

UNESCO, 2022, An introduction to the UNESCO Recommendation on Open Science, *Canadian Commission for UNESCO*, https://doi.org/10.54677/XOIR1696

OGC Open Science Persistent Demonstrator (OSPD) Initiative: Call for Participation (CFP) 2023, Version 1.3 – 10 November 2023, https://portal.ogc.org/files/?artifact_id=106421 (01 July 2024)

EOEPCA, 2021. Master System Design Document: Eoepca.Sdd.001. Https://Eoepca.Github.Io/Master-System-Design/Current/

EOEPCA, 2023a. Earth Observation Exploitation Platform Common Architecture. https://eoepca.org/.

EOEPCA, 2023b. EOEPCA Deployment Guide https://deployment-guide.docs.eoepca.org/current/quickstart/userman-deployment/ (14 July 2023)

CODATA, Committee on Data of the International Science Council, CODATA International Data Policy Committee, CODATA, C. on D. of the I. S. C., CODATA International Data Policy Committee, *CODATA and CODATA China High-level International Meeting on Open Research Data Policy and Practice*, Hodson, S., Mons, B., Uhlir, P., et al. (2019). The Beijing Declaration on Research Data., doi:10.5281/zenodo.3552330.

Schindler, F., Pari, S., Meissl, S., Smith, G., Dobrowolska, E., and Anghelea, A.: Open Science Data Catalogue, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*., XLVIII-1/W2-2023, 997–1003, https://doi.org/10.5194/isprs-archives-XLVIII-1-W2-2023-997-2023, 2023.

ESA Open Science Data Catalogue (ESA OSC) 2024 https://opensciencedata.esa.int (01 July 2024)

ESA Application Propagation Environment(s) (ESA APEx) 2024
https://remotesensing.vito.be/news/new-esa-apex-initiative-will-boost-reusability-eo-based-research-outcomes (01 July 2024)