ESA EOPF 101

Table of contents

Welcome

This online book is the go-to resource to learn everything about the EOPF Sample Service by ESA.

To be continued...

Part I

About EOPF

1 Introduction to the Earth Observation Processor Framework

What is the Earth Observation Processor Framework?

The Earth Observation Processor Framework (EOPF) is an initiative led by the European Space Agency (ESA) designed to modernise and harmonise data from the Copernicus Sentinel Missions.

With the upcoming Copernicus Expansion missions in 2028, the amount of data will significantly increase. EOPF is ESA's solution to organise Sentinel data in a way that works seamlessly with modern cloud technology. This will make it easier to find, access, and process the information you need. The new approach provides user-friendly access, simplifies maintenance, and helps keep costs down, guaranteeing reliable access to Sentinel data in the long run.

The Sentinel-1, Sentinel-2, and Sentinel-3 missions are the first to be updated with this new system.

The EOPF Data Model

The EOPF data model has been defined by following a set of principles:

- **Open standards:** Following common and community approved data standards ensure sustainability and user uptake.
- **Interoperability:** Harmonised with a clear and organised structure that describes the data itself.
- **Cloud optimisation:** Designed for efficient access and handling in cloud environments.
- **Conversion flexibility:** Providing tools to adjust the data for different applications.

To achieve this, there are the following key development areas under EOPF:

EOPF product structure

The EOPF prodcut structure defines how information about the data is structured, stored, and connected. It contains the key details we need to know about a dataset, like when it was collected or its level of detail (the attributes). As part of the EOPF, ESA is actively working on a common data structure for Sentinel data products, with the aim to define a common meta-model that can be used nacross all Sentinel and other EO missions. This approach ensures that data from several missions is consistent. Thus, it is easier to combine, relate, and evaluate data from different missions.

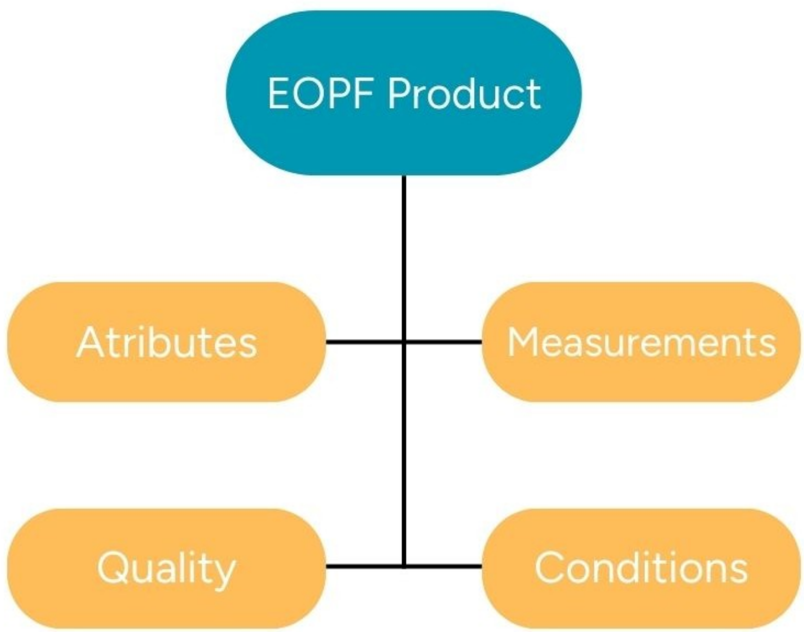

Figure 1.1: EOPF Product structure

The EOPF product structure contains of the following key parts:

- **Measurements:** The actual sensor readings (like how much light is reflected or the temperature), at different levels of detail.
- **Quality indicators:** Details that help understand how reliable the measurements are.
- **Conditions:** Information about the environment or technical aspects when the data was collected.
- **Attributes:** Global metadata, such as when it was acquired and the sensor's orbit.

Metadata Structure

Metadata can be seen as data about data, which provides essential information to understand and use data effectively. Inside the Copernicus ecosystem, it allows us to uniquely identify each Sentinel product by detailing its technical and geographic coverage information. To make sure the new data system works well and avoids redundancy, the EOPF metadata structure will be organised into:

- Discovery Metadata: Following the metadata structure defined by the SpatioTemporal Asset Catalogue (STAC), which helps to keep things consistent across different missions.
- Processing History Metadata: keeping a record of how the data has been processed.
- Other Metadata: Information like the status of the sensor and details about the satellite's orbit.

> **Note**
>
> *"Learn more about EOPF and STAC here."*

Encoding Structure

An encoding structure can be seen as the specific method used to package and store data and its associated metadata in a digital format. Building on the consistent data structure and clear metadata, the new storage system must be capable of handling various aspects of current Sentinel data (such as manifest files and tile structures from the SAFE format) while remaining fully compatible with cloud environments.

This cloud-friendly architecture will enable instant access to data, efficient processing of massive amounts of data, and seamless integration with other datasets, allowing you to work with data from multiple missions more effectively.

Currently, there are different ways to store geospatial data in the cloud, such as: - Zarr - Cloud-Optimised GeoTIFF (COG) - Hierarchical Data Format (HDF5) with Cloud Adaptations - Apache Parquet (with GeoParquet Extension)

ESA chose `.zarr` as encoding format as it allows for efficient handling of data, like selecting specific areas and combining different datasets easily (for example, by dividing the data into smaller chunks). Learn more about the EOPF Zarr format here.

> **Note**
>
> Learn more about cloud-optimised geospatial data formats in the Cloud-Optimised Geospatial Data Formats Guide

Processor Development

The way Sentinel data is processed is being updated to take advantage of modern cloud computing. This will make the processing faster and more efficient, while ensuring the scientific quality and accuracy of the Sentinel data remains the same.

> **Note**
>
> To learn more about the EOPF processor framework, visit https://eopf.copernicus.eu/eopf/

What's next?

In the following chapters, we will introduce the details of the new EOPF data structure, explain the reasons behind ESA's choice of data storage methods, and help the community become comfortable with the new EOPF data format.

2 About Cloud Optimised Formats

Why do we need to cloud-optimise geospatial data formats?

The volume of EO data has grown exponentially in recent years. The Copernicus programme alone generates ~16TB daily from the Sentinel missions. Traditional file formats, like SAFE (where each file can be hundreds of megabytes), are optimised for efficient archiving and distributing data. This means that we often download the data from an entire overpass, even if we only need to access a small part of it, for example, if we want to do an analysis of the area of a single city over a decade.

With growing data volumes, this becomes a challenge. To picture the different nature of challenges we come across, let us compare a traditional local workflow with a cloud-based workflow:

- **Traditional local workflow**: When working locally, we download much more data than we need, and we are constrained by the compute and storage capacity of the local system. However, one big advantage is that data and compute are close together, meaning that there is not much delay.
- **Cloud-based workflow**: A cloud environment offers solutions to limitations local workflows have. A cloud environment offers limitless storage and compute capacity. On the contrary, data storage, compute, and you the destination are far apart. There is an additional time for data to travel between the storage location, processing resources and us. This time is referred to as **data latency**.

> ℹ Note
>
> **Data latency** refers to the time it takes for data to be transmitted or processed from cloud storage to your computer. In local workflows, data latency is minimal, whereas in cloud-based workflows, data latency needs to be optimised.

Local workflows are similar to placing an order at the nearby pizzeria. It is quick since the 'data' (pizza) is easily accessible, but we can only choose from what they have on hand and their menu. The local alternatives limit our options. On the other hand, cloud-based workflow offers almost limitless choices and access to a wide range of speciality ingredients or distinctive styles. This makes it similar to being able to order a pizza from any pizzeria on the globe.

While we might have more options to choose from, the time between order and delivery can become a challenge.

The overall goal with cloud-based workflows is to minimise **data latency** as much as possible. This is why traditional data formats need to be cloud-optimised.

2.1 Characteristics of cloud-optimised formats

Cloud-optimised formats are optimised to minimise data latency. By allowing for an efficient retrieval of smaller, specific chunks of information rather than downloading an entire file. Accessing a smaller data subset also reduces the costs associated with data transfer and data processing.

Cloud-optimised geospatial data formats have the following characteristics:

- Data is accessible over an HTTP protocol.
- Read-Oriented, as it supports partial and parallel reads.
- Data is organised in internal groupings (such as chunks, tiles, shards) for efficient subsetting, distributed processing and data access in memory.
- Metadata can be accessed in one read.

> **Note**
>
> When accessing data over the internet (e.g., cloud storage), latency is high compared to local storage, so it is preferable to fetch lots of data in fewer reads.

Cloud-Optimised Geospatial Raster Formats

Initiatives like the Cloud Native Geospatial Forum (CNG) help develop ways to use geospatial data efficiently in the cloud, using standard cloud technologies and encouraging open collaboration on data formats. Two of the most important formats for storing and accessing geospatial data efficiently are Cloud-Optimised GeoTIFFs (COGs) and Zarr.

Cloud-optimised GeoTIFF (COG)

COGs are like snapshots of raster data (such as satellite images or elevation maps). This widely used format improves the standard GeoTIFF format by:

- Organising data into **tiles**: Dividing the data into smaller, manageable squares (like 512x512 pixels).

- Including lower-resolution previews: Having pre-generated, less detailed versions of the data. This allows for fast and efficient data visualisations.

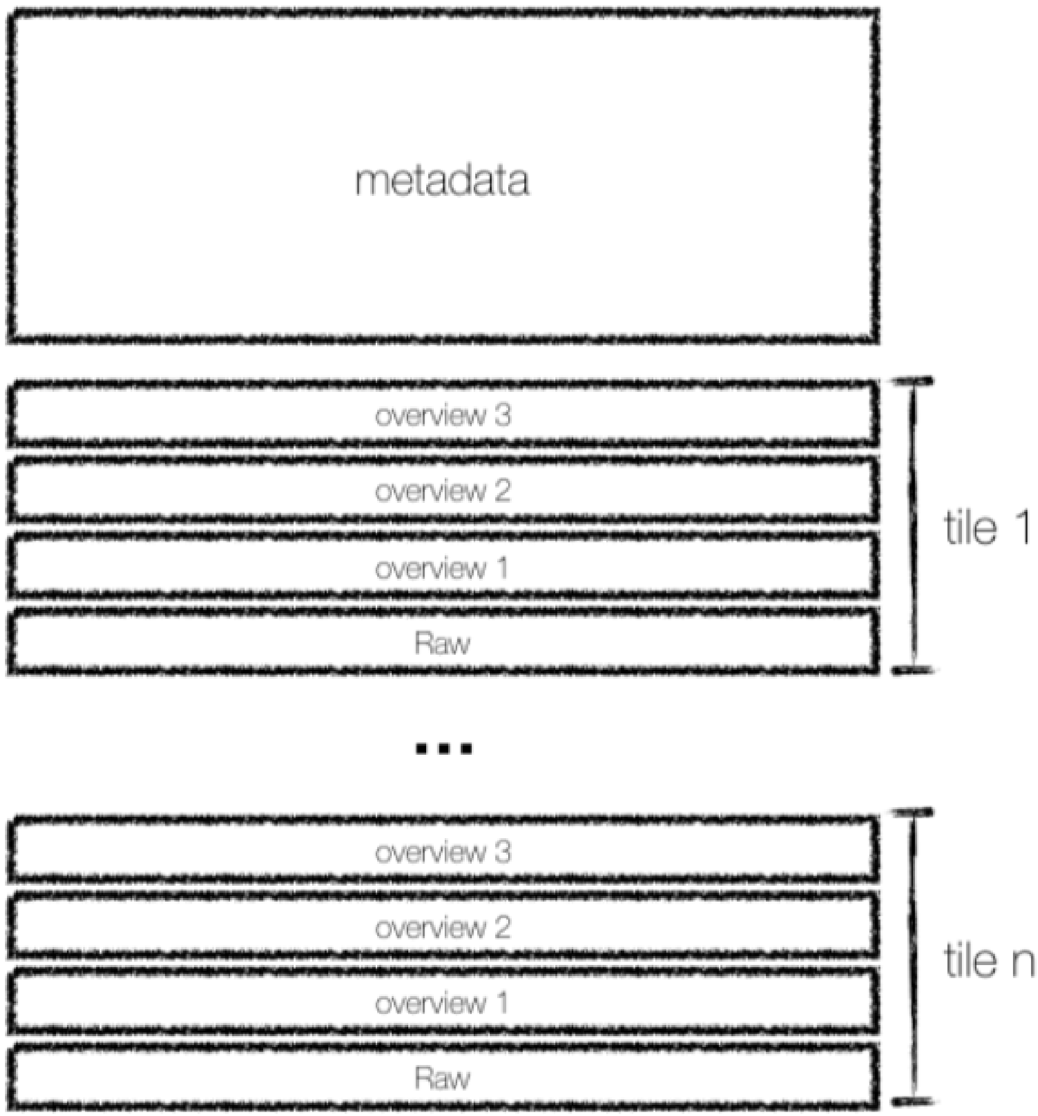

Figure 2.1: COG structure. Retrieved from CNG documentation

A key feature is the **Internal File Directory** (IFD), which acts like an internal index. This allows for retrieving only the parts of the data needed using simple web requests. For example, it is possible to access just the tiles covering Paris from a large Sentinel-2 image of Europe.

Zarr

This format is designed for handling large, multi-dimensional datasets (often called "data cubes"). Zarr, developed and maintained by the community, works by:

- Storing data as compressed **chunks** in a flexible way.
- Allowing for efficient indexing and processing of the data in parallel.
- Enabling specific descriptions (metadata) for each part of the data.

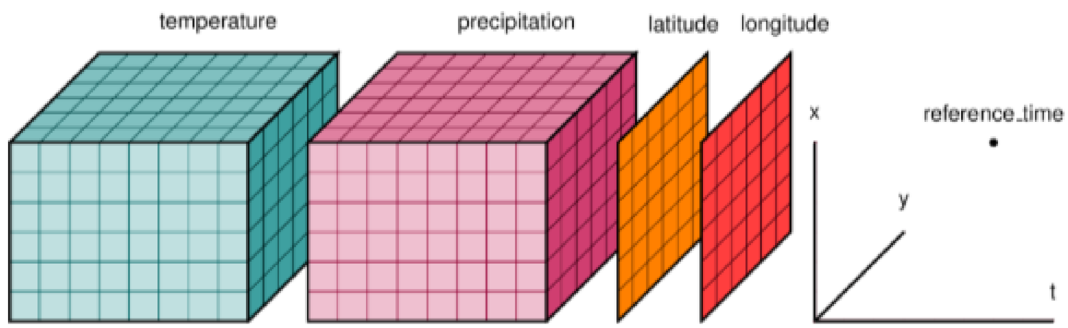

Figure 2.2: Zarr structure. Retrieved from CNG documentation.

For example, Zarr makes it possible to extract temperature data for the summer of 2023 from a climate dataset spanning 50 years without needing to load the entire dataset.

When to use COGs versus Zarr?

The table below compares some features of COG and Zarr:

Feature	Zarr	COG
Structure	Multi-file chunks	Single file
Access	Parallel	Sequential
Compression	Differently per-chunk	Whole-file
Scales	Multi-scale in single file	Separate, pre-generated lower-resolution files

Based on the structure and capabilities for each format, it is advised to use COGs when:

- Working with raster data (like images or elevation models) that have sudden changes.
- It is needed to easily visualise or access specific geographic areas without loading the entire dataset.
- Interoperability with existing GIS software is important, as COG is a widely adopted standard.

On the other hand, it is advised to use Zarr when:

- Dealing with large, multi-dimensional datasets that might be updated or modified.
- Performing complex analyses that involve accessing different parts of the data in parallel.
- Efficiently handling different resolutions or variables within a single dataset is required.
- Working in cloud environments that benefit from chunked data storage for parallel processing.

What's next?

Cloud-optimised formats bridge the gap between the vast potential of cloud storage and the need for efficient data access, ultimately making cloud-based geospatial analysis faster, more cost-effective, and more powerful.

Now that we have an idea of the available cloud-optimised formats and what cloud-optimised means, we will explore the EOPF data products.

3 The EOPF Available Datasets

The ESA Copernicus **Earth Observation Processor Framework** has made available assets as part of the Sentinel processor re-engineering in the EOPF Sentinel Sample Service. There, it is possible to explore different data sets in a Cloud Optimised format. So far, the missions integrated inside the catalogue are the Sentinel 1, Sentinel 2 and Sentinel 3. Each of them makes available to the public several Processing Levels (from raw info to composites) from which we can choose the data set that best suits our analysis needs.

> **!** Important
>
> The re-processing from the Sentinel Missions is an ongoing activity as part of the EOPF Sentinel Zarr Sample Service. This page and our tutorials will continuously be updated as soon as new data products are available.

An overview of the available data sets and the information they contain for each of the missions is explained in this document.

Available product Levels

Sentinel-1

Sentinel-1 is a radar imaging mission that is composed of a constellation of two polar-orbiting satellites providing continuous all-weather, day and night imagery.

Product	Instrument	Description	Available at
Level-1 GRD	Ground Range Detected	The Sentinel-1 Level-1 GDR products consist of focused SAR data that has been detected, multi-looked and projected to ground range using the Earth ellipsoid model WGS84.	this link
Level-1 SLC	Single Look Complex (	The Sentinel-1 Level-1 SLC products consist of focused SAR data, geo-referenced using orbit and attitude data from the satellite, and provided in slant-range geometry.	this link

Product	Instrument	Description	Available at
Level-2 OCN	Ocean	The Sentinel-1 Level-2 OCN products for wind, wave and currents applications may contain the following geophysical components derived from the SAR data: Ocean Wind field (OWI), Ocean Swell spectra (OSW), Surface Radial Velocity (RVL).	this link

Sentinel-2

Sentinel-2 acquires optical imagery at high spatial resolution (10m to 60m) over land and coastal waters. The mission supports applications such as agricultural monitoring, emergency management, land cover classifications, and water quality.

Product	Instrument	Description	Available at
Level-1C	Multi-Spectral Instrument	The Sentinel-2 Level-1C product is composed of 110x110 km2 tiles (ortho-images in UTM/WGS84 projection). Earth is subdivided on a predefined set of tiles, defined in UTM/WGS84 projection and using a 100 km step.	this link

Product	Instrument	Description	Available at
Level-2A	Multi-Spectral Instrument	The Sentinel-2 Level-2A Collection 1 product provides orthorectified Surface Reflectance (Bottom-Of-Atmosphere: BOA), with sub-pixel multispectral and multitemporal registration accuracy.	this link

Sentinel-3

Sentinel-3 is a mission that regularly measures our Earth's oceans, land, rivers, lakes, ice on land, sea ice, and the atmosphere. Its goal is to keep track of and help us understand how these large parts of our planet change over long periods.

3.0.1 Ocean and Land Colour Instrument

Product	Product	Description	Available at
Level-1 EFR	Earth Full Resolution	Provides TOA radiances at full resolution for each pixel in the instrument grid, each view and each OLCI channel, plus annotation data associated to OLCI pixels.	this link

Product	Product	Description	Available at
Level-1 ERR	Earth Reduced Resolution	The Sentinel-3 OLCI L1 ERR product provides TOA radiances at reduced resolution for each pixel in the instrument grid, each view and each OLCI channel, plus annotation data associated to OLCI pixels.	this link
Level-2 LFR	Land Full Resolution	The Sentinel-3 OLCI L2 LFR product provides land and atmospheric geophysical parameters computed for full resolution.	this link
Level-2 LRR	Land Reduced Resolution	The Sentinel-3 OLCI L2 LRR product provides land and atmospheric geophysical parameters computed for reduced resolution.	this link

3.0.2 Sea and Land Surface Temperature Radiometer

Product	Data	Description	Available at
Level-1 RBT	Radiance Brightness Temperature	The Sentinel-3 SLSTR Level-1B RBT product provides radiances and brightness temperatures for each pixel in a regular image grid for each view and SLSTR channel.	this link
Level-2 LST	LST: Land Surface Temperature	The Sentinel-3 SLSTR Level-2 LST product provides land surface temperature.	this link

In the following chapter, the `zarr` specific structure for each of the products will be explored, allowing us to better understand, how to integrate the new format into Earth Observation workflows.

Part II

About Zarr

Part III

[COMING] EOPF and STAC

Part IV

[COMING] Tools to work with Zarr

Part V

[COMING] EOPF in Action

4 Glossary

Here we introduce some important terms that are mentioned throughout the present book.

EOPF: Earth Observation Processing Framework

CDSE: Copernicus Data Space Ecosystem

CMP: Core Python Modules

HEALPix: Hierarchical Equal Area isoLatitude Pixelation.

GDR: Ground Range Detected

SLC: Singlme Look Complex

NRB: Normalized Radar Backscatter

SAFE: Standard Archive Format for Europe

STAC: Spatio Temporal Asset Catalog

References