OPEN SCIENCE DATA CATALOGUE

F. Schindler[1], S. Pari[1], S. Meissl[1], G. Smith[2], E. Dobrowolska[3], A. Anghelea[4]

[1]EOX IT Services GmbH, Vienna, Austria, fabian.schindler@eox.at
[2]Telespazio UK, Luton, UK, garin.smith@telespazio.com
[3] Serco Italy S.p.A., Frascati, Italy, ewelina.dobrowolska@serco.com
[4]European Space Agency, Frascati, Italy, anca.anghelea@esa.int

KEY WORDS: Open Science, STAC, FAIR, Cloud Platform, Earth Observation, Collaboration.

ABSTRACT:

Open Science is a catalyst for innovation. Across the Earth Observation value chain, from R&D to prototyping new products and development of commercial applications, openness can play an important role by promoting long-term sustainable, community-contributed science and technology. The FAIR principles provide essential support to implementing Open Science, by offering guidelines for how researchers can adapt their EO and Earth Science practice to enable that their work (taking place increasingly in the cloud) and results are discovered, accessed, used, and reproduced by others. The Open Science Data Catalogue (OSC) (https://opensciencedata.esa.int) 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. This paper reviews the EO OSC architecture, technology stack, and illustrates how this tool can be used to discover and publish Earth System Science products from ESA activities. It also looks at future evolutions of the product and how it contributes to ESA's EO Open Science and Innovation goals.

1. INTRODUCTION

Open Science is increasingly recognized as a catalyst for innovation. In 2016, the EC's DG-RTD laid a vision for European R&D (EC, 2016) which acknowledged that "the way that science works is fundamentally changing, and an equally important transformation is taking place in how companies and societies innovate. The advent of digital technologies is making science and innovation more open, collaborative, and global". The EC has since expanded its views and interest in Open Science and most recently has published a renewed Open Science policy (EU, 2020).

The concept of Open Science and Innovation is embraced by the European Space Agency in its Agenda 2025 (ESA, 2022), recognizing the value that such principles of innovation can bring for the space sector in terms of optimizing development cycles, accelerating time to market, and reducing cost. Adhering to principles of openness in EO and Earth System Science from the earliest stage of the value chain, i.e., the scientific research phase, can contribute to a sustainable creation of value, potentially resulting in more innovation.

The Open Science Data Catalogue is one of the elements contributing to an Open Science framework and infrastructure, with the scope to enhance the discoverability and use of products, data and knowledge resulting from Scientific Earth Observation exploitation studies.

The Open Science Data Catalogue is a publicly available platform (available at https://opensciencedata.esa.int/) that contributes to ESA's Earth Observation (EO) Open Science framework. It stores geoscience products, datasets and resources developed in the frame of scientific research projects funded by ESA EO grouped by themes (or scientific domains) associated to the ESA Science Clusters, through which ESA aims at contributing to the establishment of European research areas in close collaboration with the European Commission Directorate General for Research and Innovation and other European and international partners. Details ablut ESA's Science Clusters are available at https://eo4society.esa.int/communities/scientists/. The Open Science Data Catalogue brings new functionalities, provides discovery and open access for geospatial products and documentation (or/and code) to a scientific community of users. With common dictionary and unified metadata across heterogeneous sources, products discovery is facilitated. Published items are also open to community contribution and curation, with all activity tracked on a public GitHub project.

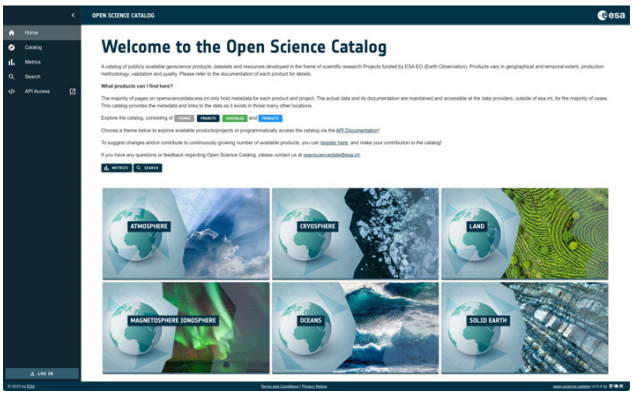

Figure 1. Open Science Data Catalogue Landing Page. The main themes of the OSC are visible on this page.

Finally, it allows for synoptic view for Earth Observation and Earth System Science gap analysis, by providing a dashboard view with statistics on the geophysical variables available in the catalogue., the EO missions providing the underlying data for the respective products and the geographical coverage of the data. Currently, the actual data and its associated documentation published on Open Science Data Catalogue are maintained and accessible by the data providers, outside of esa.int, for most cases. The catalogue provides the metadata and links to the data as it exists in those many other locations. Work in progress looks at improving the long-term availability by facilitating publication of products in community maintained and curated repositories.

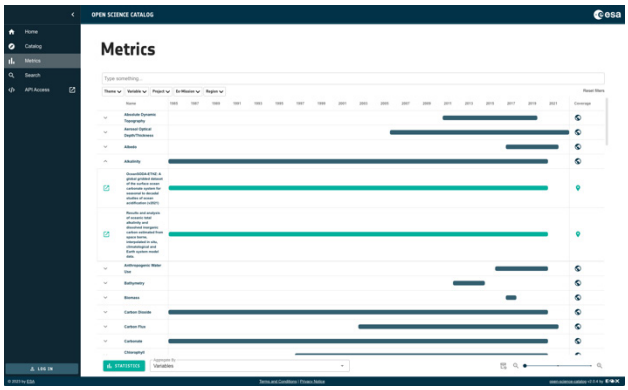

Figure 2. The OSC Metrics Page, providing an overview of the available geospatial products, and their temporal and spatial coverage.

The Open Science Data Catalogue has the capability to hold product metadata for assets that are stored externally or internally if required. Open Science Data Catalogue is also developing the capability to discover processes that can be deployed to and executed on remote EOEPCA platforms, using assets discovered by Open Science Data Catalogue.

The Open Science Data Catalogue (OSC) is based upon the EO Exploitation Platform Common Architecture (EOEPCA) and shares its basic Open Source components, but extends it with additional functionalities, including:

1. The Static Catalogue – which is a hosted STAC Catalogue, comprised of static Catalogue, Collection, and Items that represent the Themes, Variables, Projects, and Products.
2. The Open Science Data Catalogue Frontend – which is a Vue.js based client application, that allows the efficient browsing of the Open Science Data Catalogue.
3. The Backend API which allows users to make submissions to create, update, and delete Themes, Variables, Projects, Products and EO-Missions. These submissions are then handled as GitHub Pull Requests, where they can be further reviewed, discussed, and finally accepted or denied.
4. The process execution – Using the Application Deployment and Execution Service (ADES) building block of EOEPCA, it is possible to run the scientific workflows on platforms implementing ADES, to re-generate Products from the Catalogue, enabling reproducibility and contributing to more transparent research. The processing is done in a federated fashion, allowing the processing close to the input data.

Adhering by design to the "FAIR" (findable, accessible, interoperable, reproducible/reusable) principles, the Open Science Data Catalogue aims to support better knowledge discovery and innovation. It facilitates data and knowledge integration and reuse by the scientific community.

2. OPEN SCIENCE DATA CATALOGUE ARCHITECTURE

The Open Science Data Catalogue is a deployment of the EOEPCA (EOEPCA, 2023a) components in conjunction of additional components to facilitate open access to a catalogue of science projects and products.

EOEPCA is an ESA activity aiming to provide a blueprint for an EO exploitation platform that attempts to facilitate interoperability by tackling some key problems. The latest software building blocks are freely available as source code on GitHub and as docker images on DockerHub.
The reused components from EOEPCA are as follows (EOEPCA, 2023b):

- Resource Management:
  - Resource Catalogue - provides a standards-based EO metadata catalogue that includes support for OGC CSW / API Records, STAC and OpenSearch.
  - Harvester and Registrar - The Data Access provides standards-based services for access to platform hosted data - including OGC WMS/WMTS for visualisation, and OGC WCS for data retrieval. This component also includes Harvester and Registrar services to discover/watch the existing data holding of the infrastructure data layer and populate/maintain the data access and resource catalogue services accordingly.
- User Management:
  - Login-Service
  - Policy Decision Point
  - User Profile

The components supplementing the EOEPCA components are as follows:

- Fronted
- Metadata Proxy
- Backend API
- Metadata Repository
- Static Catalogue

The Figure 4 shows the interaction of the reused and supplementary components of the Open Science Data Catalogue.

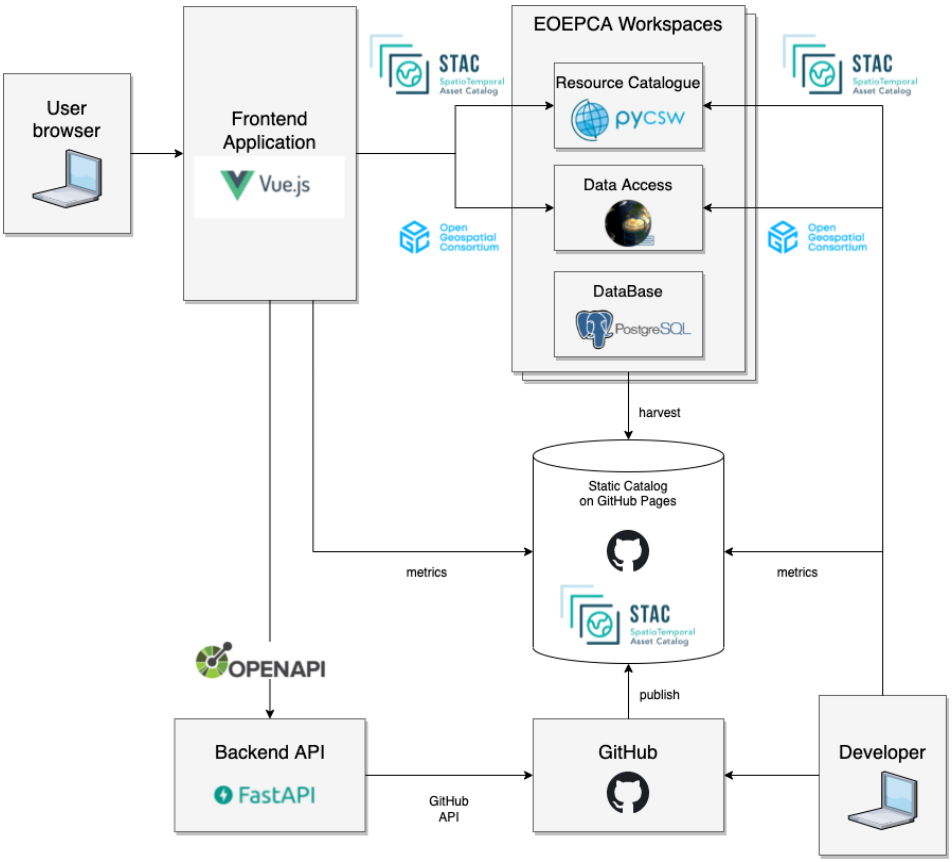

Figure 3. Open Science Data Catalogue design schema.

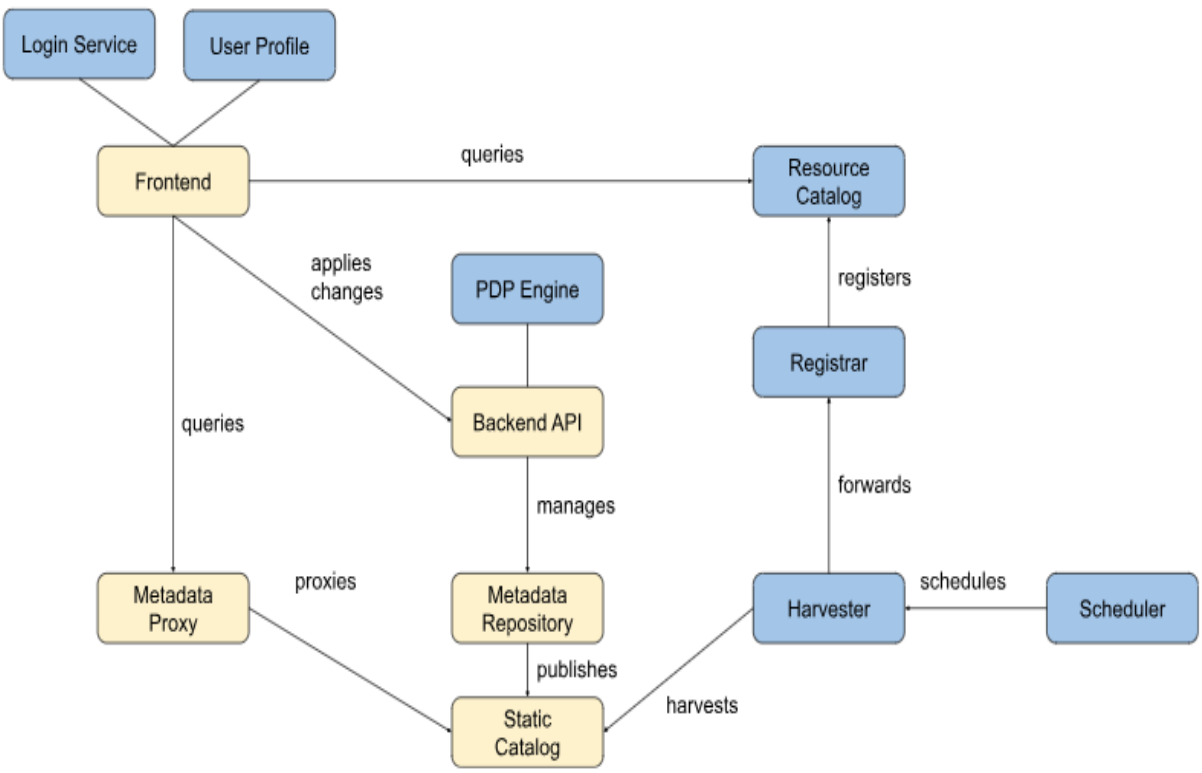

Figure 4. Interaction diagram of the Open Science Data Catalogue components.

2.1 Components

2.1.1 Frontend: The OSC frontend is the main user interface component that allows scientists and other parties to interact with the system and consume the contents. It allows users to search for scientific products, as well as to contribute to the contents of the catalogue by ingesting new products in the catalogue or submitting requests for updates of already existing content.

The Frontend is based upon the Open Source application STAC Browser (STAC Browser, 2023a) extended with functionality from the Vue framework. The main elements of the frontend are presented in the following paragraphs. On the landing page users are presented with the current OSC themes and can access the main functionalities of the OSC (Search, Catalogue, Metrics, API).

The OSC Catalogue page allows users to discover the available Collections and Items. The STAC specification describing all the components and their properties is available at (STAC, 2023f).

Since all the scientific Products in the OSC are outcomes of specific research projects, thus the OSC provides the possibility to discover the specific activities to ensure traceability and facilitate scientific exchanges between the data owners (i.e. the data producers and distributors) and the community accessing and using the data.

2.1.2 Metadata Proxy: This is a small reverse proxy to enable browser access to certain web services that do not provide the necessary CORS headers.

2.1.3 Backend API: This REST API service allows to submit contributions from either the frontend or any other compliant component to the contents of the catalogue. It translates all contributions to GitHub Pull Requests. The Backend API is built using the FastAPI software framework.

2.1.4 Metadata Repository: The metadata of the Open Science Data Catalogue items are stored in a git repository. This allows the convenient management of the contents, with the possibility to add changes in the form of branches/Pull Requests, that can be reviewed by metadata administrators and to be finally submitted to the main branch of the repository. The metadata repository is hosted and managed by GitHub.

2.1.5 Static Catalogue: This is an export of the contents of the source of truth of the Open Science Data Catalogue. Upon any change on the metadata repository, the static catalogue is re-built an exported for the consumption of the frontend or any other compliant component. The static catalogue itself is a structure of STAC objects in various inter-linked JSON files.

2.1.6 Resource Catalogue: This component loads all records from the static catalogue and provides convenient ways to search across the contents for various metadata filters. It is considered to be a cached version of the static catalogue and must be regularly synchronized. The resource catalogue is realized by the Open Source software pycsw. Internally, the records are stored in a PostgreSQL database with PostGIS extensions enabled.

2.1.7 Harvester: The harvester runs in regular intervals and reads the contents of the static catalogue. The harvested data items are them pushed forward for registration.

2.1.8 Registrar: This component is responsible to accept all harvested items and push them to the resource catalogue, either adding, altering, or removing elements as necessary.

2.2 Data Model

2.2.1 STAC Catalogue: The contents of the metadata repository are kept as a static STAC Catalogue, a collection of inter-linked JSON files and supplementary metadata. It is graph structure, with a single root STAC Catalogue as an entry point which has the following branches:

- Themes
- Variables
- EO-Missions
- Projects
- Products

Each element in turn is a listing of a number of elements of that type, which are in turn represented as a STAC Catalogue or STAC Collection. These objects use the OSC STAC extension to reference elements of other groups they are associated with. e.g a Product has an “osc:variables” field, that lists the measurement variables this product is comprised of.

To allow the easier management of the catalogue, the linking is done simply by using identifier fields. Actual STAC links are added, when the catalogue is exported.

When exported (when a change to the main branch of the metadata repository is merged), then the Static Catalogue is built. Here, the contents of the metadata repository are taken, and STAC link objects are introduced to link the related files. Additionally, search keywords are added to allow a later retrieval.

The STAC Catalogue makes use of various STAC extensions to best describe the contents. Most notably the scientific (sci), subjects, projection (proj), and datacube extensions.

2.2.2 STAC API: Once harvested into the resource management, the STAC API of the Resource Catalogue allows efficient searching using text, geospatial, temporal and other metadata attributes.

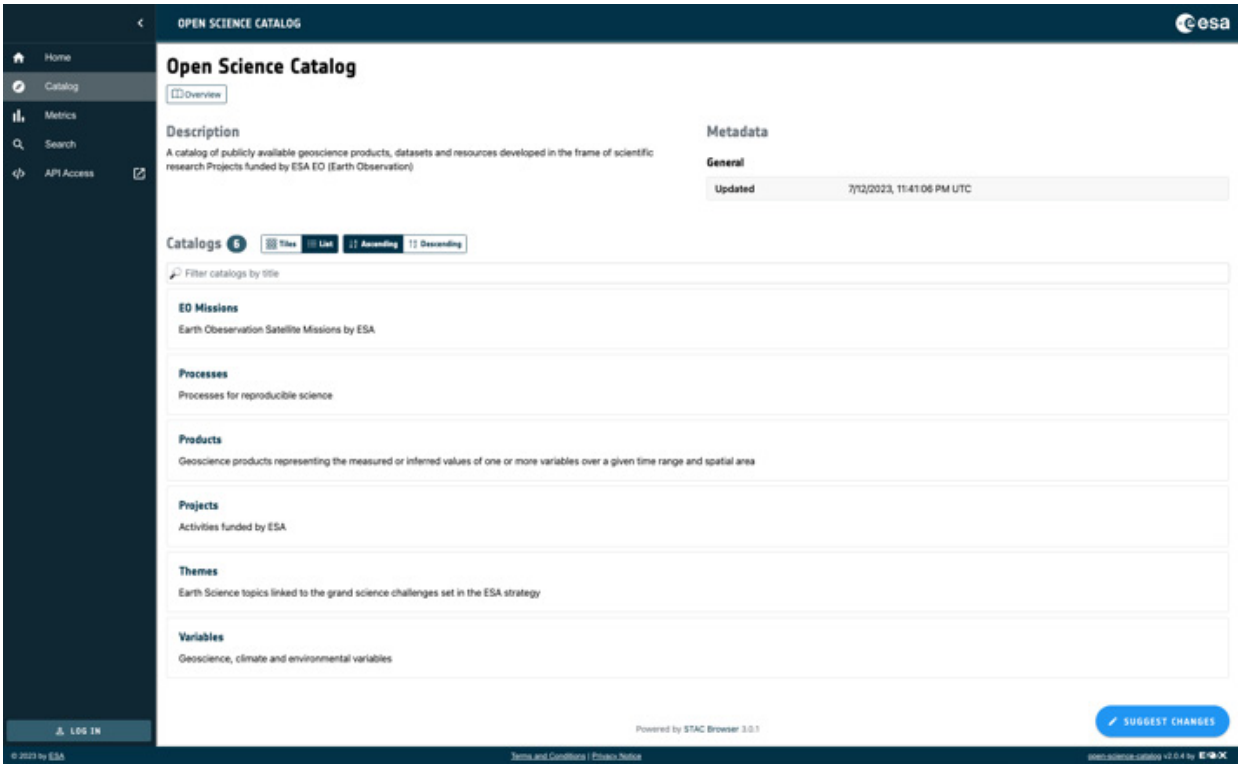

Figure 5. Open Science Data Catalogue Page

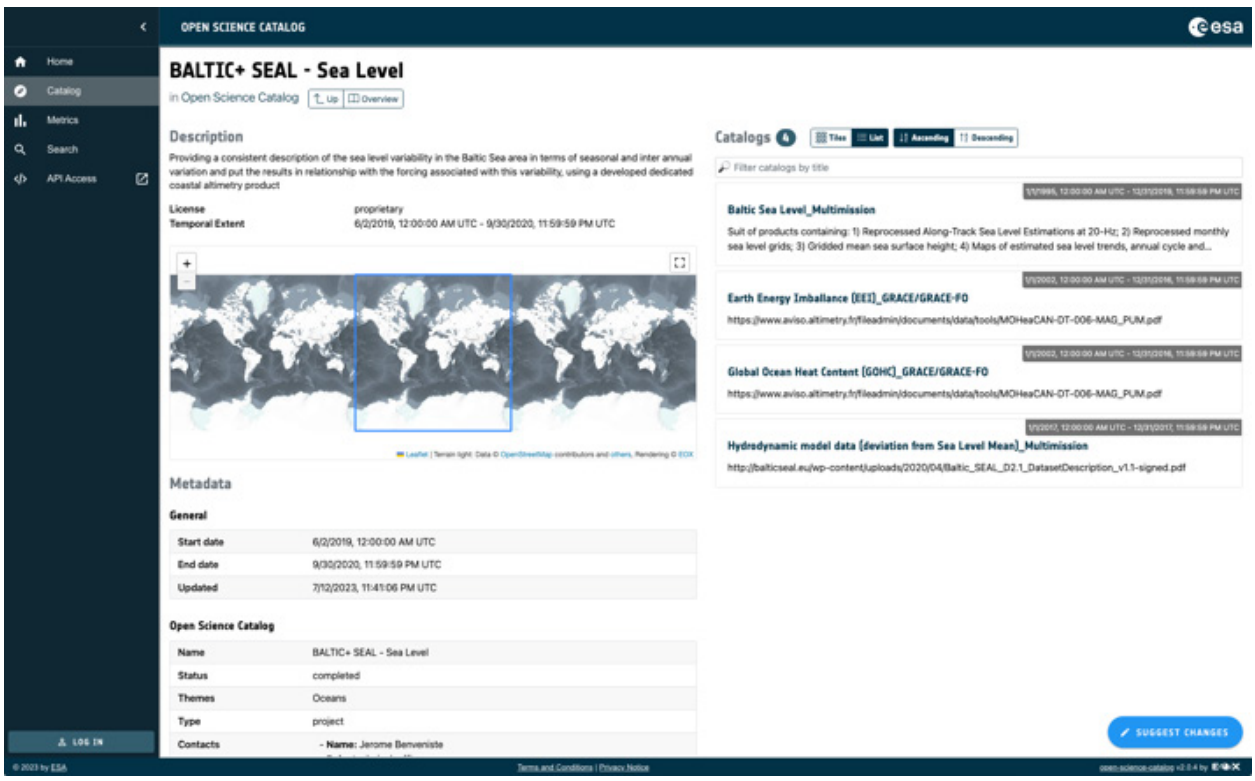

Figure 6. An example OSC Project Page. The Project metadata and the associated Products (i.e., Project outcomes) are visible on this page.

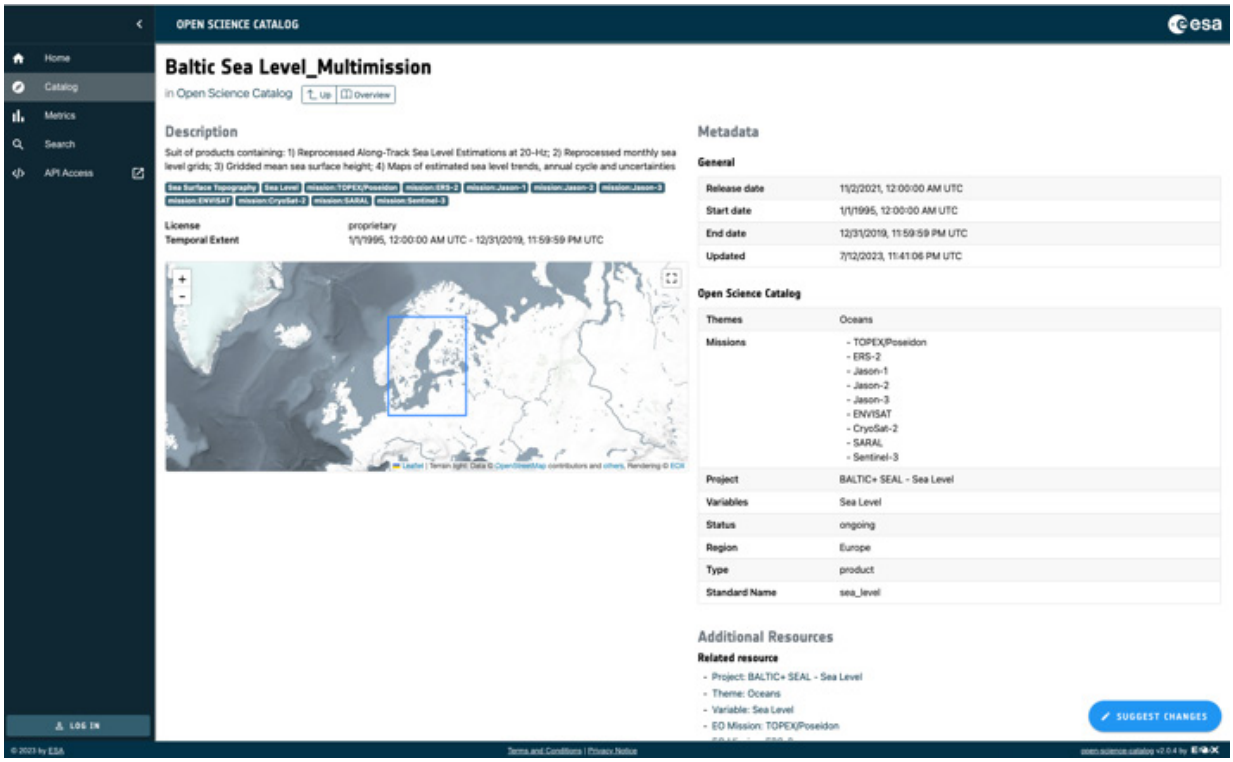

Figure 7. An OSC Product Page. A Product is a STAC item, while the various datasets included in the Product are Assets of the respective Item.

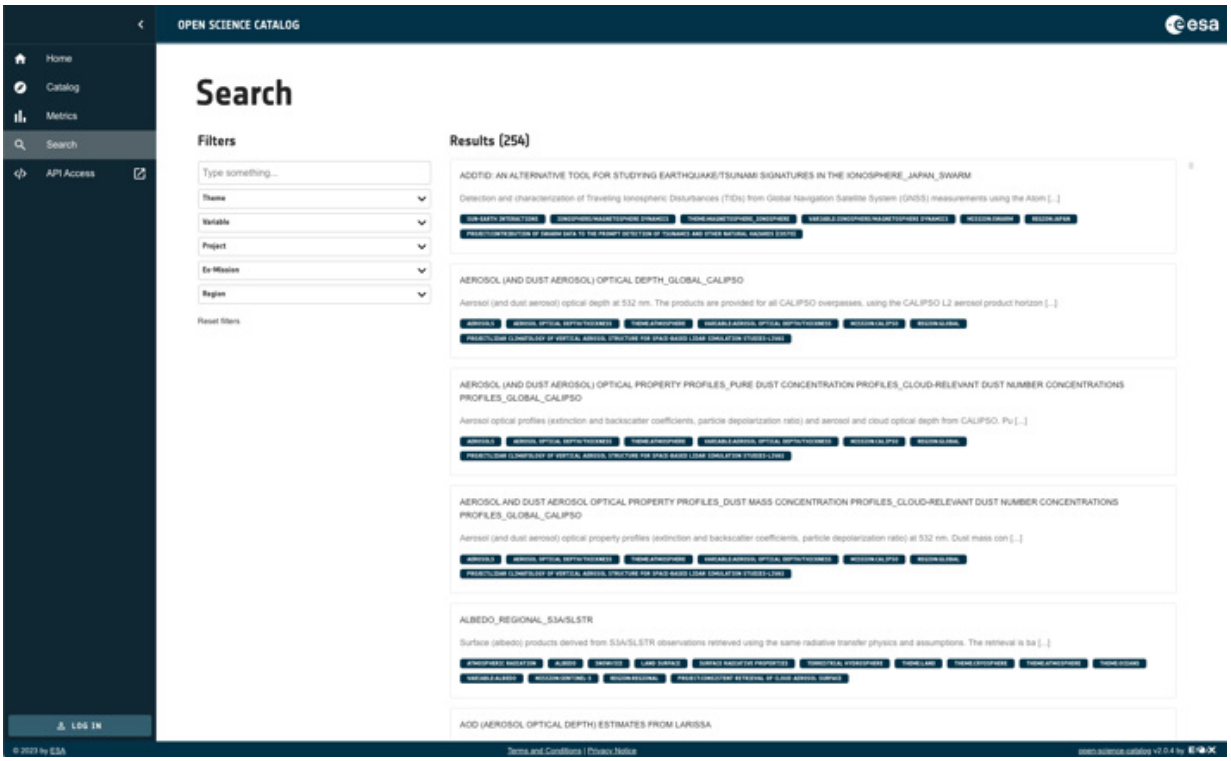

Figure 8. The OSC Search Page. In addition to searching by free text or keywords, users can search by Theme, Variable, Project, EO Mission and Region.

2.3 Deployment

The whole deployment of the Open Science Data Catalogue system is orchestrated using Flux CD, which is connected to a git repository, containing the manifests and metadata of the to be deployed components. Every change to that repository, reflects a change to the kubernetes cluster, the Open Science Data Catalogue is deployed on. This enables a convenient and reliable way to configure the cluster, with a full history of configuration changes.

3. CONCLUSIONS

The Open Science Data Catalogue is implemented using EOEPCA building blocks. It provides a catalogue of publicly available EO and Earth Science products, datasets, workflows, and other resources developed in the frame of scientific research Projects funded by ESA EO, and one of the elements of the Open Science framework and infrastructure supporting reproducible Earth System Science research with Earth Observation Data.

ACKNOWLEDGEMENTS

This work has been carried out under the EO Science for Society programme of and funded by the European Space Agency: EOEPCA (Earth Observation Exploitation Platform Common Architecture) https://eoepca.org/.

REFERENCES

EC, 2016. Open innovation, open science, open to the world. https://digital-strategy.ec.europa.eu/en/library/open-innovation-open-science-open-world (14 July 2023).

EOEPCA, 2021. Master System Design Document: EOEPCA.SDD.001. https://eoepca.github.io/master-system-design/current/ (14 July 2023).

EOEPCA, 2023a. Earth Observation Exploitation Platform Common Architecture. https://eoepca.org/ (14 July 2023).

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

ESA, 2022. ESA Agenda 2025. https://www.esa.int/About_Us/ESA_Publications/Agenda_2025 (14 July 2023).

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

Evan You, 2023. Vue.js, The Progressive JavaScript Framework. https://vuejs.org/ (14 July 2023).

Kubernetes, 2023. Kubernetes - Production-Grade Container Orchestration. https://kubernetes.io/ (14 July 2023).

OSGeo, 2023. Pycsw https://pycsw.org/. (14 July 2023).

PostGIS PSC & OSGeo, 2023. PostGIS. http://postgis.net/. (14 July 2023).

PostgreSQL Global Development Group, 2023. PostgreSQL: The World's Most Advanced Open Source Relational Database. https://www.postgresql.org/. (14 July 2023).

STAC Community, 2023a. Spatio Temporal Asset Catalogue (STAC) Browser repository. https://github.com/radiantearth/stac-browser (14 July 2023).

STAC Community, 2023b. Spatio Temporal Asset Catalogue (STAC) Datacube extension repository. https://github.com/stac-extensions/datacube (14 July, 2023).

STAC Community, 2023c. Spatio Temporal Asset Catalogue (STAC) extension repository. https://github.com/stac-extensions/ (14 July 2023).

STAC Community, 2023d. Spatio Temporal Asset Catalogue (STAC) Open Science Catalogue extension repository. https://github.com/stac-extensions/osc (14 July 2023).

STAC Community, 2023e. Spatio Temporal Asset Catalogue (STAC) Projection extension repository. https://github.com/stac-extensions/projection (14 July 2023).

STAC Community, 2023f. Spatio Temporal Asset Catalogue (STAC) Specification. https://github.com/radiantearth/stac-spec (14 July 2023).

STAC Community, 2023g. Spatio Temporal Asset Catalogue (STAC) Themes extension repository. https://github.com/stac-extensions/themes (14 July 2023).

The Flux, 2023. Flux - the GitOps family of projects. https://fluxcd.io/ (14 July 2023).

tiangolo, 2023. FastAPI. https://fastapi.tiangolo.com/ (14 July 2023).