EMBED 2SCALE

Ref. Ares(2025)896357 - 05/02/2025

Earth Observation & Weather Data Federation with AI Embeddings

D1.1:

Analysis of ARD data and APIs

Revision: v 0.1

D1.1: Analysis of ARD data and APIs

Work package	WP1
Task	Task 1.1 & Task 1.2
Due date	31/12/2024
Submission date	05/02/2025
Deliverable lead	PU
Version	0.1
Authors	Matej Batič (SIN), Grega Milčinski (SIN), Regan Koopmans (Planet), Jonas Hurst (WWU)
Reviewers	Devis Peressutti (SIN), Jonas Hurst (WWU)
Abstract	This document provides an analysis of analysis-ready data (ARD) and application programming interfaces (APIs) relevant to the Embed2Scale project. It aims to identify and evaluate existing ARD and APIs that are relevant to the project's objectives, and to assess their suitability for use in the project's pilots. The document also includes recommendations for the further development and use of ARD and APIs in the project.
Keywords	Analysis ready data, data formats, APIs

Document Revision History

Version	Date	Description of change	List of contributors(s)
V0.1	31/01/2025	1st version	Matej Batič (SIN), Grega Milčinski (SIN), Regan Koopmans (Planet), Jonas Hurst (WWU)

© 2024-2027 Embed2Scale

Grant Agreement No: 101131841 | **Topic:** HORIZON-EUSPA-2022-SPACE-02-55
Call: HORIZON-EUSPA-2022-SPACE | **Type of action:** HORIZON-RIA

DISCLAIMER

Funded by the European Union

EUSPA European Union Agency for the Space Programme

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

UK Research and Innovation

Embed2Scale (Earth Observation & Weather Data Federation With AI Embeddings) project is funded by the EU's Horizon Europe program under Grant Agreement number 101131841. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. This work has also received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI) and the UK Research and Innovation (UKRI).

COPYRIGHT NOTICE

© 2024 – 2027 EMBED2SCALE

Project funded by the European Commission in the Horizon Europe Programme		
Nature of the deliverable:	Report	
Dissemination Level		
PU	Public, fully open, e.g. web (Deliverables flagged as public will be automatically published in CORDIS project's page)	■
SEN	Sensitive, limited under the conditions of the Grant Agreement	
Classified R-UE/ EU-R	*EU RESTRICTED under the Commission Decision* No2015/ 444	
Classified C-UE/ EU-C	*EU CONFIDENTIAL under the Commission Decision* No2015/ 444	
Classified S-UE/ EU-S	*EU SECRET under the Commission Decision* No2015/ 444	

* *R: Document, report (excluding the periodic and final reports)*
DEM: Demonstrator, pilot, prototype, plan designs
DEC: Websites, patents filing, press & media actions, videos, etc.
DATA: Data sets, microdata, etc.
DMP: Data management plan
ETHICS: Deliverables related to ethics issues.
SECURITY: Deliverables related to security issues
OTHER: Software, technical diagram, algorithms, models, etc.

TABLE OF CONTENTS

© 2024-2027
Embed2Scale

EXECUTIVE SUMMARY

This document focuses on the challenges and solutions for managing and analyzing the ever-growing volume of Earth Observation (EO) data, particularly in the context of cloud-based analysis. It specifically explores the optimization of cloud data storage strategies by assessing different data formats and data access APIs, aiming to enhance accessibility, interoperability, and scalability for diverse applications, especially within the Copernicus Programme.

The rapid growth of EO data necessitates efficient cloud-based analysis, but the lack of standardized formats and access points hinders data discovery and interoperability. This document evaluates existing cloud-optimized formats like GeoTIFF, COG, and Zarr, and compares commonly used APIs like OGC and STAC, to identify optimal strategies for data storage, transfer, and processing. It also discusses the potential of emerging formats like GeoParquet for storing embeddings from geospatial foundation models. The analysis concludes that COG is currently the most suitable format for analysis-ready raster datasets, and that STAC and OpenEO are the most promising APIs for data access and processing, with potential for future extensions to support embedding generation and handling.

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

API	Application Programming Interface
ARD	Analysis-Ready Data
CDSE	Copernicus Data Space Ecosystem
CEOS	Committee on Earth Observation Satellites
CNG	Cloud-Native Geospatial Forum
COG	Cloud Optimised GeoTiff
EO	Earth Observation
EMS	Earth System Models
ESA	European Space Agency
HTTP	Hypertext Transfer Protocol
OGC	Open Geospatial Consortium
STAC	SpatioTemporal Assets Catalogue

© 2024-2027 Embed2Scale

1 INTRODUCTION

Since the 1960s, EO data has grown exponentially, with initiatives like the Copernicus Programme generating massive daily volumes. Launching advanced satellite constellations further accelerates this trend. This data deluge is essential for critical national and global services. However, data centres are overwhelmed by the sheer volume of information that requires transfer, computation, and analysis. The true potential of EO data lies in its integration with diverse modalities like weather data, surveys, and social media. Additionally, modern Earth System Models (EMS) produce petabytes of daily data, crucial for comprehending our environment and its future.

Beyond volume, the lack of a unified access point for diverse data modalities poses a significant challenge. Each modality often necessitates a unique Application Programming Interface (API), hindering data discovery and increasing system fragility. The complexities and costs associated with transferring, storing, and managing diverse data products can lead to unreliable applications and limited research capabilities. These constraints are particularly concerning in the context of climate change, where large-scale quantitative assessments are vital for informed decision-making.

Embed2Scale tackles these EO and ESM data challenges through innovative neural compression of embeddings. This approach aims to reduce data volume significantly while preserving data utility. Compressed embeddings facilitate decentralized geoinformation consumption, optimize data transfer, and reduce computational demands. The integration of Embed2Scale with the Copernicus Programme promises enhanced accessibility and scalability of geospatial data for diverse applications like ecosystem and infrastructure monitoring.

In this document, the first section assesses the most promising and widely used cloud-optimized formats for raster data from the point of view of typical access patterns against velocity, volume, value, variety and veracity of the key data collections. This should provide us with guidance on which format to choose as a default and how it should be configured to best suit the compression-related tasks. The first section presents the results of Task 1.1 of the Embed2Scale project: Assessment of analysis-ready data.

The second section provides an assessment of several existing best data access API practices used in the EO field, executing a comparison and compatibility analysis of the commonly used interfaces. The rapid growth of Earth Observation (EO) data, reaching hundreds of petabytes, presents a challenge for effective cloud-based analysis. This study explores optimal cloud data storage strategies, focusing on the strengths and weaknesses of different formats. This section presents the results of Task 1.2 of the project: Assessment of data access APIs.

The third section lists the datasets available for use within Embed2Scale, the APIs they can be accessed with, and their data format.

2 ASSESSMENT OF ANALYSIS-READY DATA FORMATS

Analysis Ready Data (ARD) are time-series stacks of overhead imagery that are prepared for a user to analyze without having to pre-process the imagery themselves. Those who don't work with satellite imagery every day likely underestimate the amount of labor involved in preparing imagery for analysis.[1] While CEOS (Committee on Earth Observation Satellites) is probably the highest authority on what ARD is, the definition of CEOS-ARD is not exclusive or prescriptive. The definition of CEOS-ARD reflects the attributes of fundamental measurement products for the majority of global remote sensing users with land imaging applications. They are the minimum level required to support time series analysis and data interoperability.

Table 1: ***Current ARD Product Family Specifications***

Product Family Type	Version	Type	Metadata Spec	Last Updated
Surface Reflectance	5.0.1	Optical	pdf	6 December 2023
Surface Temperature	5.0	Optical	pdf	8 June 2020
Aquatic Reflectance	1.0	Optical	pdf	23 February 2022
Nighttime Lights Surface Radiance	1.0	Optical	pdf	2 October 2022
Combined CEOS-ARD for Synthetic Aperture Radar	1.1	Radar	pdf	23 July 2024

In this report, we will be looking at the ARD datasets from another angle. Definitely, EO archives now host unprecedented amounts of data (enormous *volume*) that are generated daily or even faster (high *velocity*) by a wide range of sensors (data *variety*). The value of information contained in such data quantities is immense, although its *veracity* shall be carefully assessed to emphasize uncertainty and incompleteness during the acquisition process. Our point of view in this report is an assessment of these from the perspective of software solutions and tools available (or being implemented) to handle these 5Vs of Big Data.

Since the decisions to open LandSat and Sentinel data, the typical access pattern (manually looking at the available imagery) has been more and more superseded by the automatic (computerised) analyses. Cloud technologies make access and use Earth Observation data much easier. However, in many cases, data must be reorganized or reformatted to work smoothly with cloud-based tools.

1 https://medium.com/planet-stories/analysis-ready-data-defined-5694f6f48815

This study looks at the advantages and disadvantages of typical data formats used for storing EO data in the cloud.

2.1 COMMON CLOUD-READY ANALYSIS-READY DATA FORMATS

We present and briefly summarize common formats used for analysis-ready raster data, namely:

- Cloud Optimised GeoTiff (COG)
- Zarr
- HDF5
- NetCDF

In this chapter we focus on the suitability of these formats, particularly for the use in case of high data volume and velocity and typical access patterns.

Additionally, we quickly summarise a novel, but promising geospatial data format. While GeoParquet is nominally meant for vector data, it has been shown to work with raster data and more importantly for this project, it is being recognised as a suitable candidate for storing embedding data.

As the two formats (COG and Zarr) take the lead, we also provide a head-to-head comparison with some lessons we have learned working with both.

Cloud Optimised GeoTiff (COG)

GeoTIFF, a public domain metadata standard[2], enables georeferencing information to be integrated into a TIFF file. This allows for the inclusion of crucial spatial data such as map projection, coordinate systems, ellipsoids, and datums. As GeoTIFF is fully compatible with TIFF 6.0, software that cannot interpret the metadata can still open GeoTIFF format files.

TIFF files natively support many compression algorithms, customizable metadata, and color depths. TIFFs already have wide adoption and mature tooling across many industries. The flexibility of the TIFF format allows for many possible representations of physical storage.

Cloud Optimized GeoTIFF (COG) is a GeoTIFF-based standard designed to facilitate the use of GeoTIFFs hosted on HTTP web servers. COG enables users and software to utilize partial data read from within the file, removing the need to download (large file) entirely when only wanting to access a small portion/window of the file.

To achieve this, the COG format employs HTTP range requests that specify a specific data to be retrieved. The metadata layout within the GeoTIFF file allows clients to predict the necessary byte ranges for download. Since COG is a GeoTIFF specialization, COG files are TIFF files.

A COG is thus a regular GeoTIFF file, but with an internal organisation that is optimised for use in cloud-based environments. Its internal organisation enables more efficient read access to a portion of the image, by leveraging the ability of clients issuing HTTP Get range requests[3] to ask for just the portion of a file they need.

[2] https://www.ogc.org/publications/standard/geotiff/
[3] https://en.wikipedia.org/wiki/Byte_serving

© 2024-2027 Embed2Scale

COGs implement tiling and overviews for its internal organisation. **Tiling** divides the image into small tiles and stores them separately. Each tile is a regular TIFF image, with its own header and metadata, that can be requested independently. **Overviews** create reduced resolution versions of the same image. This structure allows range requests to easily find just the part of the whole file that is needed to be processed or visualised. Other significant benefits of COGs include:

- COG can be read and processed by most geospatial processing applications and libraries that also support standard GeoTIFF files e.g. GDAL.
- COG structure is optimised for efficient data fetching from cloud storage and on-the-fly processing.
- COG supports two (spatial) dimensions, but the temporal dimension can be included by creating multi-band COG or by combining several COGs in a collection (e.g., a STAC collection)

Many publicly available analysis-ready datasets have already been published with this format for its wide support and ease of use, most notably the LandSat dataset published on AWS. It is generally preferred that analysis ready GeoTIFFs also conform to being COGs since this does not incur any sizable cost and greatly improves read performance for the average case.

Core specifications for COG

- Meta-data of full resolution imagery (index) at the beginning of the file, which makes it possible to programmatically download the file index first, then fetch GET requests for the relevant data.
- (Optional) metadata of overviews – as above, for overview layers.
- (Optional) tile content of overviews (last to first).
- Tile content of full resolution data.
- Internal tiling size – unspecified default tile size.
- Compression methods allowed – unspecified default/recommended method.

Recommended options for COG creation

- **Internal tiling size** – Block size of between 256 – 2048 pixels; 1024x1024 pixel is recommended – small tiles may result in larger number of requests while large tiles will result in unnecessary data download.
- **Compression methods –** DEFLATE and LZW for lossless compression and JPG where quality loss is OK. DEFLATE is more efficient than LZW but may cause compatibility problems with some software packages.
- **Resampling –** e.g. NEAREST, AVERAGE, BILINEAR, MODE (for categorical data); a method that ensures appropriate interpretation of the content is recommended.
- **Bands -** No requirement on single-band or multi-band files; *Pixel interleaving* for multi-band raster images.

Zarr

Zarr is a newer open standard for storing multi-dimensional array data. It specifies a protocol and format, and addresses the cloud readiness (e.g. random data access) by dividing data into subsets, referred to as chunks. Zarr is not an image format but a general serialization strategy for n-dimensional arrays. This approach distinguishes it from GeoTIFFs and enables storing vector and time-series data together with raster information. Zarr allows for easy interoperability with common

Python libraries such as xarray and numpy, which are increasingly popular tools in geospatial analysis.

While the format addresses the issues of cloud readiness, the novelty of this format may pose a risk for storing large datasets. The latest version of the specification (v3) was accepted on May 15th, 2023 . Reference implementations are under active development. Future modifications to the specification could impose costly migrations. Nevertheless, nascent Zarr datasets demonstrate read performance comparable to more mature formats such as HDF5.

Zarr is conceptually similar in scope to NetCDF/ HDF5 in terms of its ability to capture and express metadata and data. However, the nature of NetCDF/HDF formats makes it nearly impossible to extract a single variable or piece of data or metadata without reading the whole file thus creating performance issues for operations in the cloud environment. Zarr chunks, when organised according to the access pattern, alleviate this issue.

Main characteristics of Zarr

- Zarr is cloud-optimised due to its chunking structure that allows independent retrieval of chunks, thus improving parallelism and granularity in access.
- Zarr can be stored in any storage system that provides a key/value interface (e.g., Object storage)
- Metadata is encoded in JSON format – metadata is kept separate from data.

Zarr files are now widely being used for cloud-optimized storage of multidimensional data. When storing EO data it is common to adhere to NetCDF/CF metadata conventions and longitude, latitude, and time dimensions.

The advantages of Zarr format

- Zarr files support arbitrary dimensions. This means that an entire (hyper)cube of data can be stored in a single Zarr file.
- Zarr files allow metadata to be stored in a single JSON object, and each chunk to be a separate object. This allows unlimited file sizes and parallel writes and reads.
- Zarr supports organizing arrays into hierarchies via groups by default, and expands to separate keys in object storage.
- Zarr files support arbitrary chunking, along each of the dimensions. This allows the chunking structure to adapt to a specific use case. Some general examples of optimal chunking scenarios include:
  - For datasets with regular updates, time chunking should be low (i.e. 1)
  - For spatial analysis of large areas, spatial chunking needs to be large (i.e. 1000)
  - For temporal analysis over a point, spatial chunking needs to be small (i.e. 1), temporal chunking needs to be large (i.e. 1000).

Multi-dimensional formats are typically designed for use in a broader scope besides geospatial arrays, therefore there are some known limitations to using Zarr in a spatial context. However, there are efforts to provide extensions for geospatial data e.g. GeoZarr specifications[4]. Some of the known limitations are listed below:

- There is no standard way of encoding CRS information in Zarr.
- There is no multi-resolution support which makes the format difficult to work with at large scales (e.g. continental).

4 https://github.com/zarr-developers/geozarr-spec

© 2024-2027 Embed2Scale

- Zarr standard supports non-equidistant arrays, which works well with time dimension (which in EO is generally non-equidistant) but might become complex with spatial coordinates.

A fairly descriptive overview of Zarr format for cloud-optimised access of EO data is provided by AWS[5].

HDF5/NetCDF4

HDF5 is a one of the Hierarchical Data Formats, developed to store and organize large amounts of data. HDF file formats (HDF4 and HDF5) can store many disparate types of data together in a hierarchical fashion. It is comparable to Zarr in ease of use and interoperability with Python. HDF5 is, however, less suitable for cloud native use. While there are some examples on making HDF5 format cloud ready[6], there is little or no support for range requests over a network and no standard to facilitate uptake in common/widely used geospatial tools. This alone makes HDF5 (and NetCDF) quite inferior to COGs and Zarr for the EO data, and it will thus not be further considered.

GeoParquet

Parquet, a columnar data store within the Hadoop ecosystem, is natively compatible with Amazon S3 Select queries and Apache Spark. Its ability to capture rich metadata on a per-column basis and support a wide range of data types, including custom ones, makes it a versatile choice.

GeoParquet is an extension of Parquet supporting geospatial data. While GeoParquet is primarily designed for traditional geospatial vector data, its flexible schema can be adapted to store embeddings from geospatial foundation models, which is of particular interest for the Embed2Scale project. By treating embeddings as additional attributes associated with geospatial features, they can be stored alongside traditional geometry data within GeoParquet files. This approach leverages GeoParquet's efficient storage and interoperability, allowing for seamless integration of embeddings into existing geospatial workflows.

However, it's important to note that GeoParquet is still under development, and its support for complex data types like embeddings may evolve. Therefore, while GeoParquet offers a promising avenue for storing embeddings from geospatial foundation models, it's advisable to stay updated with the latest developments in the GeoParquet specification and community practices.

More information on GeoParquet can be found at GeoParquet home page[7] on in guides published by Cloud-Native Geospatial Forum (CNG)[8]. We will revisit GeoParquet as file format later in the project, once we have embeddings to store.

Summary

It appears that using COGs is currently the most attractive storage strategy for analysis-ready raster datasets (see also section 4 on EO datasets in Embed2Scale). COGs strike a unique balance between maturity, performance and interoperability with modern tools.

The community around cloud native data (storage, access, …) is very much alive and thriving, and we are a part of it as well. The recently established CNG exists to unlock the potential of geospatial data through adoption of commoditized cloud technologies and open collaboration on data formats

5 https://aws.amazon.com/blogs/publicsector/decrease-geospatial-query-latency-minutes-seconds-using-zarr-amazon-s3/

6 https://guide.cloudnativegeo.org/cloud-optimized-netcdf4-hdf5

7 geoparquet.org

8 https://guide.cloudnativegeo.org/geoparquet/

© 2024-2027 Embed2Scale

– an approach that can be described as cloud-native. We are happy to say Sinergise Solutions (through Planet Labs) is a member of that initiative as well. We continue building our knowledge of latest developments, and share back the lessons learned from operating services like Sentinel-Hub, and managing large datasets coming from our own fleet of satellites.

As a summary, table 2 provides the overview of ARD file formats from the perspective of cloud support, standardisation, and other information relevant for assessing the 5V's of big data. Some data is summarized from [3].

Table 2: **Overview of ARD file formats**

	COG	**HDF5**	**(Geo)Parquet**	**Zarr**	**GeoTiff**	**NetCDF-4**
Cloud native	Yes	Potentially	Yes	Yes	No	No
Standard format?	COG is a TIFF, TIFF is ISO-standard				EOSDIS, OGC (TIFF is ISO standard)	EOSDIS, OGC
Storage Technology	Chunked numerical data, possibly pyramid overviews	multiple objects in COS: JSON storage metadata and HDF5 dataset data chunks. JSON metadata established hierarchical organisation of other stored HDF5 objects. Additionally, HSDS server: to use HDF5 library APIs for reading files	Columnar data storage, Rich columnar metadata, Rows chunked together	Array-based hierarchical data store Hierarchy storage as "logical paths" -> single conceptual file may be several files on disk/COS	arrays of numeric types	HDF5: hierarchical, array-based format
Data Type support	Raster data (projected grids)	multidimensional arrays, hierarchical data	Columnar data, possibly with vector geometries, in situ data	multidimensional arrays, hierarchical data	Raster data (projected grids)	multidimensional arrays, hierarchical data
Data integrity	not inherently in file format, but could be achieved through metadata fields fine-grained integrity checking by e.g. running gdalinfo with --checksum	capture fine-grained integrity information -> check hdf5 file format spec with h5check command line tool	no support for internal file integrity check BUT it could be achieved through metadata fields no support for fine-grained integrity checking	multi-file key/value model is conductive to finer-grained integrity checking not inherently in file format, but could be achieved through metadata fields	not inherently in file format, but could be achieved through metadata fields fine-grained integrity checking by e.g. running gdalinfo with --checksum	*(only if accessed through hdf5)* capture fine-grained integrity information -> check hdf5 file format spec with h5check command line tool

Self-descriptive format?	Yes provides application-specific metadata providing app-specific metadata may require placing it in a sidecar PAM file if it is large	Yes provides application-specific metadata	Yes provides application-specific metadata	Yes provides application-specific metadata	Yes provides application-specific metadata providing app-specific metadata may require placing it in a sidecar PAM file if it is large	Yes provides application-specific metadata
Open specification	Yes test script to validate standard conformance is provided	Yes	Yes clear diagrams and descriptions to aid in understanding of format	Yes	Yes	Yes
Ability to comply with metadata conventions	support of named layers, no widely-adopted way to express unit information for individual layer Large metadata must be offloaded into sidecar file		adequate fields to store metadata on variables and units but no accepted convention for naming of those fields	fields that allow capturing CF-compliant metadata	support of named layers, no widely-adopted way to express unit information for individual layer Large metadata must be offloaded into sidecar file	CF conventions were built to work with the format ample additional metadata fields relevant to earth science domain

2.2 COG VS ZARR

Comparing COG vs Zarr is particularly important. ESA is considering updating the Sentinel data archives to Zarr. The following table summarizes the two format differences. Later, we present a list of our concerns with switching to Zarr.

Table 3: **Comparison of Zarr and COG formats**

	Zarr	COG
Clear Specification	Yes (but still evolving)	Yes
Random remote access	Yes	Yes
Update friendly	Yes (new data reflects in addition of new chunks and change of meta-data files)	No (new data typically reflects in adding new COG files)
Easy to share	No (one has to zip the whole container as partial files cannot be read well)	Yes
Support for non-geospatial data	Yes	No real advantage to store non geo data

Overviews (enable analysis at lower resolution)	Not yet part of the core standard, there are best practices31	In core specification
CRS encoding	Not part of core specification, being added in the process of OGC workgroup. Currently used via CF conventions	In core specification
Tools available	Growing list (GDAL >=3.4, Python, Java, JavaScript, Julia, C++, NetCDF C)	All libraries reading GeoTIFF
More than 3 dimensions	Yes	No
User defined tiling (chunks)	Yes (along all dimensions)	Yes (only for spatial dimensions)

Based on the comparisons shown in Table 2, Zarr and COG file formats do exhibit differences. Zarr format is more of a Catalog storing both data and metadata and is commonly used for dense or sparse multi-dimensional data while COG format is better suited for 2-D raster data with a regular grid, e.g., EO data.

From our experiences running Sentinel Hub service, the main drawbacks of Zarr are the following:

- There are no pyramidal overview layers defined in the Zarr format, which is a blocker when it comes to efficient visualisation of large datasets.
- The size of chunking blocks is extremely important. For instance, the chunk size of 1830x1830 px that is being considered, is not a blocker per se, but makes data access for visualisation use-case inefficient. Users would typically visualise 512x512px (recommended) or 256x256px tiles. With 1830px-size chunks and 512px-size tiles users will have to read 12-times more data than needed in the optimal scenario. In case the tile is on the border of four chunks, the factor goes to 50. Based on our past experiences, we recommend 1024x1024px tiles as they provide best performance in terms of number of GET requests and volume of data read.
- Zarr format is predominantly a datacube format. In the proposed approach, however, each Sentinel product is represented as its own Zarr file, not benefiting from the “datacube approach”. The approach is therefore not using most of the benefits of the Zarr format.
- One of the benefits of Zarr is allowing different compressions to be used for different arrays of data. This would reduce the size, but make data handling more complex.
- With possible lack of open-source utilities to efficiently work with the Zarr format, users will probably resort to COG-ified versions of the data repositories, ignoring the main distribution altogether.
- Quicklooks are needed to be shown in web browsers, i.e., in the “Search tab” of the Copernicus Browser. If they are not present, every data distributor will have to generate them, resulting in lack of consistency of previews.

Specifically for Sentinel-2, we point to the following drawbacks of using Zarr:

- SAFE product contains 109 files, that number would explode to 1875 objects for Zarr. If overview layers are added to the product, the number of objects will grow further. If chunking is reduced to the recommended size of 1024x1024px, the number of objects will grow further. While this is not an issue with hyperscalers (AWS, GCP, …), it might cause troubles within the CDSE environment (on both CloudFerro and OTC infrastructure). There is a limit on the number of objects per bucket to work optimally, so the number of object storage buckets required for efficient access on CDSE would grow by orders of magnitude.
- Will the format be optimised for the time-series or for visualisation? How will the chunking reflect that?

Funded by the European Union
© 2024-2027 Embed2Scale

Our final remark is that Zarr would probably work well with Sentinel-2 quarterly mosaics and other temporally regular data. Mosaics could be stored as one big hyper cube (split per UTM zones, due to non-uniform CRS).

2.3 ACCESS PATTERNS - LARGE SCALE DATA PROCESSING

As part of the efforts to allow (very) large scale EO data processing using Sentinel Hub services, we have implemented the Batch API (now already at version 2). Within Global Earth Monitor project (Grant Agreement 101004112), we have created an ARD datacube, Sentinel-2 L2A 120m mosaic[9]. To process it, we had to access a whole year worth of Sentinel-2 data. During the processing, we nearly overloaded the access to public dataset on AWS, performing about 400,000 requests per second, as can be seen in the image below.

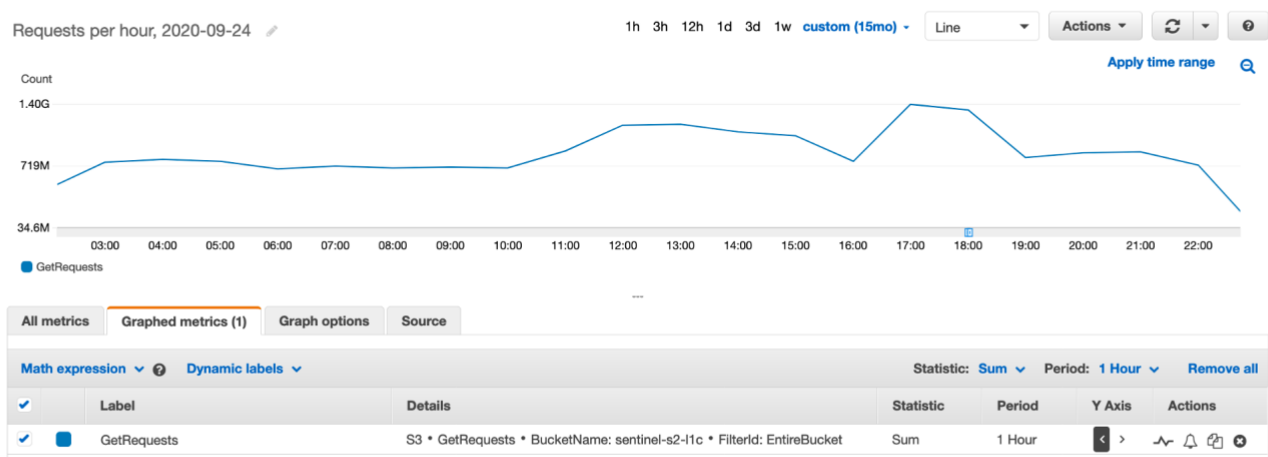

Figure 1: Number of requests (per hour), processing 1 year worth of Sentinel-2 data

The main point we are trying to raise here is actually the suitability of cloud-native format, in this case JPEG2000, which has the same properties in terms of cloud-nativeness as COGs. Typical access patterns in terms of data retrieval are significantly lower for normal use. The fact that properly set up cloud infrastructure can deal with enormous, unplanned data access spikes, makes it possible to replicate such structure on infrastructure as is CDSE, provided the datasets are in suitable format.

9 https://collections.sentinel-hub.com/sentinel-s2-l2a-mosaic-120/

3 ASSESSMENT OF DATA ACCESS APIS

After comparing different analysis-ready data formats for storing and transferring the generated embeddings, we will now focus our attention on how the EO data can be discovered, accessed and processed. This is essential, as work package teams must be able to access EO data in a unified and streamlined manner.

As data federation requires interoperability between machines, we are only taking Application Processing Interfaces (APIs) into account, and we are neglecting all software intended to be interacted by users through a visual frontend (e.g. data browsers).

Instead of designing a novel Application Processing Interface (API) for accessing EO data, we are going to focus on preexisting, well-established API standards. Therefore, we are going to compare existing API standards used within the geospatial domain in general and the field of earth observation specifically. After, we are going to present a feasibility study on how these existing API standards could be extended to foster compatibility and interoperability (section 3.1), before we draw a conclusion to fix a decision regarding which API we are going to build our further work on (section 3.2).

3.1 COMPARISON OF DATA ACCESS APIS

Within the and EO domain, a wide variety of APIs are commonly used to catalog, discover and access geospatial data. In this section, we are going to compare them with respect to the requirements within the Embed2Scale project.

We define our requirements for an API that we choose as follows:

- **Well-established:**
  It is important to us that the API we choose is well-known in the Geospatial domain and more specifically in the Earth Observation community. This will make data access familiar.
- **Modern (RESTful):**
  We want to build our API to use state-of-the-art paradigms. We therefore require the API to follow a RESTful, resource-oriented approach, which at the time of writing is the state of the art.
- **Management of metadata:**
  Besides providing access to the EO data itself, the API must be able to expose metadata on the data it serves.
- **Queryable:**
  The API must allow querying the data by metadata to allow consumers to only return the data matching specific criteria.
- **Extensibility:**
  The novelty about embeddings stemming from general-purpose foundation models is that it encodes image content semantically: Within this embedding vector space, image content of the same “type” is closer together than dissimilar content. Therefore, neighborhood searching around an embedding will allow to find similar images. As this is a novel concept, we must assume that this form of querying is not reflected in any existing API standard. Therefore, the standard we will choose must be extensible to perform such an operation to allow content-aware data querying and retrieval.

OGC Web Services

The OGC Web Services are a set of standards for accessing spatial resources on the web. Due to their widespread use, they are well-known within the geospatial community and understood with many client applications available.

The OGC Web Services are a product of the time when they were developed. Therefore, they rely on paradigms and technologies which were considered modern at the time they were introduced but are seen outdated as of today: Instead of a RESTful design pattern with individual URL endpoints for each resource, the Web Services expose a single endpoint, and interaction with the service is done through URL query parameters. Further, the OGC Web Services use XML to transmit information from the Server to the clients, resulting in overhead in the transmission due to redundancy in encoding data. Therefore, they are slowly superseded by OGC APIs (*c.f.* below) and we are not going to elaborate on the OGC Web Services here further.

OGC APIs

The OGC has acknowledged the limitations of the OGC Web services and created a new set of standards to supersede the antiquated Web Services: They are called OGC APIs and follow a RESTful approach, where each resource has their own URL endpoint and use JSON for communication between client and server.

Due to their novelty, they are not as widely adopted yet as the legacy OGC Web Services. Nonetheless, open-source implementations exist for both an OGC API compliant server (e.g. Geoserver, pyGeoAPI) and client (e.g. QGIS, camptocamp/ogc-client). As members of the OGC are research, government and the industry leaders, the standards it defines are well-regarded, and we expect OGC APIs to be used as widely as OGC Web Services in the future.

It must be noted that at the time of writing, some while OGC API standards are finalized, some others are still in draft state and may be subject to changes. In the following, we are going to go over the OGC API standards that are relevant to our use case and discuss their benefits and drawbacks.

OGC API - Coverage

This API specification is designed specifically to provide access to so-called coverages. As per definition in ISO-19123, a coverage is a "*feature that acts as a function to return values from its range for any direct position within its spatial, temporal or spatiotemporal domain"*. Given that the earth observation imagery we are working with (namely Sentinel 1 and Sentinel 2) is available globally, we could produce global coverages of these imagery data. However, this would come at a loss of information when reprojecting, warping and aggregating the images.

At the time of writing, this API standard is still a draft and not an OGC-approved standard yet.

OGC API - Features

This specification is designed to give access to spatial features with properties, i.e. spatial geometries with associated non-spatial attributes. Features belonging together thematically are grouped together in a collection.

While this does not director correspond to our needs for providing access to imagery data, individual features could be treated as a proxy for the real image: The geometry can be used to indicate the footprint of the image, metadata as well as a Link to download the image can be given in the properties of the feature.

Part 1 of the OGC API - Features standard defines how a collection can be queried temporally (i.e. return all features that fall within a certain time period, if a timestamp is given) and spatially using a bounding box (i.e. only return Feature that fall into the bounding box). More precise filtering based on properties is defined in Part 3 of the standard.

While the idea outlined above to enable access to imagery data through OGC API - Features is feasible in general, it is worth pointing out that we'd be using it in a way it was never designed to do: to serve as a data catalog. Other API standards exist that were created to serve as exactly this, so we should rather rely on them.

- ### OGC API - Processes

This specification is designed to invoke geospatial processing on a remote server in a standardized way. Each process can reveal a set of input parameters through which the user is able to control certain aspects of the process.

While this is not an API to access data, it could still be useful to us to invoke the on-demand generation of embeddings . The foundation model that shall be used to generate the embeddings, and the spatial and temporal bounds for which the embeddings should be generated, could be defined using process input parameters.

One limitation of this standard is that users are limited to executing the processes that are exposed by the API, they cannot define their own, which makes developing downstream applications for embeddings in the cloud environment challenging. The API could provide some processes as building blocks to develop downstream tasks, but that would require the API standard to allow for chaining of processes. This is addressed in Part 3 of the OGC API - Processes standard, which at the time of writing is in draft state.

- ### OGC API - Records

The purpose of this specification is to catalog spatial and spatio-temporal resources, provide appropriate metadata for these resources, and make them discoverable by ordering them thematically into collections. We could use this API specification to create a catalog of images. This specification also defines a standardized way to query saved records by their metadata.

However, this specification is still in draft state and therefore could be subject to change.

- ### OGC API - Environmental Data Retrieval

This specification defines standardized access to environmental data, which is characterized by being highly dimensional (e.g. lat, lon, elevation level, time, parameter, ensemble run).

This API provides a wide range of geometric querying options such as querying by Point geometry or by Bounding box, but also by cube, by radius or by 3D-trajectory in space. While the former definitely apply to our use case, it is questionable whether we really need all these sampling methods for our use cases. More likely, having to implement such functionality is a waste of time for us.

Funded by the European Union
EUSPA
© 2024-2027 Embed2Scale

SpatioTemporal Asset Catalog

In contrast to the formerly mentioned API standards, the SpatioTemporal Asset Catalog (STAC) is not a standard developed by a governing organization (e.g. OGC), but is rather developed, maintained and governed by the community that uses it. As the name suggests, it is used to catalog spatio-temporal assets and make them discoverable.

In its most basic form, a STAC catalog can be a collection of statically inter-linked JSON files. Coupling with a STAC-API allows for a more dynamic approach, including querying the STAC by metadata. It is considered a flexible schema, as STAC only describes a core schema which can be customized through application-specific schema extensions. Similarly, the STAC-API can be extended, which in our case could allow for content-aware embedding querying.

Within the earth observation community, STAC and STAC-API are used widely to catalog earth observation products. To exemplify this, the following is a short list of notable STAC deployments:

- Copernicus Data Space Ecosystem: https://catalogue.dataspace.copernicus.eu/stac
- Element84 EarthSearch (AWS): https://earth-search.aws.element84.com/v1/
- Microsoft Planetary Computer: https://planetarycomputer.microsoft.com/api/stac/v1
- Synergize SentinelHub: https://sh.dataspace.copernicus.eu/api/v1/catalog/1.0.0/

OpenEO

An increase in EO satellites orbiting earth has led to data catalogs containing EO data products on petabyte-scale. Performing analysis on these catalogs with large spatial and temporal extent necessitates offloading the computations into cloud environments. Preferably, the same cloud environment is used, which already stores the data catalogs to perform the computations close to the data.

To facilitate the interaction between the backend offering EO computation capabilities in the cloud and clients, OpenEO API has emerged as an open standard, enabling a decoupling of technologies used on the client and on the server. This API is well-known in the field of earth observation. For example, it is part of the new Copernicus Data Space Ecosystem (CDSE): https://openeo.dataspace.copernicus.eu/openeo/

OpenEO constructs EO image analysis workflows by chaining so-called OpenEO-Processes together, each of which performs a single operation on the data, to work in unison to generate the desired analysis results. OpenEO already has capability to query a dataset by its spatial and temporal extent in the load_collection and load_stac processes. Querying by embedding similarity could be added by creating a new process or adding additional parameters to existing processes for this task.

3.2 COMPATIBILITY OF APIS

Compatibility OGC APIs <-> STAC-API

Both OGC API (Features and Records) and STAC-API organize the data they serve thematically as items into collections. For both, the endpoint URLs to query collections and items within a collection are the same. Collections in both are represented as supersets of GeoJSON FeatureCollections.

Funded by the European Union
EUSPA
© 2024-2027
Embed2Scale

Likewise, both represent items as supersets of GeoJSON Features. Both do not conflict with each other, meaning that APIs could be built that conform to both standards.

However OGC API - Records and STAC-Items both have properties to store the same information but are named differently, leading to redundancy when the API response should conform to both standards. For example, both contain a property to store the date of creation of a Record (OGC API Records) or of an asset (STAC): In OGC API - Records, this property is called created, in STAC, it is called datetime.

Compatibility OGC API - Processes <-> OpenEO

Both OGC API - Processes and OpenEO are built to invoke processing on a remote server. A push for compatibility between the two can be seen in the early draft of Part 3 of OGC API - Processes specification, as it defines a conformance class to invoke an OpenEO process graph, but only limited information is available about this conformance class at the point of writing. However, it must be noted that the two may never be fully compatible with each other (Link).

Compatibility STAC <-> OpenEO

Both STAC and OpenEO are standards that are driven by its community of users. OpenEO tightly integrates STAC-API, as it implements STAC API's /collections and /collections/collection-id endpoints, which can be loaded into the processing workflow using the load_collection process. External STAC-Collections can be loaded using the load_stac process.

The results of an OpenEO workflow are written a valid STAC collection, enabling them to be loaded into a new OpenEO process graph.

3.3 FEASIBILITY STUDY ON POSSIBLE EXTENSIONS

STAC

STAC's core data model consists of only a minimal core schema, which only defines a limited set of properties. Other properties needed can be added using STAC extensions to form a metadata schema that fits the needs of the application.

To provide access to embeddings with STAC, proper care must be taken to provide appropriate metadata on the embeddings. A STAC extension is necessary to follow the STAC standard in this regard. To query for embedding similarity, an extension to the STAC-API standard could enable such operations. As STAC is community-driven, anyone who is interested in it can create and use such extensions.

OpenEO

We can use OpenEO to access EO data and process it on location within the HPC cluster without needing to download the imagery first. At its current state, however OpenEO is lacking crucial functionality needed for the processing of EO data regarding embeddings:

A process is needed that can execute EO foundation models to compute embeddings from source imagery. Then, the save_result process must be altered to support writing embeddings in the chosen ARD data format from Section 2. Further, processes capable of offering advanced machine learning

© 2024-2027 Embed2Scale

capabilities must be created to allow for training and applying application-specific embedding decoder heads to be able to perform downstream tasks.

3.4 INTERMEDIARY CONCLUSION

Based on the analysis within this section of the report, we fix the decision to build our solution upon STAC and OpenEO as they fulfill all five criteria lined out above. Both are sufficiently developed in their current state to give access to EO imagery, yet need development with respect to generating and handling embeddings later on in the project.

Funded by the European Union
© 2024-2027 Embed2Scale

EO Datasets available within Embed2Scale

The following table lists Earth Observation (EO) datasets and derived products available through the Sentinel Hub platform. It includes a short description of the data, spatial and temporal availability, API access information, and other relevant notes.

Table 4: **EO datasets in Embed2Scale**

Dataset	Description	Availability		STAC	API	Link	File format
		Spatial	Temporal				
Sentinel-2 L1C	L1C processed Sentinel-2 data	Global	November 2015 onwards	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	Sentinel-2 L1C	Jpeg2000
Sentinel-2 L2A	L2A processed Sentinel-2 data	Global	November 2016 onwards	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	Sentinel-2 L2A	Jpeg2000
Sentinel-1 GRD	Sentinel 1 Ground Range Detected Imagery	Global	October 2014 onwards	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	Sentinel-1 GRD	COG
CNES Land Cover	The CNES Land Cover Map (Occupation des Sols, OSO) produces land classification for Metropolitan France at 10 m spatial resolution based on Sentinel-2 L2A data within the Theia Land Cover CES framework.	Metropolitan France	2016 - 2021, yearly product	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	CNES Land Cover Map - Sentinel Hub Collections	COG
Impact Observatory LULC	Impact Observatory, Microsoft, and Esri produce the 10m Annual Land Use Land Cover (LULC) map collaboratively. The data collection is derived from ESA Sentinel-2 imagery at 10m resolution globally using Impact Observatory's state of the art deep learning AI land classification model which is trained by billions of human-labeled image pixels. There are 9 LULC classes generated by the algorithm, including Built, Crops, Trees, Water, Rangeland, Flooded Vegetation, Snow/Ice, Bare Ground, and Clouds.	Global	2017-2021, yearly product	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	10m Annual Land Use Land Cover (9-class) - Sentinel Hub Collections	COG
Sentinel S2 L2A Mosaic 120m	Sentinel-2 L2A 120m mosaic is a derived product, which contains best pixel values for 10-daily periods, modelled by removing the cloudy pixels and then performing interpolation among remaining values. As clouds can be missed and as there are some parts of the world which have lengthy cloudy periods, clouds might be remaining in some parts.	Global	2020	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	Sentinel-2 L2A 120m Mosaic	COG
ESA World Cover	The European Space Agency (ESA) WorldCover is a global land cover map with 11 different land cover classes produced at 10m resolution based on a combination of both Sentinel-1 and Sentinel-2 data. In areas where Sentinel-2 images are covered by clouds for an extended period of time, Sentinel-1 data then provides complimentary information on the structural characteristics of the observed land cover. Therefore, the combination of Sentinel-1 and Sentinel-2 data makes it possible to update the land cover map almost in real time. WorldCover Map was first produced for 2020 using v100 of the algorithm and later for 2021 with v200 algorithm. Due to the different algorithm versions, it should be noted that changes between WorldCover map 2020 and WorldCover map 2021 are a result of both actual changes in land cover and in the used algorithm. The WorldCover map is provided as part of the 5th Earth Observation Envelope Programme (EOEP-5). It provides global coverage of valuable information for applications such as biodiversity, food security, carbon assessment and climate modelling.	Global	2020, 2021, yearly product	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	ESA WorldCover - Sentinel Hub Collections	COG
Sentinel 3 OLCI	Sentinel 3 imagery captured by OLCI sensor	Ocean and Land	April 2016 onwards	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	Sentinel-3 OLCI L1B	NetCDF
Sentinel 3 SLSTR	Sentinel 3 imagery captured by SLSTR sensor	Ocean and Land	April 2016 onwards	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	Sentinel-3 SLSTR L1B	NetCDF
Sentinel 5 Precursor	Sentinel 5 Precursor imagery captured by TROPOMI sensor	Ocean and Land	April 2018 onwards	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	Sentinel-5P L2	NetCDF
Copernicus DEM	The Copernicus DEM is a Digital Surface Model (DSM) which represents the bare-Earth surface and all above ground natural and built features. It is based on WorldDEM™ DSM that is derived from TanDEM-X and is infilled on a local basis with the following DEMs: ASTER, SRTM90, SRTM30, SRTM30plus, GMTED2010, TerraSAR-X Radargrammetric DEM, ALOS World 3D-30m. Copernicus Programme provides Copernicus DEM in 3 different instances: COP-DEM EEA-10, COP-DEM GLO-30 and COP-DEM GLO-90 where COP-DEM GLO-90" tiles and most of the "COP-DEM GLO-30 " tiles are available worldwide with free license. Sentinel Hub provides two instances named COPERNICUS_90 which uses "COP-DEM GLO-90" and COPERNICUS_30 which uses "COP-DEM GLO-30 public" and "COP-DEM GLO-90" in areas where "COP-DEM GLO-30 Public" tiles are not yet released to the public by Copernicus Programme. Copernicus DEM provides elevation data and can also be used to orthorectify satellite imagery (e.g. Sentinel 1).	European and Worldwide coverage (https://spacedata.copernicus.eu/collections/copernicus-digital-elevation-model)	static product (two releases 2019 & 2020) for data acquired in (2011-2015), updated updated as soon as new release is available publicly	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	Copernicus DEM - Sentinel Hub Collections	COG

CORINE Land Cover Accounting Layers	The CORINE Land Cover (CLC) Accounting Layers are CLC status layers modified for the purpose of consistent statistical analysis in the land cover change accounting system at EEA. The modification combines CLC status and change layers in the 100m raster to create a homogeneous quality time series of CLC / CLC-change layers for accounting purposes. The CLC inventory consists of 44 land cover and land use classes derived from a series of satellite missions since it was first established. There are altogether 5 mapping inventories implemented since 1986, producing five status layers (CLC1990, CLC2000, CLC2006, CLC2012, CLC2018), 4 CLC-Change layers for the corresponding periods (1990-2000, 2000-2006, 2006-2012, 2012-2018) and 4 CLC Accounting layers for 2000, 2006, 2012 and 2018. More information can be found in Annex 5 of the [Product User manual](https://land.copernicus.eu/user-corner/technical-library/clc-product-user-manual).	EEA39 region	2000, 2006, 2012, 2018, updated every 6 years.	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	CORINE Land Cover Accounting Layers - Sentinel Hub Collections	COG
CORINE Land Cover	The CORINE Land Cover (CLC) inventory consists of 44 land cover and land use classes derived from a series of satellite missions since it was first established.	Pan-European, French overseas regions and departments (DOMs)	1990, 2000, 2006, 2012, 2018, updated every 6 years.	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	CORINE Land Cover - Sentinel Hub Collections	COG
Global Human Settlement Layer - GHS-BUILT-S2	The Global Human Settlement (GHS) framework produces global maps of built-up, population density and settlements to monitor human presence on Earth over time. The Global Human Settlement Layer GHS-BUILT-S2 is a global map of built-up areas (expressed as probabilities) at 10 m spatial resolution. It was derived from a Sentinel-2 global image composite for the reference year 2018 using Convolutional Neural Networks.	Global coverage with longitude from 180°W to 180°E and latitude from 72°N to 56°S	Reference year 2018, updated irregular	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	Global Human Settlement Layer - GHS-BUILT-S2 - Sentinel Hub Collections	COG
Global Land Cover	Global Land Cover products at 100 m resolution are delivered annually by the [global component of the Copernicus Land Service](https://land.copernicus.eu/global/products/lc). The most recent collection 3 (version 3.0.1) of 100 m Land Cover products for the years 2015 - 2019 were generated from the PROBA-V 100 m and 300 m satellite observations and several other ancillary datasets with global coverage. These Land Cover products provide a main discrete land cover classification map according to the UN-FAO Land Cover Classification System [LCCS](https://land.copernicus.eu/global/products/lc). Additional continuous fractional layers for all basic land cover classes which give the percentage of a 100 m pixel that is filled with a specific land cover class, are also included in the Land Cover products to provide more detailed information on each land cover class.	Global coverage of land surface area with longitude from 180°W to 180°E and latitude from 78.25°N to 60°S.	2015 - ongoing, updated annually	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	Global Land Cover - Sentinel Hub Collections	COG
Global Surface Water	The European Commission's Joint Research Centre developed the Global Surface Water dataset within the framework of the Copernicus Programme. It is derived from Landsat imagery and shows different aspects of the spatio-temporal distribution of surface water between 1984 and 2021 at the global scale.	Global coverage from longitude 170°E to 180°W and latitude 80°N to 50°S.	1984 - 2019, 1984 - 2020, 1984 - 2021, updated annual revision	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	Global Surface Water - Sentinel Hub Collections	COG
High-Resolution Snow & Ice Monitoring; Fractional Snow Cover	The High-Resolution Snow & Ice Monitoring service (HR-S&I) is part of the Copernicus Land Monitoring Service (CLMS). The snow aspect of the service provides products measuring Snow cover (FSC, FSTOC, FSCOG, GFSC), Snow state conditions (WDS, SWS) and persistent snow area (PSA). There are also ice products that measure ice cover (RLIE) and aggregated river and lake ice extent (ARLIE). More information about the data can be found [here](https://land.copernicus.eu/pan-european/biophysical-parameters/high-resolution-snow-and-ice-monitoring) and The snow products user manual is located [here](https://land.copernicus.eu/user-corner/technical-library/hrsi-snow-pum). The Fractional Snow Cover (FSC) product provides the snow fraction at the Top Of Canopy (FSCTOC) and On Ground (FSCOG). NDSI is also provided as part of this service.	Europe (Lat; 34N 66N Lon; -26W 44E)	September 2016 - ongoing, updated 5 days, in near real-time, according s2 revisit	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	High-Resolution Snow & Ice Monitoring; Fractional Snow Cover - Sentinel Hub Collections	COG

High-Resolution Snow & Ice Monitoring; Daily cumulative Gap-filled Fractional Snow Cover	The High-Resolution Snow & Ice Monitoring service (HR-S&I) is part of the Copernicus Land Monitoring Service (CLMS). The snow aspect of the service provides products measuring Snow cover (FSC, FSTOC, FSCOG, GFSC), Snow state conditions (WDS, SWS) and persistent snow area (PSA). There are also ice products that measure ice cover (RLIE) and aggregated river and lake ice extent (ARLIE). More information about the data can be found [here](https://land.copernicus.eu/pan-european/biophysical-parameters/high-resolution-snow-and-ice-monitoring) and The snow products user manual is located [here](https://land.copernicus.eu/user-corner/technical-library/hrsi-snow-pum). The daily cumulative Gap-filled Fractional Snow Cover (GFSC) product is a more complete version of the FSC product; gap-filled both at a spatial and temporal scale.	Europe (Lat; 34N 66N Lon; -26W 44E)	September 2016 - ongoing, updated 5-7 days, in near real-time, at each release of fsc, wds or sws	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	High-Resolution Snow & Ice Monitoring; Daily cumulative Gap-filled Fractional Snow Cover - Sentinel Hub Collections	COG
High-Resolution Snow & Ice Monitoring; Persistent Snow Area	The High-Resolution Snow & Ice Monitoring service (HR-S&I) is part of the Copernicus Land Monitoring Service (CLMS). The snow aspect of the service provides products measuring Snow cover (FSC, FSTOC, FSCOG, GFSC), Snow state conditions (WDS, SWS) and persistent snow area (PSA). There are also ice products that measure ice cover (RLIE) and aggregated river and lake ice extent (ARLIE). More information about the data can be found [here](https://land.copernicus.eu/pan-european/biophysical-parameters/high-resolution-snow-and-ice-monitoring) and The snow products user manual is located [here](https://land.copernicus.eu/user-corner/technical-library/hrsi-snow-pum). The Persistent Snow Area (PSA) product is generated annually from FSC products and provides the extent of persistent snow cover for that year (the area where snow is present throughout the hydrological year). The PSA corresponds to pixels with a snow probability greater than 0.95, i.e. where snow was observed in at least 95% of the observations (under clear-sky conditions). For optimisation reasons, the retrieval methodology assumes that the minimal snow cover is reached across Europe between 1 May and 30 September; thus the algorithm performs the aggregation only over this period.	Europe (Lat; 34N 66N Lon; -26W 44E)	May 2017 - ongoing, updated annually (released in October after end of the hydrological year)	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	High-Resolution Snow & Ice Monitoring; Persistent Snow Area - Sentinel Hub Collections	COG
High-Resolution Snow & Ice Monitoring; River and Lake Ice Extent - Sentinel 1	The River and Lake Ice Extent - Sentinel 1 (RLIE S1) product is part of the Copernicus Land Monitoring Service (CLMS), pan-European High Resolution Snow and Ice Monitoring [(HR-S&I)](https://land.copernicus.eu/pan-european/biophysical-parameters/high-resolution-snow-and-ice-monitoring) Ice product suite.The River and Lake Ice Extent (RLIE) products measure the presence of ice in rivers and lakes described by the EU-HYDRO network database, in particular the area of snow-free or snow-covered ice. The RLIE S1 product is released in near real time for the entire EEA39 zone based on the revisit time of the Sentinel 1 synthetic aperture radar constellation, and it is generated at a spatial resolution of 20 m x 20 m. Because it is time critical for several applications, every effort is made to provide near real time ice products between 6 and 12 hours after the data sensing time. More information about the RLIE S1 product methodology and workflow is outlined in the [product user manual](https://land.copernicus.eu/user-corner/technical-library/hrsi-ice-pum). A detailed explanation about the accuracy of the product can be found in the [quality assessment report](https://land.copernicus.eu/user-corner/technical-library/hrsi-ice-qar).	Europe (Lat; 34N 66N Lon; -26W 44E)	September 1, 2016 - ongoing, updated nrt, between 6 and 12 hours after the data sensing time.	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	High-Resolution Snow & Ice Monitoring; River and Lake Ice Extent - Sentinel 1	COG
High-Resolution Snow & Ice Monitoring; River and Lake Ice Extent S1+S2	The River and Lake Ice Extent S1+S2 (RLIE S1+S2) product is part of the Copernicus Land Monitoring Service (CLMS), pan-European High Resolution Snow and Ice Monitoring [(HR-S&I)](https://land.copernicus.eu/pan-european/biophysical-parameters/high-resolution-snow-and-ice-monitoring) Ice product suite.The River and Lake Ice Extent (RLIE) products measure the presence of ice in rivers and lakes described by the EU-HYDRO network database, in particular the area of snow-free or snow-covered ice. The RLIE S1+S2 product is computed from the combination of RLIE S1 and RLIE S2 products.The RLIE S2 product from a given day is gap-filled using the RLIE S1 information from the same day. Because the RLIE S1+S2 is only produced when RLIE S2 and RLIE S1 products are acquired on the same day, it is delivered in delayed-time for the entire EEA39 zone at a spatial resolution of 20 m x 20 m. More information about the RLIE S1+S2 product methodology and workflow is outlined in the [product user manual](https://land.copernicus.eu/user-corner/technical-library/hrsi-ice-pum). A detailed explanation about the accuracy of the product can be found in the [quality assessment report](https://land.copernicus.eu/user-corner/technical-library/hrsi-ice-qar).	Europe (Lat; 34N 66N Lon; -26W 44E)	September 1, 2016 - ongoing, updated delayed time, when s1 and s2 are acquired on the same day	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	High-Resolution Snow & Ice Monitoring; River and Lake Ice Extent S1+S2 - Sentinel Hub Collections	COG

High-Resolution Snow & Ice Monitoring; River and Lake Ice Extent - Sentinel 2	The River and Lake Ice Extent - Sentinel 2 (RLIE S2) product is part of the Copernicus Land Monitoring Service (CLMS), pan-European High Resolution Snow and Ice Monitoring [(HR-S&I)](https://land.copernicus.eu/pan-european/biophysical-parameters/high-resolution-snow-and-ice-monitoring) Ice product suite.The River and Lake Ice Extent (RLIE) products measure the presence of ice in rivers and lakes described by the EU-HYDRO network database, in particular the area of snow-free or snow-covered ice. The RLIE S2 product is released in near real time for the entire EEA39 zone based on the revisit time of the Sentinel 2 constellation, and it is generated at a spatial resolution of 20 m x 20 m. Because it is time critical for several applications, every effort is made to provide near real time ice products between 6 and 12 hours after the data sensing time. More information about the RLIE S2 product methodology and workflow is outlined in the [product user manual](https://land.copernicus.eu/user-corner/technical-library/hrsi-ice-pum). A detailed explanation about the accuracy of the product can be found in the [quality assessment report](https://land.copernicus.eu/user-corner/technical-library/hrsi-ice-qar).	Europe (Lat; 34N 66N Lon; -26W 44E)	September 1, 2016 - ongoing, updated nrt, between 6 and 12 hours after the data sensing time.	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	High-Resolution Snow & Ice Monitoring; River and Lake Ice Extent - Sentinel 2	COG
High-Resolution Snow & Ice Monitoring; SAR Wet Snow in high mountains	The High-Resolution Snow & Ice Monitoring service (HR-S&I) is part of the Copernicus Land Monitoring Service (CLMS). The snow aspect of the service provides products measuring Snow cover (FSC, FSTOC, FSCOG, GFSC), Snow state conditions (WDS, SWS) and persistent snow area (PSA). There are also ice products that measure ice cover (RLIE) and aggregated river and lake ice extent (ARLIE). More information about the data can be found [here](https://land.copernicus.eu/pan-european/biophysical-parameters/high-resolution-snow-and-ice-monitoring) and The snow products user manual is located [here](https://land.copernicus.eu/user-corner/technical-library/hrsi-snow-pum). The SAR Wet Snow (SWS) product provides information on the wet snow extent in high-mountain areas.	Europe (Lat; 34N 66N Lon; -26W 44E)	September 2016 - ongoing, updated in near real-time, according s1 revisit	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	High-Resolution Snow & Ice Monitoring; SAR Wet Snow in high mountains - Sentinel Hub Collections	COG
High-Resolution Snow & Ice Monitoring; Wet / Dry Snow - Snow state classification	The High-Resolution Snow & Ice Monitoring service (HR-S&I) is part of the Copernicus Land Monitoring Service (CLMS). The snow aspect of the service provides products measuring Snow cover (FSC, FSTOC, FSCOG, GFSC), Snow state conditions (WDS, SWS) and persistent snow area (PSA). There are also ice products that measure ice cover (RLIE) and aggregated river and lake ice extent (ARLIE). More information about the data can be found [here](https://land.copernicus.eu/pan-european/biophysical-parameters/high-resolution-snow-and-ice-monitoring) and The snow products user manual is located [here](https://land.copernicus.eu/user-corner/technical-library/hrsi-snow-pum). The Wet/Dry Snow (WDS) product differentiates the snow state conditions within the snow mask defined by the FSCTOC information.	Europe (Lat; 34N 66N Lon; -26W 44E)	September 2016 - ongoing, updated in near real-time, according s1 revisit and fsc release	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	High-Resolution Snow & Ice Monitoring; Wet / Dry Snow - Snow state classification - Sentinel Hub Collections	COG
Sea Ice Index	[Sea Ice Index](https://nsidc.org/data/G02135/versions/3) products offer information about ice cover and its trends in the Arctic and Antarctica. In this collection, sea ice extent and concentration maps are provided at a resolution of 25 km on a daily basis. The products are computed using images from the Special Sensor Microwave Imager (SSM/I) and Special Sensor Microwave Imager/Sounder (SSMIS) instruments on Defense Meteorological Satellite Program (DMSP) satellites. Sea ice concentrations are generated from [brightness temperature data](https://nsidc.org/data/nsidc-0051). Sea ice extent is computed as areas covered by ice with a concentration above 15 percent. This dataset is for demonstration purposes only and is not maintained.	Longitude from 180°W to 180°E and latitude from 39.23°N to 90°N and 30.98°S to 90°S.	2017 - May 2021, updated none. for demonstration purposes only.	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	Sea Ice Index - Sentinel Hub Collections	COG
Seasonal Trajectories, 10-daily	Seasonal Trajectories (ST) product is a filtered time series of [Plant Phenology Index (PPI)](https://www.nateko.lu.se/research/remote-sensing-and-earth-observation/lund-earth-observation-research-group/vegetation-phenology) provided yearly on a 10-daily basis at 10m resolution . It is part of the Copernicus Land Monitoring Service (CLMS), pan-European High Resolution Vegetation Phenology and Productivity [(HR-VPP)](https://land.copernicus.eu/pan-european/biophysical-parameters/high-resolution-vegetation-phenology-and-productivity) product suite. The Seasonal Trajectories PPI is derived through fitting a smoothing and gap filling function to the yearly time-series raw PPI values generated from Sentinel-2 satellite observations. In addition to the PPI band, a QFLAG band indicating the quality of the smoothing process is included. The Seasonal Trajectories provide the vegetation status for each pixel on a regular 10-day time step from January 1 2017 onwards over the EEA39 region.	Europe (Lat; 26N 72N Lon; -25W 45E)	January 1, 2017 - ongoing, updated annually	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	Seasonal Trajectories, 10-daily - Sentinel Hub Collections	COG

	More information about ST product is outlined in the [product user manual](https://land.copernicus.eu/user-corner/technical-library/product-user-manual-of-seasonal-trajectories/)						
Vegetation Indices, Daily	Vegetation Indices (VI) product is part of the Copernicus Land Monitoring Service (CLMS), pan-European High Resolution Vegetation Phenology and Productivity [(HR-VPP)](https://land.copernicus.eu/pan-european/biophysical-parameters/high-resolution-vegetation-phenology-and-productivity) product suite. The product is comprised of 4 raw Vegetation Indices; (1) Normalized Difference Vegetation Index (NDVI), (2) Leaf Area Index (LAI), (3) Fraction of Absorbed Photosynthetically Active Radiation (FAPAR) and (4) Plant Phenology Index (PPI) generated near real-time (NRT) from Sentinel-2 satellite observations. The raw Vegetation Indices are provided on a daily basis at 10m resolution from October 2016 onwards. Therefore VI products are available over the EEA39 region for the respective observation day based on the 5-day revisit period of Sentinel-2. The VIs are accompanied by a related QFLAG2 band that flags clouds, shadows, snow, open water and other areas where the VI retrieval is less reliable. More information about raw vegetation indices is outlined in the [product user manual](https://land.copernicus.eu/user-corner/technical-library/product-user-manual-of-vegetation-indices/)	Europe (Lat; 26N 72N Lon; -25W 45E)	October 1, 2016 - Ongoing, updated daily	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	https://collections.sentinel-hub.com/vegetation-indices/	COG
Vegetation Phenology and Productivity Parameters Season 1, Yearly	Vegetation Phenology and Productivity Parameters (VPP) product is part of the Copernicus Land Monitoring Service (CLMS), pan-European High Resolution Vegetation Phenology and Productivity [(HR-VPP)](https://land.copernicus.eu/pan-european/biophysical-parameters/high-resolution-vegetation-phenology-and-productivity) product suite. The VPP product consists of 13 parameters that describe specific stages of the seasonal vegetation growth cycle. These parameters are extracted from Seasonal Trajectories of the [Plant Phenology Index (PPI)](https://www.nateko.lu.se/research/remote-sensing-and-earth-observation/lund-earth-observation-research-group/vegetation-phenology) derived from Sentinel-2 satellite observations at 10m resolution. Since growing seasons can traverse years, VPP parameters are provided for a maximum of two growing seasons per year. The parameters include (1) start of season (date, PPI value and slope), (2) end of season (date, PPI value and slope), (3)length of season, (4) minimum of season, (4) peak of the season (date and PPI value), (5) amplitude, (6) small integrated value and (7) large integrated value. VPP parameters are generated over the EEA39 region on a yearly frequency from January 1 2017 onwards. Among other applications, the high-resolution phenology data provides a detailed assessment of the impacts of human or climate change on the ecosystem through monitoring of vegetation responses to disturbances, e.g. droughts, storms, insect infestations, and to human influence from global to local levels. More information about VPP product is outlined in the [product user manual](https://land.copernicus.eu/user-corner/technical-library/product-user-manual-of-seasonal-trajectories/)	Europe (Lat; 26N 72N Lon; -25W 45E)	January 1, 2017 - ongoing, updated annually	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	Vegetation Phenology and Productivity Parameters Season 1, Yearly - Sentinel Hub Collections	COG

Vegetation Phenology and Productivity Parameters Season 2, Yearly	Vegetation Phenology and Productivity Parameters (VPP) product is part of the Copernicus Land Monitoring Service (CLMS), pan-European High Resolution Vegetation Phenology and Productivity [(HR-VPP)](https://land.copernicus.eu/pan-european/biophysical-parameters/high-resolution-vegetation-phenology-and-productivity) product suite. The VPP product consists of 13 parameters that describe specific stages of the seasonal vegetation growth cycle. These parameters are extracted from Seasonal Trajectories of the [Plant Phenology Index (PPI)](https://www.nateko.lu.se/research/remote-sensing-and-earth-observation/lund-earth-observation-research-group/vegetation-phenology) derived from Sentinel-2 satellite observations at 10m resolution. Since growing seasons can traverse years, VPP parameters are provided for a maximum of two growing seasons per year. The parameters include (1) start of season (date, PPI value and slope), (2) end of season (date, PPI value and slope), (3)length of season, (4) minimum of season, (4) peak of the season (date and PPI value), (5) amplitude, (6) small integrated value and (7) large integrated value. VPP parameters are generated over the EEA39 region on a yearly frequency from January 1 2017 onwards. Among other applications, the high-resolution phenology data has a high potential to give a detailed assessment of the impacts of human or climate change on the ecosystem through monitoring of vegetation responses to disturbances, e.g. droughts, storms, insect infestations, and to human influence from global to local levels. More information about VPP product is outlined in the [product user manual](https://land.copernicus.eu/user-corner/technical-library/product-user-manual-of-seasonal-trajectories/)	Europe (Lat; 26N 72N Lon; -25W 45E)	January 1, 2017 - ongoing, updated annually	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	Vegetation Phenology and Productivity Parameters Season 2, Yearly - Sentinel Hub Collections	COG
Water Bodies	The Global Water Bodies product shows the surface extent covered by inland water on a permanent, seasonal or occasional basis. The product available here is the Water Bodies 100m Version 1 collection which is derived from Sentinel-2 level 1C data, starting from October 2020 after the end of the PROBA-V mission and is delivered as a monthly composite product at 100m resolution.The Water Bodies product contain one main Water Bodies detection layer (WB) and one Quality layer (QUAL) that provides information on the seasonal dynamics of the detected water bodies. Water Bodies detection layer (WB) shows water bodies detected using the Modified Normalized Difference Water Index (MNDWI) derived from Sentinel-2 Level 1C data. The Quality layer (QUAL) is generated from water body occurrence statistics computed from previous monthly Water Bodies products.The occurrence statistics is ranked from low occurrence to permanent occurrence.	Global coverage from longitude -180°E to +180°W and latitude +80°N to -60S . Depending on the month, some high latitude areas are not covered by Sentinel-2 satellites.	October 2020 - ongoing, updated monthly	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	Water Bodies - Sentinel Hub Collections	COG
Harmonized Landsat Sentinel	Harmonized Landsat Sentinel is a NASA initiative to produce a Virtual Constellation of surface reflectance (SR) data from the Operational Land Imager (OLI) and Multi-Spectral Instrument (MSI) aboard the Landsat 8-9 and Sentinel-2 remote sensing satellites, respectively. The combined measurement enables global observations of the land every 2–3 days. Input products are Landsat 8-9 Collection 2 Level 1 top-of-atmosphere reflectance and Sentinel-2 L1C top-of-atmosphere reflectance, which NASA radiometrically harmonizes to the maximum extent, resamples to common 30-meter resolution, and grids using the Sentinel-2 Military Grid Reference System (MGRS) UTM grid. Because of this, the products are different from Landsat 8-9 Collection 2 Level 2 surface reflectance and Sentinel-2 L2A surface reflectance.	Global land	April 2013 - ongoing., updated 2-3 days	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	Harmonized Landsat Sentinel	COG
Landsat 1-5 MSS L1	The Landsat Multispectral Scanner System (MSS) sensors were carried onboard Landsats 1 to 5. It provides 4 spectral bands. See [USGS EROS Archive](https://www.usgs.gov/centers/eros/science/usgs-eros-archive-landsat-archives-landsat-1-5-multispectral-scanner-collection?qt-science_center_objects=0#qt-science_center_objects) for more information. MSS Level-1 data provides Top of Atmosphere Reflectance products for the period from July 1972 to October 1992 and from June 2012 to January 2013.	Global land	Landsat 1 from July 1972 to January 1978 Landsat 2 from January 1975 to February 1982 Landsat 3 from March 1978 to March 1983 Landsat 4 from July 1982 to December 1993 Landsat 5 from 1984 to October 1992, and from June 2012 to January 2013 , updated archived data	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	Landsat 1-5 MSS L1 - Sentinel Hub Collections	COG

Landsat 4-5 TM L1	The Landsat Thematic Mapper (TM) sensor was carried onboard Landsats 4 and 5. TM collected data in 7 spectral bands; from the blue, green, red, near-infrared, mid-infrared (2) and thermal infrared portions of the electromagnetic spectrum. Visit [USGS EROS Archive - Landsat Archives - Landsat 4-5 Thematic Mapper Collection 2 Level-1 Data](https://www.usgs.gov/centers/eros/science/usgs-eros-archive-landsat-archives-landsat-4-5-thematic-mapper-collection-2?qt-science_center_objects=0#qt-science_center_objects) webpage for more information.	Global Land	Landsat 4 from July 1982 to December 1993 Landsat 5 from March 1984 to May 2012 , updated archived data	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	Landsat 4-5 TM L1 - Sentinel Hub Public Collections	COG
Landsat 4-5 TM L2	The Landsat Thematic Mapper (TM) sensor was carried onboard Landsats 4 and 5. TM collected data in 7 spectral bands; from the blue, green, red, near-infrared, mid-infrared(2) and thermal infrared portions of the electromagnetic spectrum. L2 data include Surface Reflectance and Surface Temperature scene-based products. Visit [USGS EROS Archive - Landsat Archives - Landsat 4-5 TM Collection 2 Level-2 Science Products](https://www.usgs.gov/centers/eros/science/usgs-eros-archive-landsat-archives-landsat-4-5-tm-collection-2-level-2-science?qt-science_center_objects=0#qt-science_center_objects) for more information.	Global Land	Landsat 4 from July 1982 to December 1993 Landsat 5 from March 1984 to May 2012 , updated archived data	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	Landsat 4-5 TM L2 - Sentinel Hub Collections	COG
Landsat 7 ETM+ L1	The Landsat 7 Enhanced Thematic Mapper (ETM+) sensor is carried onboard Landsat 7. ETM+ provides 7 spectral bands and 1 thermal band. See [USGS EROS Archive](https://www.usgs.gov/centers/eros/science/usgs-eros-archive-landsat-archives-landsat-7-enhanced-thematic-mapper-plus?qt-science_center_objects=0#qt-science_center_objects) for more information. Landsat 7 Level-1 data provides Top of Atmosphere Reflectance and Top of the Atmosphere Brightness Temperature products. Level 1 data have been available since April 1999. All scenes collected since May 30, 2003 have data gaps due to the Scan Line Corrector (SLC) failure.	Global land	April 1999 - ongoing, updated landsat data are added regularly,as soon as they are available in usgs aws cloud	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	Landsat 7 ETM+ L1 - Sentinel Hub Collections	COG
Landsat 7 ETM+ L2	The Landsat 7 Enhanced Thematic Mapper (ETM+) sensor is carried onboard Landsat 7. ETM+ provides 7 spectral bands and 1 thermal band. See [USGS EROS Archive](https://www.usgs.gov/centers/eros/science/usgs-eros-archive-landsat-archives-landsat-7-etm-plus-collection-2-level-2?qt-science_center_objects=0#qt-science_center_objects) for more information. Landsat 7 level-2 data provides atmospherically corrected Surface Reflectance and Surface Brightness Temperature products. Level-2 data have been available since April 1999.	Global land	April 1999 - ongoing, updated landsat data are added regularly,as soon as they are available in usgs aws cloud	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	Landsat 7 ETM+ L2 - Sentinel Hub Public Collections	COG
Landsat 8-9 L1	[Landsat 8-9 Level 1](https://www.usgs.gov/centers/eros/science/usgs-eros-archive-landsat-archives-landsat-8-9-operational-land-imager-and) collection includes both [Landsat 8](https://www.usgs.gov/landsat-missions/landsat-8?qt-science_support_page_related_con=0#qt-science_support_page_related_con) and the most recently launched [Landsat 9](https://www.usgs.gov/landsat-missions/landsat-9) satellites (provided by NASA/USGS), both carrying the Operational Land Imager (OLI) and the Thermal Infrared Sensor (TIRS) instruments, with 9 optical and 2 thermal bands. These two sensors provide seasonal coverage of the global landmass. [Landsat 8-9 Level 1 data](https://www.usgs.gov/core-science-systems/nli/landsat/landsat-collection-2-level-1-data?qt-science_support_page_related_con=1#qt-science_support_page_related_con) from the most recently released [collection 2](https://www.usgs.gov/faqs/what-updates-are-being-made-landsat-collection-2?qt-news_science_products=0#qt-news_science_products), provides Top of Atmosphere Reflectance and Top of the Atmosphere Brightness Temperature products. Level 1 data are available since February 2013 for Landsat 8 and since January 2022 for Landsat 9.	Land ([more info](https://landsat.gsfc.nasa.gov/landsat-8/mission-details))	February 2013 - ongoing, updated landsat data are added regularly,as soon as they are available in usgs aws cloud	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	Landsat 8-9 L1 - Sentinel Hub Collections	COG

Landsat 8-9 L2	[Landsat 8-9 Level 2](https://www.usgs.gov/centers/eros/science/usgs-eros-archive-landsat-archives-landsat-8-9-operational-land-imager-and) collection includes both [Landsat 8](https://www.usgs.gov/landsat-missions/landsat-8?qt-science_support_page_related_con=0#qt-science_support_page_related_con) and the most recently launched [Landsat 9](https://www.usgs.gov/landsat-missions/landsat-9) satellites (provided by NASA/USGS), both carrying the Operational Land Imager (OLI) and the Thermal Infrared Sensor (TIRS) instruments, with 9 optical and 2 thermal bands. These two sensors provide seasonal coverage of the global landmass. [Landsat 8-9 Level 2 data](https://www.usgs.gov/core-science-systems/nli/landsat/landsat-collection-2-level-2-science-products) from the most recently released [collection 2](https://www.usgs.gov/faqs/what-updates-are-being-made-landsat-collection-2?qt-news_science_products=0#qt-news_science_products), provides atmospherically corrected Surface Reflectance and Surface Brightness Temperature products generated from Collection 2 Level-1 scenes that have been processed to Tier 1 or Tier 2. Collection 2 level 2 data are available since February 2013 for Landsat 8 and since January 2022 for Landsat 9 and new data are continuously added usually within 24 hours after Level 1 scenes are available.	Land ([more info](https://landsat.gsfc.nasa.gov/landsat-8/mission-details))	February 2013 - ongoing, updated landsat data are added regularly,as soon as they are available in usgs aws cloud	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	Landsat 8-9 L2 - Sentinel Hub Collections	
Mapzen DEM	Mapzen DEM is based on [Mapzen's terrain tiles](https://github.com/tilezen/joerd/tree/master/docs) that provides global DEM and bathymetry data. Mapzen terrain tiles is a composite of elevation data of varying resolutions from [multiple open data sources](https://github.com/tilezen/joerd/blob/master/docs/data-sources.md) including [SRTM](https://www.usgs.gov/centers/eros/science/usgs-eros-archive-digital-elevation-shuttle-radar-topography-mission-srtm-1-arc?qt-science_center_objects=0#qt-science_center_objects), [ETOPO1](https://www.ngdc.noaa.gov/mgg/global/global.html), and other higher resolution sources for some parts of the world. Mapzen DEM provides bare-earth terrain heights and can also be used for the orthorectification of satellite imagery (e.g. Sentinel 1).	Ocean and Land ([more info](https://github.com/tilezen/joerd/blob/master/docs/data-sources.md#what-is-sourced-at-what-zooms))	static [current version was built in 2017](https://github.com/tilezen/joerd/blob/master/docs/data-sources.md#data-updates), updated [depends on updates by mapzen](https://github.com/tilezen/joerd/blob/master/docs/data-sources.md#data-updates)	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	Mapzen DEM - Sentinel Hub Collections	COG
MODIS MCD43A4.006	[MODIS](https://modis.gsfc.nasa.gov/about/) (Moderate Resolution Imaging Spectroradiometer) is the main instrument operating on both NASA's [Terra](https://terra.nasa.gov/) and [Aqua](https://aqua.nasa.gov/) satellites. It acquires images of the earth in 36 bands within the visible and the infrared regions of the spectrum at low to medium spatial resolutions. MODIS is designed to provide at least daily observations of land, oceans and lower atmosphere that contribute to local or global scale land or water applications. There are several products derived from MODIS which include land, atmosphere, cryosphere and ocean products. MCD43A4 Nadir BRDF-Adjusted Reflectance (NBAR) product Version 6 is a MODIS land product available in Sentinel Hub. It is computed daily for each of MODIS bands 1 - 7 and it contains the most representative pixels from the 16 day period temporarily weighted to the ninth day. MCD43A4.006 has been available since February 2000 on a daily basis but with a 8 days delay. Sentinel Hub no longer has the latest MODIS data because the underlying data provider has ceased operations. Data from February 24, 2000 to February 10, 2023 remains accessible.	MCD43A4.006 product covers Land (more info)	February 24, 2000 - February 10, 2023, updated archived data. Sentinel hub no longer has the latest modis data.	STAC compliant API (CatalogAPI)	SentinelHub API, OGC, limited openEO	MODIS MCD43A4.006 - Sentinel Hub Collections	COG

© 2024-2027 Embed2Scale

4 CONCLUSIONS

The analysis of ARD formats and data access APIs highlighted the complexities and challenges of managing and analyzing the ever-growing volume of EO data, especially in cloud-based environments. The lack of standardized formats and access points hinders data discovery and interoperability, making it difficult to fully leverage the potential of EO data.

This document has explored optimal cloud data storage strategies by assessing different data formats and data access APIs, aiming to enhance accessibility, interoperability, and scalability for diverse applications, especially within the Copernicus Programme. The key findings include:

- **COG** is currently the most suitable format for analysis-ready raster datasets due to its maturity, performance, and interoperability with modern tools.
- **STAC** and **OpenEO** are the most promising APIs for data access and processing. They have potential for future extensions to support embedding generation and handling.

Based on these findings, it is recommended to adopt COG as the default format for storing analysis-ready raster datasets and to utilize STAC and OpenEO APIs for data access and processing. Further research and development efforts should focus on extending STAC and OpenEO to support embedding generation and handling, and on exploring the potential of emerging formats like GeoParquet for storing embeddings from geospatial foundation models.

In the future work, we will update this deliverable with findings that are more pertinent to the project: on the data formats and structures suitable for accessing the embeddings data.

© 2024-2027
Embed2Scale

REFERENCES

[1] R. Wilkinson, M. M. Mleczko, R. J. W. Brewin, K. J. Gaston, M. Mueller, J. D. Shutler, X. Yan, and K. Anderson, “Environmental Impacts of Earth Observation Data in the Constellation and Cloud Computing Era,” in Science of The Total Environment, vol. 909, p. 168584, 2024, https://doi.org/10.1016/j.scitotenv.2023.168584

[2] C. Gomes and T. Brunschwiler, “Neural Embedding Compression for Efficient Multi-Task Earth Observation Modelling,” in Proceedings of the 2024 IEEE International Geoscience and Remote Sensing Symposium, Athens, Greece, 2024, pp. 8268–8273, https://doi.org/10.1109/IGARSS53475.2024.10642535

[3] C. Durbin, Patrick Quinn, Dana Shum, “Task 51 - Cloud-Optimized Format Study”, EED2-TP-125, Revision 01, Technical Paper, January 2020, Prepared Under Contract NNG15HZ39C, https://ntrs.nasa.gov/api/citations/20200001178/downloads/20200001178.pdf

Funded by the European Union
EUSPA
© 2024-2027 Embed2Scale