EarthCODE: A Federated FAIR and Open Science Infrastructure for Next-Generation Spatial Data Infrastructures

Executive Summary

The next generation of Spatial Data Infrastructure (SDI) must meet the increasing demand for interoperable, machine-actionable, and reproducible data workflows that can operate across platforms, disciplines, and governance domains.

ESA's vision for EO Open Science and Innovation captures this opportunity, providing a structured framework to embed FAIR and Open Science practices across its Earth Observation activities. EarthCODE is part of this larger panorama of strategic initiatives. Through its ecosystem of tools and platforms, it aims to transform FAIR and Open principles from an aspiration to routine practice for Earth Science activities funded through its programme and beyond (e.g., including collaborations with EU funded research).

At a high-level, EarthCODE's federation of platforms supports open science by empowering scientists to 1. Access and process satellite and in-situ data in collaborative cloud environments, 2. Develop and publish reusable code and workflows across the federation of EO platforms, 3. Validate outputs and share reproducible results, and 4. Collaborate across institutional, disciplinary, and national boundaries.

SDI Challenges EarthCODE Addresses

EarthCODE faces common Spatial Data Infrastructures (SDIs) challanges which limit effectiveness, including fragmented metadata, and datasets and workflows tightly coupled to specific platforms. The measure–understand–predict–decide–act cycle in Earth Observation (EO) critically depends on digital research objects—data, software, workflows, models, and services—being findable, accessible, interoperable, and reusable (FAIR). Despite the growing focus on creating open science data catalogues from various institutions, FAIR implementation remains fragmented.

EO datasets are often cataloged but isolated in system-specific silos with inconsistent metadata and weak links to tools and workflows. Tools like PySTAC and stactools help improve metadata generation, but require platform-specific adaptations, limiting

automation and discoverability. Although open APIs and cloud services have improved data accessibility, challenges persist due to inconsistent authentication, documentation gaps, and unclear licensing. These issues also affect software and workflows, which are often transient and poorly documented, hindering their accessibility.

Interoperability remains a major barrier due to variations in formats, metadata standards, and execution environments. Even FAIR-compliant components often cannot be integrated without shared APIs, containers, or workflow engines.

Reusability is further limited by missing provenance, inadequate documentation, and the absence of best practices like semantic versioning or persistent identifiers. This hampers reproducibility and broader reuse.

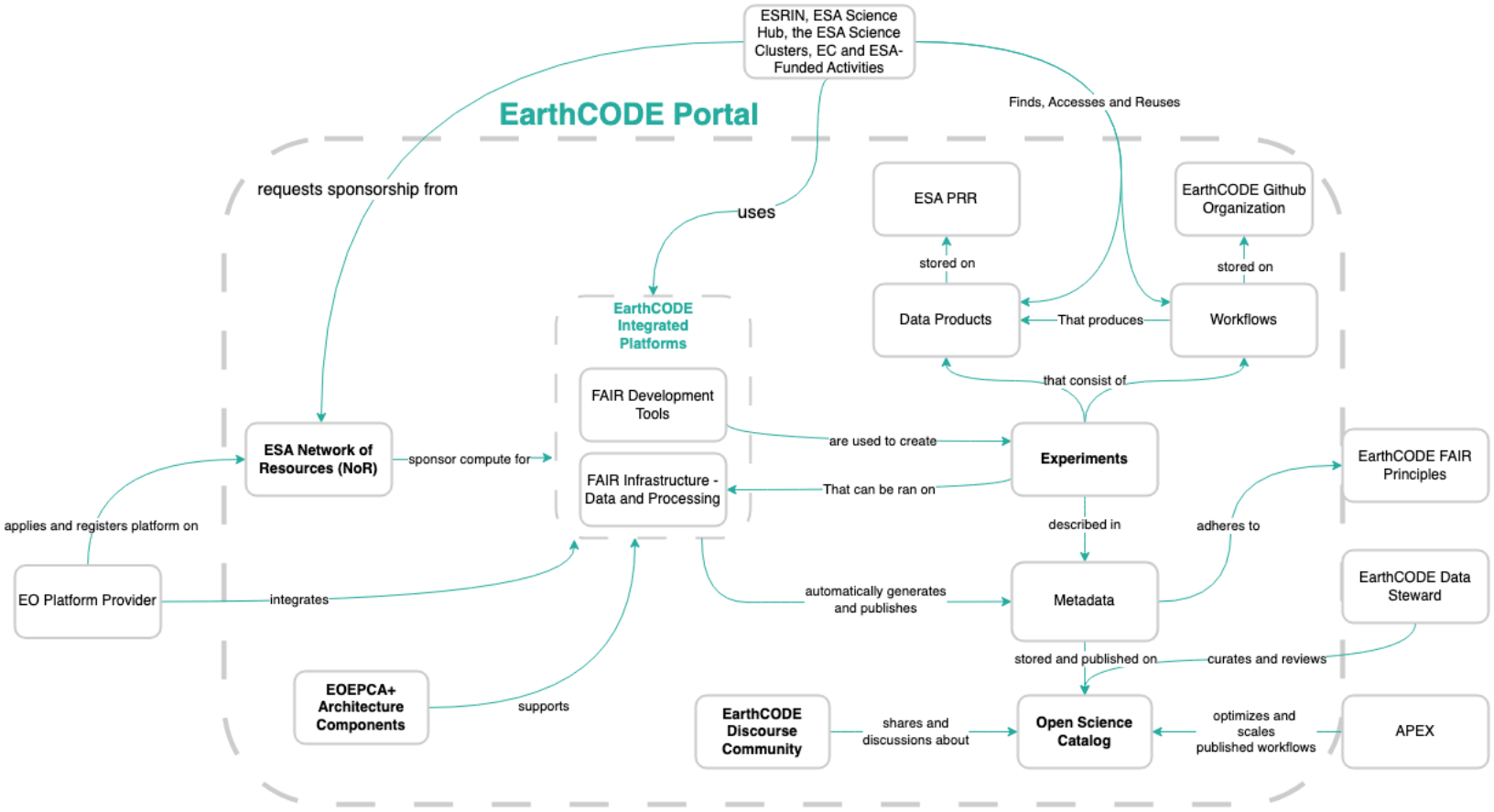
While progress has been made, especially in data accessibility, significant work is still needed to improve metadata quality and FAIRness for data and especially for workflows.

Implementing Open Science principles in the cloud era introduces additional technical complexities. Traditional paradigms assume “one-click” retrieval, but high-resolution global datasets from cloud-optimized EO platforms are too large for cost-effective download and separating them from native infrastructure undermines reproducibility and reusability.

Moving compute next to data solves this, but requires orchestration across diverse data infrastructures, each with unique storage layouts, metadata schemas, access protocols, and authorization. Additionally, Earth Observation workflows are often tightly coupled to specific data, infrastructure, and execution environments, limiting reuse and reproducibility. Although most EO cloud support STAC for data discovery, their compute interfaces remain highly heterogeneous. This fragmentation leads to a lock-in effect: code and pipelines built for one platform often require substantial modification to function on another’s infrastructure or to be reused, forcing scientists to repeatedly reengineer workflows rather than advancing science.

EarthCODE Federated Approach

EarthCODE overcomes these challenges by providing scientists with accessible tools and guidelines to practice FAIR & Open Science. It promotes coordination among various EO cloud providers to enable portable and reproducible science across a federation of platforms by using open standards. The federated ecosystem of EarthCODE goes far beyond openness, it strives to be open, FAIR, and reproducible.



In summary, EarthCODE offers a comprehensive and integrated environment that addresses the persistent challenges of modern SDIs:

- Provides a central access point via the EarthCODE Portal, offering a rich catalog of data and workflows, integrated platforms for scientific development and analysis, and FAIR tools to publish research outputs.
- Integrates platforms for workflow development, reproducibility, and publication.
- Leverages the ESA Network of Resources (NoR) to sponsor projects and enable access to these integrated platforms.
- Facilitates discovery and reuse through a metadata-rich Open Science Catalog.
- Promotes FAIR principles at all stages of research, supported by data stewards and governance mechanisms.

- Ensures long-term preservation of data results in the ESA Project Results Repository (PRR), source code on GitHub, and metadata within the Open Science Catalog.
- Supports community engagement and knowledge exchange via the EarthCODE Discourse forum.

EarthCODE's Users

EarthCODE is designed for scientists and research teams engaged in ESA-funded Earth Observation (EO) activities. Its primary aim is to enable the adoption of FAIR and Open Science practices across all stages of scientific development, ensuring that data, workflows, and documentation are reusable, transparent, and persistently available for long-term use.

EarthCODE directly supports and connects communities within the ESA Science Clusters, streamlining research efforts, promoting interoperability, and amplifying the impact of funded projects. The EarthCODE Portal is envisioned as the central hub for these scientific communities, including the ESRIN Science Hub, the ESA Science Clusters, and individual ESA-funded EO research projects. These groups currently form the main user base of EarthCODE.

As EarthCODE evolves, its scope will expand to support the broader Earth Observation science community, fostering cross-disciplinary collaboration and making high-quality, open scientific resources accessible to a global audience.

High-Level Use Cases

Projects leverage cloud platforms integrated with EarthCODE to develop, execute, and document their experiments, ensuring that results can be stored, shared, and reproduced.

The EarthCODE Portal (earthcode.esa.int) is the principal entry point, offering SSO access to the integrated platforms, the Open Science Catalog, and the EarthCODE Discourse forum.

The key components of EarthCODE that make this possible are described below:

EarthCODE Portal

The EarthCODE Portal serves as the central entry point to the EarthCODE federation —a web-based hub designed to facilitate open and collaborative Earth science. It is the primary interface through which scientists discover, develop, and publish FAIR and Open resources.

The portal offers a unified access point for all EarthCODE services, integrating user authentication, data and workflow discovery, and project publishing, while connecting researchers to a growing suite of platforms - it is the gateway to EarthCODE's federated ecosystem.

Key functions of the EarthCODE Portal include:

- Discovery of published research via the Open Science Catalog, encompassing data products and reproducible experiments.
- Publishing of new scientific outputs, supporting the linking of inputs, workflows, configurations, and results to promote reuse and reproducibility, whether via the EarthCODE publishing GUI or automated publishing through integrated platforms.
- Guided access to integrated FAIR Open Science Platforms, with SSO authentication via EarthCODE accounts and integration of tools for workflow development and data processing, including redirection to appropriate ESA Network of Resources (NoR) requests.
- Engagement with the EarthCODE Discourse Community, a dedicated forum for collaboration, knowledge sharing, and community-driven support across science clusters and research teams.
- Users who do not log in can still explore published resources and engage with the community. Logged-in users benefit from platform tools for workflow development, execution, and publication to foster transparent and reproducible science.

Open Science Catalog

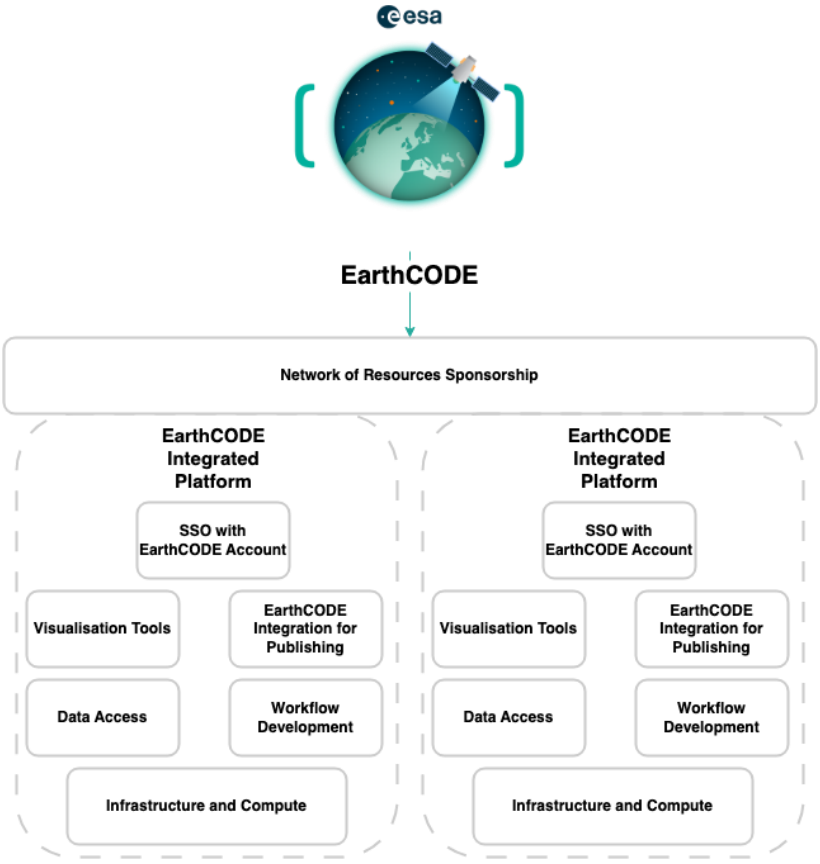
The Open Science Catalog (OSC - https://opensciencedata.esa.int/catalog) serves as the central interface for publishing, discovering, and accessing scientific resources generated through ESA-funded Earth Observation research. It is where metadata describing published data products, experiments, and workflows is made publicly available and reusable. Integrated platforms may also provide re-usable workflow services that support FAIR and Open Science practices within the catalog. It is built using STAC Browser and is a re-usable EOEPCA+ building block.

Researchers use EarthCODE platforms to generate and submit metadata, which—after validation and a manual review—is published to the OSC. Metadata describing datasets (Products) is published using the SpatioTemporal Asset Catalog (STAC) specification, while workflows and experiments are described using the OGC API - Records standard.

The OSC is further integrated with EarthCODE's integrated platforms. Users with appropriate access—such as NoR-sponsored compute—can reproduce experiments directly from the catalog, running the same workflow with the same input and configuration on a compatible platform.

EarthCODE Integrated Platforms

EarthCODE provides access to a suite of integrated EO cloud platforms, each offering tools, data, and compute environments tailored for scientific Earth Observation workflows.

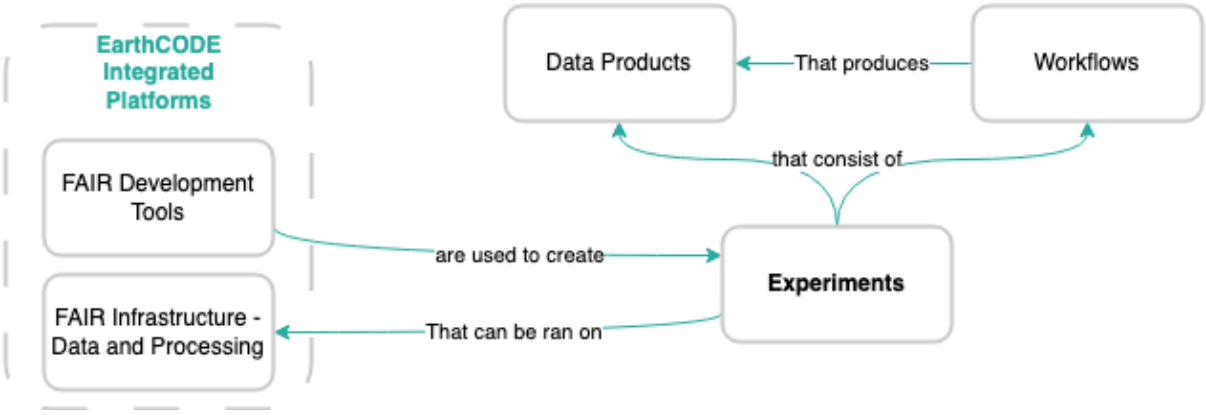


EarthCODE Integrated Platform Capabilities

From the user's perspective, EarthCODE's integrated platforms serve as collaborative environments for developing reproducible workflows, accessing data, and publishing items to the Open Science Catalog. Platform providers play a crucial role in this ecosystem by contributing in two key ways:

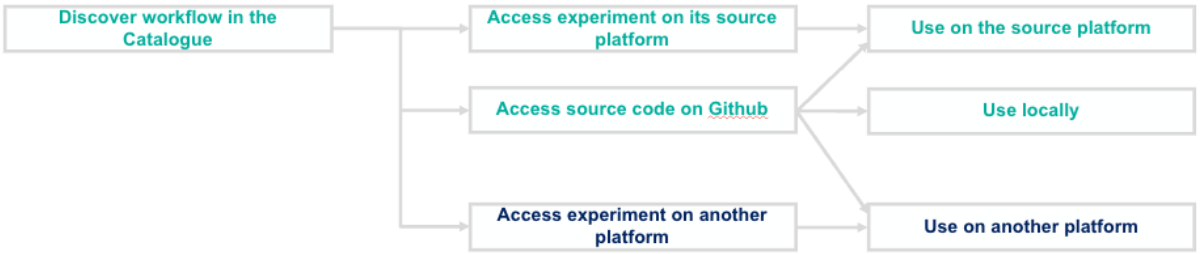
1. FAIR Open Science Platforms: These platforms enable researchers to create, discover, and reuse data and workflows, while also publishing results to the EarthCODE catalog.
2. FAIR Infrastructure Platforms: These provide the compute resources and data proximity needed to run or reproduce workflows at scale, supporting the EarthCODE paradigm of bringing code to the data.

Providers may offer (and typically provide both) capabilities for infrastructure and FAIR Open Science together or separately.



EarthCODE Integrated Platforms – Infrastructure and FAIR Tools

Through this separation of concerns between the tools that users use to build scientific workflows and the infrastructure that runs them, EarthCODE aims towards cross platform execution capabilities. With this, a workflow developed on one platform could then be executed on another one. The platforms need to implement the execution interface and fulfil the data access requirements described in the asset metadata. Users can also access and run workflows locally as the source code is all available and hosted on GitHub.



The ESA Network of Resources

Scientific projects can make use of the EarthCODE integrated platforms with sponsorship from ESA's Network of Resources (NoR). This mechanism allows eligible research and development activities to access integrated EarthCODE platforms, including sponsored processing, data access, and storage capabilities. Sponsorship is available for activities that do not generate revenue, such as scientific research, demonstration projects, and pre-commercial development.

By simplifying access to computing resources across a network of European platforms, NoR helps bring users to the data—supporting both scientific and operational uses of EO

data. Within EarthCODE, NoR plays a critical role in providing the computing resources needed to develop, reuse, or reproduce research as part of a modern SDI.

FAIR Open Science Platforms

FAIR Open Science Platforms within EarthCODE provide researchers with tools and environments to perform scientific research in line with FAIR (Findable, Accessible,Interoperable, and Reusable) and Open Science principles. These platforms support the complete research lifecycle—from initial data discovery, visualization and data exploration, reuse of existing datasets and code, through to workflow development and publication of resulting workflows and data to the ESA PRR and the Open Science Catalog.

Researchers have the choice to develop/run their scientific workflows entirely within EarthCODE's integrated FAIR Infrastructure Platforms or use their own local, institutional computing resources instead. Regardless of the computing environment, these platforms ensure researchers adhere to FAIR guidelines, enable them to develop and manage workflows, and publish FAIR items to EarthCODE.

Their key responsibilities are:

- Discovering, accessing, and reusing scientific datasets and code.
- Developing scientific workflows and producing experiments and data.
- Automated build, test and deployment (as container images) pipelines and source code management
- Providing tools to explore and visualize results data
- Integrating with the **ESA Project Results Repository (PRR)** for long-term data storage
- Publishing datasets, code, and results to the EarthCODE Open Science Catalog.
- Integrating with EarthCODE's Single Sign-On (SSO)
- Providing documentation about the integration and usage of their platform

FAIR Infrastructure Platforms

FAIR Infrastructure Platforms integrate scalable computational resources required to execute scientific workflows developed by researchers using FAIR Open Science Platforms or other tools (through standard protocols). These platforms offer robust cloud-based services, providing users access to compute, storage and data access to EO/Geopspatial datasets, significantly simplifying the transition from traditional on-premise research to scalable, cloud-native processing.

They focus on providing computational resources located close to the data, minimizing data transfer and ensuring efficient execution of scientific workflows at scale. Platforms also integrate necessary APIs, connectors, and interfaces to ensure smooth interoperability with the EarthCODE Portal and other integrated services. They provide an offering through the Network of Resources and are already production ready services. Their key responsibilities are:

- Provide scalable compute resources for big data Earth Observation (EO) analysis.
- Offering efficient data access to hosted EO/Geospatial datasets.
- Facilitate the execution of workflows developed on integrated FAIR Open Science Platforms.
- Supporting workflow execution close to data to minimize data transfer and enhance performance - to support the key EarthCODE paradigm of bringing code to the data.
- Integrating with EarthCODE's Single Sign-On (SSO)
- Register to the Network of Resources to ensure visibility and accessibility
- Providing documentation about the integration and usage of their platform

Publishing Experiments Data and Workflows

Once a research activity is complete, results can be published to the EarthCODE Open Science Catalogue, making them findable, reproducible, and reusable by the broader scientific community. For users working on integrated platforms, this publishing process is typically automated: once an experiment is finalized, the platform generates the appropriate metadata and pushes it to the EarthCODE Open Science Catalog.

The core units of data modelled and shared in EarthCODE are:

- Data Products: final outputs of scientific analysis - geospatial datasets.
- Workflows: the code or processing steps used to generate those products and the computing environment required to run them.

Data

In EarthCODE, the final outputs of research—referred to as Products or data products—are stored, described, and published in a way that ensures long-term FAIRness and availability.

These data products can be hosted in the ESA Project Results Repository (PRR) or in an external persistent repository of choice. The PRR serves as ESA’s dedicated long-term storage service for project results. Uploading to the PRR is recommended to ensure FAIR

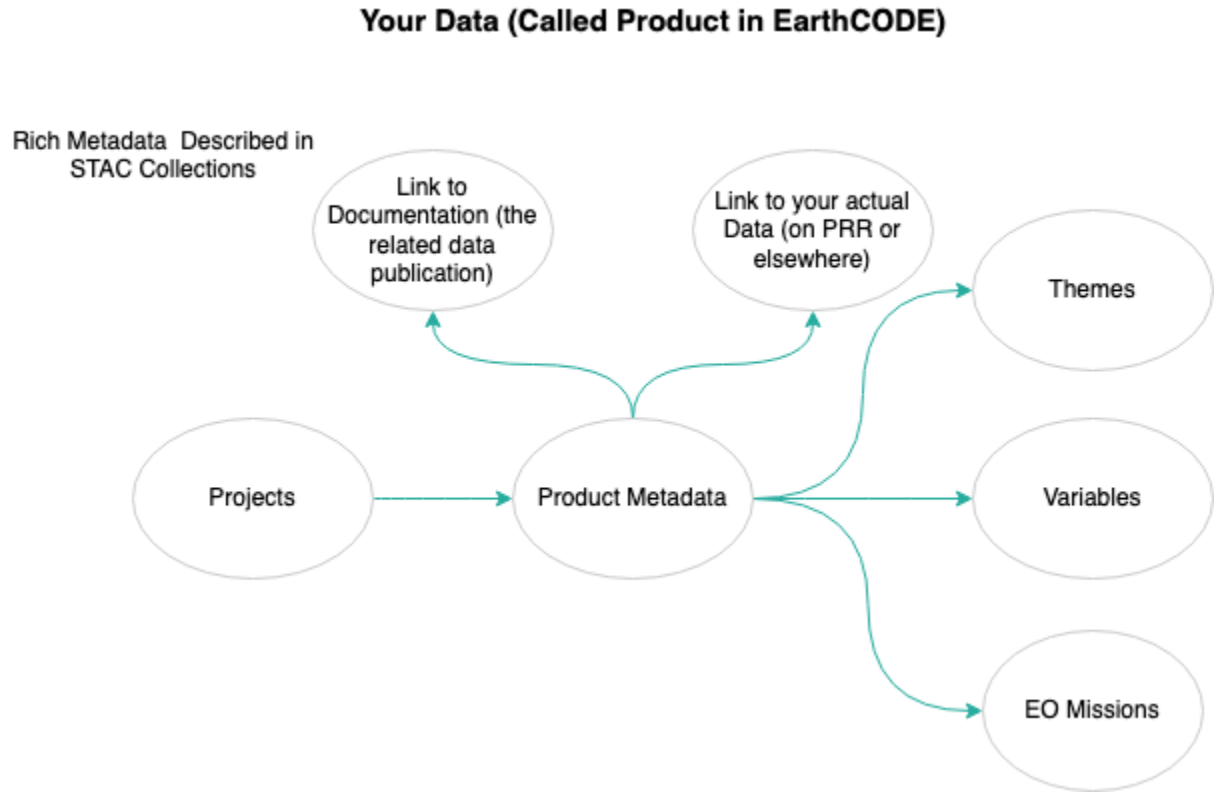
compliance and persistent access, but users may opt for other repositories that are established in their communities.

Each product is described using STAC (Spatio-Temporal Asset Catalog) metadata, captured in a Collection that includes attributes such as spatial and temporal extent, scientific context, and provenance.

A product in EarthCODE typically includes:

- A dataset representing measured or derived environmental or geoscience variables.
- Documentation links that describe the methodology or related publications.
- Metadata fields capturing EO satellite missions, project affiliation, and classification tags.

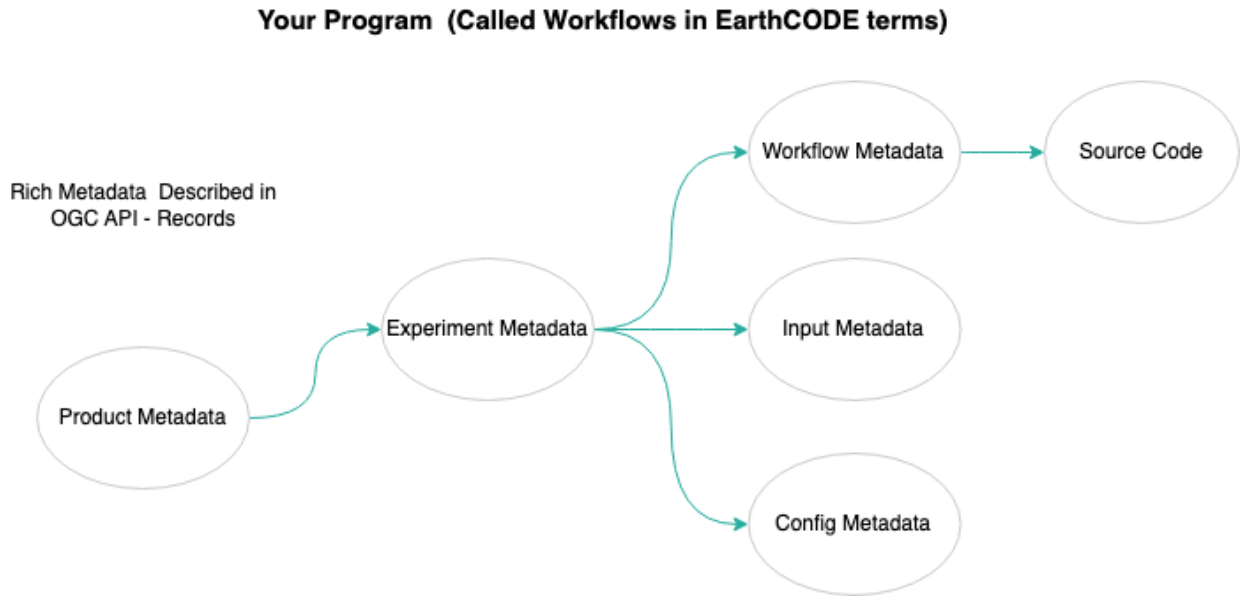
To ensure FAIRness, the catalog uses a shared dictionary and metadata standard. This structure enables exploration across sources by theme, variable, and mission. The terms are described as a common STAC extension which can be found at: https://github.com/stac-extensions/osc, or summarised as below:



- Product: a geoscience dataset covering specific variables, spatial extent, and time period.
- Project: the ESA-funded project associated with the product.
- Variables: measured or derived scientific variables in the dataset.
- Themes: top-level science topics aligned with ESA's strategic challenges (e.g. climate, biodiversity, atmosphere).
- Keywords: hierarchical tags to aid product discovery.
- EO Mission: the satellite mission or sensor referenced in the metadata.
- Documentation: links to publications or supporting materials.
- Data Link: a pointer to the dataset's actual location (PRR, institutional archive, or external repository).

Workflows

In EarthCODE, a published product is more than a dataset—it is the result of a defined process captured as an Experiment. Products include references to the experiments that generated them, supporting understanding, reuse, and reproducibility.



An Experiment is defined using the OGC API Records schema and defines:

- A human-readable summary of purpose and context.
- A machine-executable workflow to transform inputs into outputs.
- Definitions of input datasets, referencing other published products.
- A configuration specifying execution parameters.

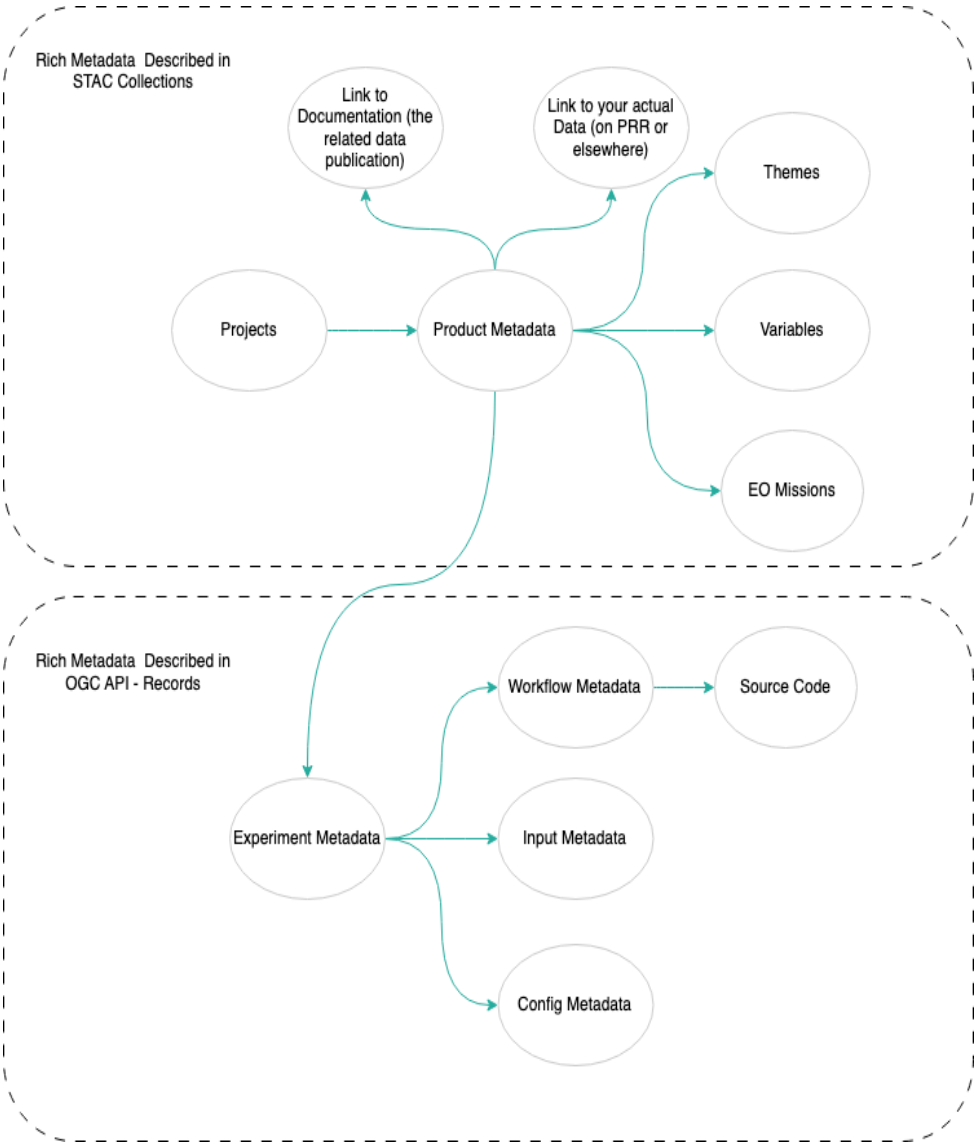
A workflow in EarthCODE defines the processing steps of an experiment. Unlike mere source code, a workflow is formally executable within a platform using a defined interface. Workflow types include:

- openEO Process Graphs
- OGC API Processes (e.g. CWL, Application Packages)
- MLflow models
- Jupyter Notebooks and others

Source code needs to be referenced, but the workflow itself must be described for execution on EarthCODE platforms, ensuring reproducibility and cross-platform compatibility. Workflows are typically stored in the EarthCODE GitHub organization and referenced in the Open Science Catalog as part of the metadata.

Experiments also declare input datasets and a configuration—a set of parameters for runtime execution. Inputs are referenced by unique identifiers to support validation and re-execution. Configuration values typically follow simple name–value pairs but can vary in complexity.

In summary, EarthCODE combines workflows and products: A product is the result of a successfully executed experiment. The product metadata links back to the experiment metadata, which references the workflow, input, and configuration—together ensuring reproducibility, FAIRness, and openness.



EarthCODE Global Collaboration

EarthCODE is integrated with the larger picture of innovative initiatives, such as for example APEx (https://apex.esa.int) , EOEPCA+ and the Open Science Persistent Demonstrator (OSPD). APEx, developed by the European Space Agency (ESA), complements EarthCODE by transforming EO research outcomes—such as algorithms and workflows—into interoperable and scalable cloud-ready services. APEx optimizes these outputs and integrates them into a dynamic Algorithm Services Catalogue, enhancing discoverability and operational impact.

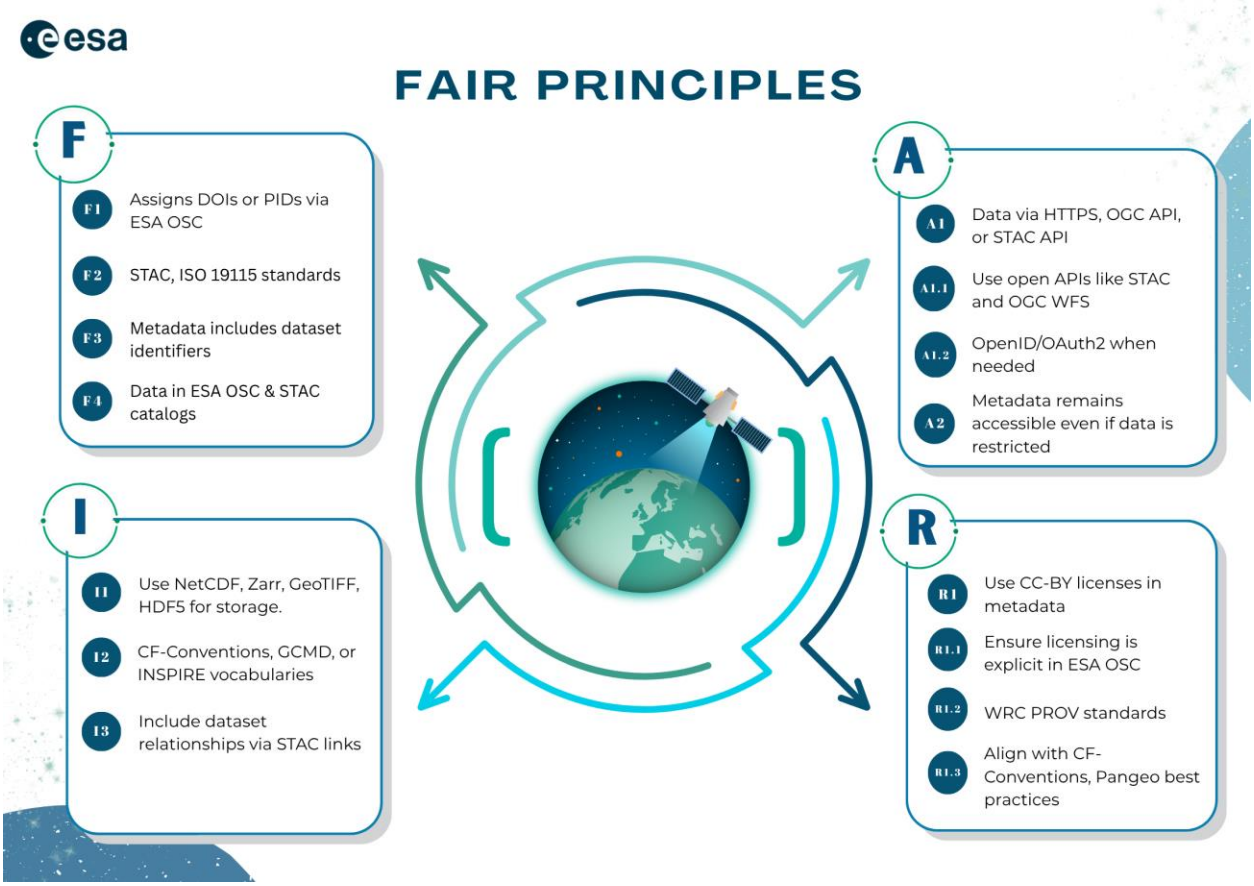
The Open Science Persistent Demonstrator (OSPD, https://www.ogc.org/initiatives/open-science/) is a long-term inter-agency initiative aiming to enable and communicate

reproducible Earth Science across global communities of users and amplify inter-agency Earth Observation mission data, tools, and infrastructures.

Demonstrative SDI Example of EarthCODE Assets

EarthCODE aims to fully embrace FAIR principles, following guidance for workflows as described in Applying the FAIR Principles to Computational Workflows by S. R. Wilkinson et al. (Sci. Data, vol. 11, no. 1, 2025; doi: 10.1038/s41597-025-04451-9) and for data as described in The FAIR Guiding Principles for Scientific Data Management and Stewardship by M. D. Wilkinson et al. (Sci. Data, vol. 3, 160018, 2016; doi: 10.1038/sdata.2016.18).

To illustrate how these principles are practically implemented within EarthCODE, this section presents a demonstrative example of an experiment published in the Open Science Catalog (OSC), highlighting the metadata standards, reproducibility practices, and cross-platform capabilities that EarthCODE enables.



The results from the execution of the experiment (https://opensciencedata.esa.int/experiments/worldcereal-experiment/record) are described as a data product, published as a STAC Collection (F2, I1) enriched with EarthCODE taxonomy elements (Themes, Variables, EO Missions) (I2). Metadata explicitly includes standardized references to the dataset and its components (F3), data is indexed in the ESA Project Results Repository(F4). Qualified references to related datasets and workflows are included (I3), and standardized EO formats (COG/TIFF) ensure scalable access (A1.1). Licensing and provenance are recorded using open, standardized practices (R1.1, R1.2), and all metadata aligns with community standards widely adopted in Earth Observation (R1.3) such as STAC.

All metadata including input, configuration, workflows, experiments and products on the catalog are assigned persistent, globally unique identifiers (F1, F1.1) and are indexed and searchable through the Open Science Catalog (F4), held separately from the data (A2). The catalog exposes data via open, standardized protocols such as HTTPS, STAC API, and OGC WCS (I.1, A1, A1.1, A1.2).

Together, the WorldCereal resources form a fully FAIR research object chain, where data, infrastructure, methods, and outputs remain findable (F1–F4), accessible (A1–A2),

interoperable (I1–I4), and reusable (R1–R3) across platforms, infrastructures, and research communities. This metadata is automatically generated and published via integrated platforms.

Future Vision for SDI

EarthCODE stands as a living exemplar of moder SDIs, demonstrating how interoperable, FAIR-compliant data and workflows can be made operational, reproducible, and impactful at scale. Its federated, standards-based approach directly addresses persistent challenges in SDIs—bridging data silos, enabling cross-platform execution, and promoting collaboration.

Aligned with national and international spatial strategies, EarthCODE is committed to contributing further under OGC review—sharing lessons learned, best practices, and actionable insights that advance the global SDI community as well as integrating into the larger picture of the SDI Hub. Through EarthCODE, we aim to support and inspire new models of spatial data infrastructure that are open, inclusive, and designed for the scientific challenges of today and tomorrow.