**CAMEO**

**Architecture for Manipulating Earth Observation data**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**D1.1**

**WORK PACKAGE: WP1 Data Warehouse**

**DELIVERABLE: D1.1 Data Warehouse Design Document**

**SUBMISSION DATE: 30 June 2022**

**NATURE: WORK PACKAGE OUTPUT REPORT**

| **Document Control** | |
| --- | --- |
| **Creator** | **Dell Technologies** |
| **Description** | **Data Warehouse Design Document** |
| **Publisher** | **Dell Technologies** |
| **Audience** | **Consortium Partners** |
| **Status** | **Work Package Output Report** |

| **Version** | **Issue date** | Stage | Changes | Contributor |
| --- | --- | --- | --- | --- |
| **0.1** | **17th June 2022** | Draft | Initial draft | Merry Globin, David Bowden, Aidan O’Mahony, Alan Barnett, Michal Sworzeniowski, Matthew Keating, Anh Vu Vo, Namita Agarwal, Michela Bertolotto |
| **1.0** | **30th June 2022** | Final | Final version | Merry Globin, David Bowden, Aidan O’Mahony, Alan Barnett, Michal Sworzeniowski, Matthew Keating, Anh Vu Vo, Namita Agarwal, Michela Bertolotto |

**Contents**

[1](#_heading=h.lsdne2ccpf9) Introduction 5

[1.1](#_heading=h.44sinio) Purpose of this document 5

[1.2](#_heading=h.1jlao46) Scope 5

[1.3](#_heading=h.43ky6rz) Definitions, acronyms, and abbreviations 5

[1.4](#_heading=h.2iq8gzs) Reference Documents 7

[1.5](#_heading=h.xvir7l) General Constraints 7

[1.6](#_heading=h.3hv69ve) Assumptions 7

[2](#_heading=h.1x0gk37) Data Sources 9

[2.1 Live data feeds 9](#_heading=h.4h042r0)

[2.1.1 Sentinel satellite data 9](#_heading=h.2w5ecyt)

[2.1.2 EUMETSAT satellite data 13](#_heading=h.1baon6m)

[2.1.3 NASA/USGS’s Landsat satellite data 15](#_heading=h.3vac5uf)

[2.1.4 NASA’s MODIS satellite data 16](#_heading=h.2afmg28)

[2.1.5 Other satellite data sources 16](#_heading=h.pkwqa1)

[2.1.6 Information products from Copernicus thematic services 18](#_heading=h.39kk8xu)

[2.1.7 Irish meteorological data 18](#_heading=h.1opuj5n)

[2.1.8 Irish marine buoy observations and forecasts 19](#_heading=h.48pi1tg)

[2.1.9 Other live data sources 20](#_heading=h.2nusc19)

[2.2](#_heading=h.1302m92) Static Datasets 20

[2.2.1 Archived/offline/historic satellite datasets 21](#_heading=h.3mzq4wv)

[2.2.2 Information products from Copernicus thematic services 21](#_heading=h.2250f4o)

[2.2.3 Archived weather data from Met Eireann 22](#_heading=h.haapch)

[2.2.4 Elevation data products 22](#_heading=h.319y80a)

[2.2.5 Other sources of static data 23](#_heading=h.1gf8i83)

[2.3](#_heading=h.40ew0vw) Data Formats 24

[2.3.1 Container formats 24](#_heading=h.2fk6b3p)

[2.3.2 Imagery/raster data formats 25](#_heading=h.upglbi)

[2.3.3 Vector data formats 26](#_heading=h.3ep43zb)

[2.3.4 Data formats for array-oriented data 26](#_heading=h.1tuee74)

[2.3.4 Point cloud data formats 27](#_heading=h.4du1wux)

[2.3.5 Text-based formats for documenting metadata, annotation and other purposes 27](#_heading=h.2szc72q)

[2.4](#_heading=h.184mhaj) Data Processing Tools 28

[2.4.1 GDAL (Geospatial Data Abstraction Library) 28](#_heading=h.3s49zyc)

[2.4.2 Orfeo Toolbox (OTB) 28](#_heading=h.279ka65)

[2.4.3 SNAP Toolbox 28](#_heading=h.meukdy)

[2.4.4 GeoTools 29](#_heading=h.36ei31r)

[2.4.5 QGIS 29](#_heading=h.1ljsd9k)

[2.4.6 ArcGIS 29](#_heading=h.45jfvxd)

[3](#_heading=h.2koq656) Requirements 30

[3.1](#_heading=h.zu0gcz) Functional Requirements: 30

[3.2 Non-functional requirements: 31](#_heading=h.3jtnz0s)

[3.3](#_heading=h.1yyy98l) Summary 31

[4](#_heading=h.gaohqgfoc05g) Data Warehouse Architecture 36

[4.1 Main Design Features 36](#_heading=h.4iylrwe)

[4.1.1 EO data for ingestion 36](#_heading=h.2y3w247)

[4.1.2 Extract 36](#_heading=h.1d96cc0)

[4.1.3 Transform 37](#_heading=h.3x8tuzt)

[4.1.4 Load 37](#_heading=h.2ce457m)

[4.1.5 CAMEO Data Store 37](#_heading=h.rjefff)

[4.1.6 Metadata/Schema 37](#_heading=h.3bj1y38)

[4.2 Data Warehouse Layers 37](#_heading=h.1qoc8b1)

[4.2.1 Data Warehouse Access Layer 38](#_heading=h.4anzqyu)

[4.2.2. Storage Abstraction Layer 38](#_heading=h.2pta16n)

[4.2.3. Extract Transform Load Layer 38](#_heading=h.14ykbeg)

[4.3 Data Warehouse Components 39](#_heading=h.3oy7u29)

[4.3.1 CAMEO Webpage 39](#_heading=h.243i4a2)

[4.3.2 Login Authorization 39](#_heading=h.j8sehv)

[4.3.3 Security Layer 39](#_heading=h.338fx5o)

[4.3.4 Data Quality Monitoring 40](#_heading=h.1idq7dh)

[4.3.5 API Gateway 40](#_heading=h.42ddq1a)

[4.3.5.1 Data Retrieval API 40](#_heading=h.2hio093)

[4.3.5.2 Metadata Retrieval API 40](#_heading=h.wnyagw)

[4.3.5.3 Search API 40](#_heading=h.3gnlt4p)

[4.3.6 Storage Abstraction Layer 40](#_heading=h.1vsw3ci)

[4.3.6.1 Metadata abstraction & ID allocation 41](#_heading=h.4fsjm0b)

[4.3.6.2 Event recognition & notification 41](#_heading=h.2uxtw84)

[4.3.6.3 Event Emitter/Receiver 41](#_heading=h.1a346fx)

[4.3.6.4 Storage Service 41](#_heading=h.3u2rp3q)

[4.3.7 Object Store 41](#_heading=h.2981zbj)

[4.3.7.1 Namespace 41](#_heading=h.odc9jc)

[4.3.8 Pre-processing 41](#_heading=h.38czs75)

[4.3.8.1 Pipeline Engine 41](#_heading=h.1nia2ey)

[4.3.8.2 Additional Processes 42](#_heading=h.47hxl2r)

[4.3.9 Metadata Index 42](#_heading=h.2mn7vak)

[4.3.10 Stream/Data Ingestion Mechanism 42](#_heading=h.11si5id)

[4.3.11 Metadata Crawler 42](#_heading=h.3ls5o66)

[4.4 Tools Used 42](#_heading=h.20xfydz)

[4.5 Standards 43](#_heading=h.4kx3h1s)

[4.6 Database design 43](#_heading=h.302dr9l)

[4.7 Files 43](#_heading=h.1f7o1he)

[4.8 Reports 43](#_heading=h.3z7bk57)

[4.9 Error Handling 44](#_heading=h.2eclud0)

[4.10 Interfaces 44](#_heading=h.thw4kt)

[4.11 Help 44](#_heading=h.3dhjn8m)

[4.12 Performance 44](#_heading=h.1smtxgf)

[4.13 Security 44](#_heading=h.4cmhg48)

[4.14 Reliability 44](#_heading=h.2rrrqc1)

[4.15 Maintainability 44](#_heading=h.16x20ju)

[4.16 Portability 45](#_heading=h.3qwpj7n)

[4.17 Reusability 45](#_heading=h.261ztfg)

[4.18 Application compatibility 45](#_heading=h.l7a3n9)

[5](#_heading=h.ycnptowlzsjz) Conclusion 46

# 1 Introduction

## Purpose of this document

This document is the design document for the Data Warehouse. This document includes the requirements and architecture of the proposed Data Warehouse. The requirements might be refined as the project develops according to the project goals and objectives.

The rest of the document is organised as follows: Chapter 2 briefs the data sources for EO data which will be considered in the project CAMEO. Chapter 3 lists the requirements for the Data Warehouse for the project CAMEO. Finally, Chapter 4 provides the high-level architecture of the Data Warehouse.

## Scope

This document defines the requirements and high-level architecture of the Data Warehouse. Insights and suggestions from other partners during the survey process have been incorporated in this document. This document is not intended as design requirements for the analytical engine which is part of Work Package 2 and security requirements which is part of Work Package 4.

## 1.3 Definitions, acronyms, and abbreviations

| **Acronym** | **Acronym meaning** |
| --- | --- |
| API | Application Programming Interface |
| AWS | Amazon Web Service |
| BDAP | JRC Big Data Analytics Platform |
| CAMS | Copernicus Atmosphere Monitoring Service |
| COG | Cloud Optimised GeoTIFF |
| CNES | Centre National D'Etudes Spatiales |
| Copernicus | European Programme for the establishment of a European capacity for Earth Observation |
| DIAS | Data and Information Access Services |
| DSM | Digital Surface Model |
| DTM | Digital Terrain Model |
| EC | European Commission |
| ECS | Elastic Cloud Server |
| EEA | European Environmental Agency |
| ENS | Elastic Node Server |
| EO | Earth Observation |
| EP | Exploitation Platforms |
| EPA | Environmental protection agency |
| ESA | European Space Agency |
| EU | European Union |
| EUMETSAT | European Organisation for the Exploitation of Meteorological Satellites |
| FTP | File Transfer Protocol |
| GDAL/OGR | Geospatial Data Abstraction Library |
| GEE | Google Earth Engine |
| GEO | Group on Earth Observations |
| GEOSS | Global Earth Observation System of Systems |
| GeoTiff | Georeferenced Tagged Image File Format |
| GIF | Graphics Interchange Format |
| GIS | Geographic Information System |
| GML | Geography Markup Language |
| GRD | Ground Range Detected |
| GSI | Geological Survey Ireland |
| GUI | Graphical User Interface |
| HDF | Hierarchical Data Format |
| HTML | HyperText Markup Language |
| HTTP | HyperText Transfer Protocol |
| IaaS | Infrastructure as a Service |
| IFORIS | Integrated Forestry Information System |
| IOT | Internet of things |
| ISO | Organisation for Standardisation |
| JEODPP | JRC Earth Observation Data and Processing Platform |
| JPEG | Joint Photographic Experts Group |
| JRC | Joint Research Centre |
| JSON | JavaScript Object Notation |
| KML | Keyhole Markup Language |
| LiDAR | Light Detection And Ranging |
| MODIS | Moderate Resolution Imaging Spectroradiometer |
| NASA | National Aeronautics and Space Administration |
| netCDF | Network Common Data Form |
| NRT | Near Real Time |
| NTC | Non Time Critical |
| OCN | Ocean |
| OData | Open Data Protocol |
| OGC | Open Geospatial Consortium |
| OLCI | Ocean and Land Colour Instrument |
| OPW | Office of Public Works |
| OSI | Ordnance Survey Ireland |
| OTB | Orfeo Toolbox |
| PaaS | Platform as a Service |
| PB | Petabyte |
| PNG | Portable Network Graphics |
| SaaS | Software as a Service |
| SAFE | Standard Archive Format for Europe |
| SEPAL | Processing and Analysis for Land Monitoring |
| SH | Sentinel Hub |
| SLC | Single Look Complex |
| SLSTR | Sea and Land Surface Temperature Radiometer |
| SNAP | Sentinel Application Platform |
| SQL | Structured Query Language |
| STC | Slow Time Critical |
| SVG | Scalable Vector Graphics |
| TEP | Thematic Exploitation Platforms |
| TIFF | Tagged Image File Format |
| TII | Transport Infrastructure Ireland |
| USGS | United States Geological Survey |
| XML | Extensible Markup Language |

## 1.4 Reference Documents

[CAMEO DTIF Proposal](https://drive.google.com/file/d/1eQWUumvRr19Zg4pZ0kLCdlSbhySKAUgj/view?usp=sharing)

[CAMEO Workshop Report February 2022](https://drive.google.com/file/d/18DoN_ymUuy5g2r_0Amxnu8BlX2n8z6J7/view?usp=sharing)

[CAMEO Requirements Summary](https://drive.google.com/file/d/1HswBomcRk_EkW8GvsSY0iKZQnTrAw-dR/view?usp=sharing)

[CAMEO Work Programme](https://drive.google.com/file/d/1VuVzIBtur7HOHC2KYlAWidINz490oRIv/view?usp=sharing)

## 1.5 General Constraints

A number of constraints on this document exists. These are:

1. The specific software packages, libraries, and applications have not been decided upon. Therefore, the design of the data warehouse is agnostic of specific software.
2. The precise mechanisms for interacting with both the EO data sources and the WP2 analytic engine are not completely determined.

## 1.6 Assumptions

A number of assumptions have been made when creating this document. These are:

1. The cost for acquiring EO data and the licence for use of that data is not so prohibitive as to prevent the DW from being implemented.
2. The underlying hardware use for the DW remains to be scoped.

# 

# 2 Data Sources

Earth observation is the gathering of information about the physical, chemical and biological systems of the Earth[[1]](#footnote-0). EO can be performed via remote sensing technologies (e.g. photogrammetry, radar, lidar), through direct-contact sensors (e.g. meteorological stations, ocean buoys, seismic stations), or through non-sensing methods (e.g. census, forecast, modelling). EO can be conducted from a space-based, airborne, seaborne, or from a ground-based platform. EO data include data directly derived from EO activities (e.g. satellite images) and their derived data products (e.g. land-use, land-cover maps). This chapter briefs some of the EO data sources for consideration in project CAMEO.

The data sources presented in this Chapter are derived from the information contributed by CAMEO SMEs through a survey conducted in February 2022. Additional data sources are gathered by reviewing relevant international and national data repositories. Within this chapter, Section 2.1 describes live data feeds that continuously provide recent data from operational sensors. Some of the live data feeds are real-time (available immediately after acquisition). Others are available from a few hours up to a month after acquisition. Section 2.2 describes static data sources, which include datasets that are fixed, rarely require an update, or are archives of historical observations. Section 2.3 summarises and describes the formats of the data sources in 2.1 and 2.2. Section 2.4 lists some data pre-processing tools useful for preparing data for ingestion and for manipulating data inside the data warehouse.

## 2.1 Live data feeds

The majority of live data feeds presented in this section are data and information products derived from operational satellites. Other data comes from in-situ measurements, forecasting and modelling.

### 2.1.1 Sentinel satellite data

Sentinels are a set of satellites of the Copernicus programme. Among these satellites, Sentinel 1, 2, 3, 5-P, and 6 are dedicated satellites, while Sentinel 4, and 5 are instruments onboard EUMETSAT’s weather satellites. ESA operates the Sentinel 1, 2, and 5-P, while EUMETSAT operates Sentinel 3, and 6 satellites, and Sentinel 4, and 5 instruments onboard the weather satellites[[2]](#footnote-1). Table 1 Sentinel satellites summarises key information about the Sentinel missions.

*Table 1 Sentinel satellites*

| Satellite Name | Purpose | No of satellites | Instruments | Processing Levels | Resolution | Launch date | Lifetime & status | Repeat cycle |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sentinel-1 | Marine monitoring, land monitoring | 2 | SAR | • Level-0  •Level-1 Single Look Complex  •Level-1 Ground Range Detected  •Level-2 | •Strip Map Mode: 80 km swath, 5 x 5 m spatial resolution  •Interferometric Wide Swath: 250 km swath, 5 x 20 m spatial resolution  •Extra-Wide Swath Mode: 400 km swath, 20 x 40 m | 3 April 2014 & 25 April 2016 | 7-year design life.  Operational | 12 days |
| Sentinel-2 | Land monitoring | 2 | MSI | •Level-0  •Level-1A •Level-1B  •Level-1C •Level-2A | •Four bands at 10-meter spatial resolution  •Six bands at 20-meter spatial resolution  •Three bands at 60-meter spatial resolution | 23 June 2015 & 7 March 2017 | 7.25-year design life.  Operational | 10 days |
| Sentinel-3 | Marine monitoring, land monitoring | 3 | •SLSTR  • OLCI  • SRAL  • MWR  • POD | •Level-0  •Level-1 •Level-2 | •300-meter full resolution  •1200-meter reduced resolution | 16 Feb 2016 & 25 April 2018 | 7.25-year design life.  Operational | 27 days |
| Sentinel-4 | Air quality monitoring | Instrument on EUMETSAT satellites | Spectrometer | •Level-0 •Level-1B  •Level-2 | 8x8 km2 spatial resolution | 2019 | 8.25-year design life.  Operational | 1 hour |
| Sentinel-5 | Air quality monitoring | Instrument on EUMETSAT satellites | Spectrometer | •Level-0 •Level-1B  •Level-2 | 7x7 km spatial resolution | 13 Oct 2017 | 7.25-year design life.  Operational |  |
| Sentinel-5P | Air quality monitoring | 1 | Tropomi | •Level-0 •Level-1B  •Level-2 | 5.5\* km x 3.5 km spatial resolution | 13 Oct 2017 | 7-year design life |  |
| Sentinel-6 | Sea-surface height measurements | 2 | •Dual-frequency radar altimeter :Poseidon-4  •AMR-C  • GNSS POD Receiver, DORIS  • Laser Retroreflector Array  •GNSS-RO TriG Receiver for Radio Occultation  •Radiation Environment Monitor |  |  | 21 Nov 2020 | 5.5-year design life.  Operational | 10 days |

The Sentinel programme provides more than 40 different data products categorised by the satellite missions, instruments and processing levels. A complete list of Sentinel data products as of 2014 is available in the ESA’s Sentinels Data Products List[[3]](#footnote-2). The delivery timelines for the different data products vary from under 1 hour up to 1 month after data acquisition. The ESA uses the concepts of Near Real Time (NRT), Slow Time Critical (STC) and Non Time Critical (NTC) to indicate different data delivery timelines of less than 3 hours, within 48 hours and within 1 month from the data acquisition time.

The ESA provides Sentinel-1, -2, -3, and -5P data in full, open, and free-of-charge, through its Copernicus Open Access Hub. The Open Access Hub provides synchronous access to the latest data, and asynchronous access to the historical data. Synchronous access is immediate and is available for the latest month of products at the minimum. Asynchronous access to historical data requires submitting a request to restore the data. Requests are processed within 1 hour after submission and data are available for download within the 3 subsequent days. The Open Access Hub provides graphical user interfaces for searching, previewing, and requesting data. Search criteria include data acquisition modes, product types, product levels, and geographical areas. The Open Access Hub also exposes two APIs, Open Search (Apache Solr)[[4]](#footnote-3) and Open Data Protocol (OData). The Open Search API can be used to identify the required products which then can be downloaded by using OData[[5]](#footnote-4).

In addition to the Open Access Hub, ESA provides five Data and Information Access Service platforms (i.e. CREODIAS, Mundi, ONDA, WEkEO and Sobloo) for Sentinel data access. Alternatively, Sentinel data can be accessed using commercial platforms such as Sentinel Hub. A thorough review of the five DIASs and Sentinel Hub is available in CAMEO Deliverable 7.1.

Table 2 shows the Sentinel data products which are of most interest to the CAMEO SMEs. They include Level-1 and Level-2 products from Sentinel-1, -2, and -3 satellites. Some SMEs indicated their plans to use Sentinel-5 and Sentinel-6 products in the future even though the specific data products have not been decided. All Sentinel data are delivered in the ESA’s SAFE package format. The measurement data inside the package are in GeoTIFF, JPEG2000, and netCDF formats.

*Table 2 Sentinel data products relevant to CAMEO partners*

| **Data products** | **Description** | **Format** |
| --- | --- | --- |
| Sentinel-1 SLC L1 | Single Look Complex Level-1 imagery with amplitude and phase | SAFE/GeoTIFF |
| Sentinel-1 GRD L1 | Ground Range Detected Level-1 data with multi-looked intensity only | SAFE/GeoTIFF |
| Sentinel-1 OCN L2 | Ocean Level-2 data for retrieved geophysical parameters of the ocean | SAFE/netCDF |
| Sentinel-2 L1C | Top of atmosphere reflectance | SAFE/JPEG2000 |
| Sentinel-2 L2A | Bottom of the atmosphere reflectance | SAFE/JPEG2000 |
| Sentinel-3 SLSTR L1 | Radiances and brightness temperatures | SAFE/netCDF |
| Sentinel-3 OLCI L1 | Top of atmosphere radiometric measurements, radiometrically corrected, calibrated and spectrally characterised | SAFE/netCDF |
| Sentinel-5 products | Air quality data - processing levels have not been decided |  |
| Sentinel-6 products | Sea surface height data - processing levels have not been decided |  |

### 2.1.2 EUMETSAT satellite data

In addition to several Sentinel satellites and Sentinel instruments mentioned in 2.1.1, EUMETSAT operates Meteosat-8 (retires on1/7/2022), -9, -10, and –11, Metop-B, -C, and Jason-3. While not all EUMETSAT data are free, Ireland has access to data from EUMETSAT. EUMETSAT data are provided either in near real-time or in historical mode. Access to EUMETSAT data is available via EUMETSAT Data Store[[6]](#footnote-5), which provides a web interface and a suite of APIs. Data products can be explored by using the Product Navigator[[7]](#footnote-6). The data products are disseminated in various formats, including HRIT, netCDF, HDF, etc[[8]](#footnote-7). EUMETSAT also provides a real-time data push service called EUMETCast[[9]](#footnote-8).

*Table 3 EUMETSAT satellites*

| Satellite name | Purpose | No of satellites | Resolution | Instruments | Launch date | Lifetime & status | Repeat cycle |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Meteosat-8 | Weather nowcasting, forecasting, climate monitoring for Europe and Africa | 1 | 1 km Ch 12 (HRV)  3 km Ch 1–11 (VIS, IR, WV) | * GERB * SEVIRI | 28/08/2002 | 20 years. To be retired on 1/7/2022 | 15-mins (full disc)  Rapid Scan 5-mins (Europe) |
| Meteosat-9 | Weather nowcasting, forecasting, climate monitoring for Europe and Africa | 1 | 1 km Ch 12 (HRV)  3 km Ch 1–11 (VIS, IR, WV) | * GERB * SEVIRI | 22/12/2005 | 20 years. Operational | 15-mins (full disc)  Rapid Scan 5-mins (Europe) |
| Meteosat-10 | Weather nowcasting, forecasting, climate monitoring for Europe and Africa | 1 | 1 km Ch 12 (HRV)  3 km Ch 1–11 (VIS, IR, WV) | * GERB * SEVIRI | 05/07/2012 | 18 years. Operational | 15-mins (full disc)  Rapid Scan 5-mins (Europe) |
| Meteosat-11 | Weather nowcasting, forecasting, climate monitoring for Europe and Africa | 1 | 1 km Ch 12 (HRV)  3 km Ch 1–11 (VIS, IR, WV) | * GERB * SEVIRI | 15/07/2015 | 18 years. Operational | 15-mins (full disc)  Rapid Scan 5-mins (Europe) |
| Metop-B | Global weather forecasting and climate monitoring | 1 | 50km | * A-DCS * AMSU-A * ASCAT * AVHRR/3 * GOME-2 * GRAS * HIRS/4 * IASI * MHS * SEM/MEPED * SEM/TED | 17/09/2012 | 12 years. Operational | 29 days |
| Metop-C | Global weather forecasting and climate monitoring | 1 | 50km | * A-DCS * AMSU-A * ASCAT * AVHRR/3 * GOME-2 * GRAS * IASI * MHS | 07/11/2018 | 9 years. Operational | 29 days |
| Jason-3 | Measure sea surface height, wind speed, ocean surface and significant wave height | 1 |  | * AMR * DORIS * LRA * Poseidon-3B * TRSR * CARMEN * LPT | 17/1/2016 | 7 years. Operational | 9.9 days |

* GERB Geostationary Earth Radiation Budget
* SEVIRI Spinning Enhanced Visible Infra-Red Imager
* A-DCS Advanced Data Collection System (also called "Argos-3")
* AMSU-A Advanced Microwave Sounding Unit - A
* ASCAT Advanced Scatterometer
* AVHRR/3 Advanced Very High Resolution Radiometer / 3
* GOME-2 Global Ozone Monitoring Experiment - 2
* GRAS GNSS Receiver for Atmospheric Sounding
* HIRS/4 High-resolution Infra Red Sounder / 4
* IASI Infrared Atmospheric Sounding Interferometer
* MHS Microwave Humidity Sounding
* SEM/MEPED SEM / Medium energy proton detector
* SEM/TED SEM / Total Energy Detector
* AMR Advanced Microwave Radiometer
* DORIS Doppler Orbitography and Radiopositioning Integrated by Satellite
* LRA Laser Retroreflector Array
* Poseidon-3B Poseidon 3B
* TRSR Turbo Rogue Space Receiver
* CARMEN Environment Characterization and Modelisation
* LPT Light Particle Telescope

### 2.1.3 NASA/USGS’s Landsat satellite data

The Landsat programme provides the continuous space-based record of the Earth’s land since 1972. Two satellites currently operating are Landsat-8, and -9. Landsat Level-1, -2, and -3 products are available for download at no charge from the USGS archive through platforms such as USGS’s Earth Explorer[[10]](#footnote-9). A complete list of Landsat data products and instructions for downloading the data are available on USGS website[[11]](#footnote-10). Bulk download of the data is possible via the USGS’s file system access API[[12]](#footnote-11). Many Landsat data products are also available on Sentinel Hub and on several DIAS platforms such as Mundiwebservice and Creodias. Several CAMEO SMEs indicated their use of Landsat-8 data at processing Levels-1 and -2.

*Table 4 Operational Landsat satellites*

| Satellite name | Purpose | No of satellites | Instruments | Resolution | Launch date | Lifetime & status | Repeat cycle |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Landsat-8 | Land imaging |  | OLI, TIRS | 30m for Bands 1-7 & 9; 15m for Band 8; 100m for Bands 10&11 | 11/2/2013 | 5 years. Operational | 16 days |
| Landsat-9 | Monitoring, understanding and managing the land resources |  | OLI-2; TIRS-2 | 30m for shortwave infrared spectral bands, 15m for panchromatic band, and 100m for thermal bands | 27/9/2021 | 5 years. Operational | 16 days |

* OLI Operational Land Imager
* TIRS Thermal Infrared Sensor

### 2.1.4 NASA’s MODIS satellite data

Data from the MODIS (Moderate Resolution Imaging Spectroradiometer) instruments aboard NASA’s Terra and Aqua satellites is another source of data considered in CAMEO. The data are images of the Earth in 36 bands within the visual 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. Instructions for downloading MODIS data products are available on MODIS’s website[[13]](#footnote-12). Some MODIS data products (e.g. MCD43A4 Nadir BRDF-Adjusted Reflectance (NBAR) product Version 6) are computed daily and made available in Sentinel Hub.

### 2.1.5 Other satellite data sources

**MAXAR’s WorldView satellite data**

The WorldView series (WorldView-1,2,3,&4) consists of high-resolution commercial Earth-imaging satellites[[14]](#footnote-13). The satellites are owned by DigitalGlobe, a subsidiary of Maxar Technologies. All four WorldView satellites are currently operational, and they have been supplying imagery since 2007.

* WorldView-1 launched in 2007, it was the world's first 50 cm resolution commercial imaging satellite
* WorldView-2 launched in 2009, as a follow-on spacecraft to WorldView-1. It was the first high resolution 8-band multispectral commercial satellite.
* WorldView-3 launched in 2014, it is very similar to WorldView-2 but positioned in a lower orbit
* WorldView-4 launched in 2016, and provided 31 cm panchromatic imagery and 1.23 m multispectral imagery

The WorldView series is part of ESA's Third Party Missions Programme in which ESA has an agreement with European Space Imaging to distribute data products from the missions. Data from WorldView-3&4 are of interest to CAMEO SMEs as indicated in the survey responses. As members of ESA Member States, users in Ireland have access to the ESA archives of WorldView data[[15]](#footnote-14),[[16]](#footnote-15). Data are available for on demand ordering, upon submission of a Project Proposal subject to evaluation and acceptance by ESA and the data owner. Product programming quotas, if applicable, are assigned during the evaluation. The process usually takes up to two weeks and there are no deadlines for submitting a proposal. WorldView data are also available for purchase, order and access on Sentinel Hub[[17]](#footnote-16). WorldView data are distributed in GeoTIFF and NITF formats.

**ICEYE**

ICEYE is a Finnish company that develops and operates a constellation of small, commercial radar satellites. As of the beginning of 2022, there were 16 satellites in the constellation[[18]](#footnote-17). The constellation as a whole has the ability to observe a location at different times of the day rather than the more conventional dawn-dusk sun-synchronous orbit. That high revisit rate is one of the distinguishable capabilities ICEYE provides. As indicated in the survey responses, ICEYE data are of interest to some CAMEO SMEs. The data are available for scientific research and application development through the ESA archive. Data access can be requested by submitting a proposal to the ESA13. Details about ICEYE data products can be found in the ICEYE Product Specs[[19]](#footnote-18). ICEYE data are distributed as HDF5 files (SLC Complex data) and GeoTiff files (GRD Amplitude data).

**Capella Space**

Capella Space is an American space company that develops and operates small, commercial X-SAR satellites[[20]](#footnote-19). The Capella constellation will consist of 36 microsatellites that provide average imaging revisit times of less than one hour. Capella itself offers a portion of its historical archived imagery as open data[[21]](#footnote-20). The company provides a web-based data portal, Capella Console, and an API to access its data[[22]](#footnote-21). Some CAMEO SMEs are using data from Capella Space.

**PlanetScope and Skysat**

PlanetScope and Skysat are two optical constellations owned by Planet Labs PCB (Planet). The satellites are able to image the entire land surface of the Earth every day at a high resolution (approximately 3 meters per pixel). ESA is offering, for scientific research and application development, access to data from the full archives of the PlanetScope and Skysat missions upon submission and acceptance of a project proposal [[23]](#footnote-22),[[24]](#footnote-23). The data are distributed in GeoTIFF format with auxiliary data in JSON and other text formats. Some CAMEO SMEs indicated their use of PlanetScope data in the SMEs survey responses.

**SPOT and Pléiades**

The SPOT (from French "Satellite pour l'Observation de la Terre") series consists of seven satellites (SPOT-1 to -7). The series, initiated by Centre National D'Etudes Spatiales (CNES) of France, has been providing high-resolution, wide-area optical imagery for various applications in mapping, vegetation monitoring, land use and land cover, and the impacts of natural disasters. Two satellites currently operating are SPOT-6 and SPOT-7, which are commercial satellites owned by Airbus Defence and Space. Pléiades is an environment-focused constellation of two satellites from CNES. The satellites deliver very-high optical resolution images (0.5-2.0 m) with a swath of 20 km and a repeat cycle of 26 days[[25]](#footnote-24). Pléiades satellites complements the imaging capabilities of the SPOT satellites with a higher spatial resolution and a narrower field of view.

The SPOT series is part of ESA's Third Party Missions Programme, in which ESA has an agreement with Airbus Defence and Space to distribute data products from the mission. Users from EU member states can request SPOT-6 and -7 data via the ESA through the Fast (2-day) approval process[[26]](#footnote-25). The ESA also maintains an archive of Pléiades data. Access to the archive can be requested for scientific research and application development by submitting a project proposal to the ESA[[27]](#footnote-26). SPOT and Pléiades data products are delivered in DIMAP V2 format. Imagery data in the products are in either GeoTIFF or JPEG2000 format[[28]](#footnote-27). Data from SPOT and Pléiades are among the data sources mentioned in the SMEs survey responses.

**ESA Earth Explorers**

Together with the Copernicus Sentinels and EUMETSAT meteorology missions, Earth Explorers is another family of ESA’s EO satellites. Earth Explorers consists of small research missions dedicated to specific aspects of our Earth environment whilst demonstrating new technology in space. Under the programme are GOCE (Gravity Field and Steady-State Ocean Explorer), SMOS (Soil Moisture and Ocean Salinity), CryoSat (monitor icecaps shrinkage), Swarm (monitors the Earth's magnetic field), ADM-Aeolus (Atmospheric Dynamics Mission), EarthCARE (Earth Clouds, Aerosols and Radiation Explorer), and Biomass. Data from the Biomass mission, which is scheduled for launch in 2023, may be of interest to some CAMEO SMEs.

### 2.1.6 Information products from Copernicus thematic services

The Copernicus programme provides free-of-charge information products corresponding to six thematic services: Atmosphere, Marine, Land, Climate Change, Security, and Emergency[[29]](#footnote-28). Some of the services describe the current situation and forecast the situation a few days ahead. An example of dynamic information products is the information on active fires in the European Forest Fire Information System, which is updated 6 times a day[[30]](#footnote-29). Another example is the Ocean Physics Analysis and Forecast for the North-West European Shelf provided as hourly instantaneous, quarter-hourly, and daily 25-hour, de-tided, averages[[31]](#footnote-30). Others are more static as they describe analyses of historical data. The services have reached different degrees of maturity. Not all services are operational. Information products from the services can be accessed from the Copernicus services catalogue[[32]](#footnote-31) and the DIAS platforms.

The following data sources from Copernicus thematic services are of interest to CAMEO SMEs:

* Wave observation, analysis, forecast, hindcast, and reanalysis data[[33]](#footnote-32) from the Copernicus Marine service
* ERA5 and ERA-Interim (hourly estimates of many atmospheric, land and oceanic climate variables derived from reanalysis of the global climate)[[34]](#footnote-33) from the Copernicus Climate service
* Corine Landcover data (see Section 2.2 for more details)

### 2.1.7 Irish meteorological data

Met Éireann provides several live data sources through Ireland’s Open Data Portal. The data sources include:

* Weather forecast[[35]](#footnote-34)
* Text forecast[[36]](#footnote-35)
* Weather warnings[[37]](#footnote-36)
* Weather observations at synoptic stations[[38]](#footnote-37),[[39]](#footnote-38)
* Rainfall radar data[[40]](#footnote-39)

Weather forecast is provided for any point location (latitude and longitude) and is available in (i) 1-hour intervals out to 90 hours, (ii) three-hour intervals from 90 hours to 144 hours, and (iii) six-hour intervals from 144 hours out to 240 hours. The meteorological parameters for each timestep are timing, location, temperature, wind direction and speed, humidity, pressure, cloud, dewpoint, precipitation, and weather symbol. The data are provided in XML format via an API.

The text forecast includes information meant for human readers/consumers. The data include sea area forecast, farming commentary, inland lake forecast, current observations, town forecast, county forecast, etc. The data are provided in XML format via an API.

Weather warnings on a three-level scale (Yellow, Orange, Red) are issued on a county-by-county basis. The information is meant for mitigating damage to property and to reduce disturbance to economic activity and so enhance support for impact-based decision making for weather events. The provided information include time, location (county regions), type of weather event, severity, level of certainty, description of the warning, etc. The information is provided in JSON format via an API.

In addition, Met Éireann provides detailed weather observation data (e.g. temperature, wind speed, wind gust, wind direction, humidity, rainfall, pressure) recorded at about 20 weather observing stations across the country for every current day and the day before. Data are available in JSON and CSV formats via an API.

Near real-time rainfall data recorded by the radar stations at Dublin Airport and Shannon Airport is another data source. The data can be requested by email and is available in HDF5 format.

### 2.1.8 Irish marine buoy observations and forecasts

Marine Institute Ireland provides several real-time observation and forecast data sources through Ireland’s Open Data Portal. Examples include:

* Weather buoy real-time data[[41]](#footnote-40)
* Wave buoy near real-time data[[42]](#footnote-41)
* Buoy wave forecast[[43]](#footnote-42)
* SmartBay Ireland buoy wave (Galway Bay[[44]](#footnote-43), Cork[[45]](#footnote-44))

Weather buoy data are meteorological and oceanographic data recorded at 5 weather buoy stations around Ireland[[46]](#footnote-45) (including buoys M2-M6 located in Dublin, Wexford, Donegal, Cork, and Galway). The weather buoy network has been established to improve predictive weather forecasting. Parameters collected include: DateTime (yyyy-mm-ddThh:mm:ss.sss), Atmospheric Pressure (mbar), Air Temperature (degreeCelsius), DewPoint Temperature (degreeCelsius), Wind Speed (knots), Max Gust Wind Speed (knots), Wind Direction (degreeTrue), Sea Surface Temperature (degreeCelsius), Wave Period (seconds), Wave Height (metres) and Relative Humidity (%). The data can be accessed via the Marine Institute’s ERDDAP system[[47]](#footnote-46), which provides an API as well as a web form for data requests. Data are available in multiple formats including CSV, JSON, GeoJSON, and netCDF.

Wave buoy near real-time data are data recorded from the stations in the Wave Buoy Network[[48]](#footnote-47). The parameters collected include: Longitude (decimal degrees), Latitude (decimal degrees), dateTime (yyyy-mm-ddThh:mm:ss.sss), Station ID, Peak Period (secs), Peak Direction (Degrees True), Upcross Period (secs), Significant Wave Height (cm) and Sea Temperature (degreesC). The data are supposed to be available on the Marine Institute’s ERDDAP system. However, access to the data was unavailable at the time of writing. Alternative access to the data is available at the Marine Institute Data Catalogue website[[49]](#footnote-48). The data formats include CSV, TSV, and JSON.

In addition to the actual measurements, Marine Institute provides wave climate forecasts including significant wave height, mean wave period, mean wave direction, wave per unit crest length, peak period and energy period at the weather and wave buoy locations. The forecast data are also provided through ERDDAP and are available in CSV, TSV, and JSON formats. Ireland’s Open Data Portal has several records of data from the Marine Institute SmartBay facilities in Galway and Cork. Some data were not accessible on the portal at the time of writing. The SmartBay website[[50]](#footnote-49) indicates that the recorded data includes a live video stream, the depth of the observatory node, the sea temperature and salinity, and estimates of the chlorophyll and turbidity levels in the water. Underwater acoustic data are also recorded[[51]](#footnote-50). Wave data from the data buoy is also available to view in real-time. The video data are available as a set of static files in MP4 format. Notably, as of June 2022, the latest available videos were from 2021 even though the data are supposed to be available in real-time[[52]](#footnote-51).

Data from the weather and wave buoys are being extensively used by some CAMEO SMEs.

### 2.1.9 Other live data sources

Apart from the data sources described in 2.1.7 and 2.1.8, there are many other real-time datasets available on Ireland’s Open Data Portal. Examples include the sea level recordings at the 19 tide gauges around the coastline[[53]](#footnote-52), the real-time modelling of bottom current[[54]](#footnote-53), temperature[[55]](#footnote-54), and salinity[[56]](#footnote-55), etc. in Northeast Atlantic ocean, the real-time water level data for water monitoring stations in the Office of Public Works hydrometric network[[57]](#footnote-56), and the real-time air quality monitoring data in Dublin[[58]](#footnote-57). Some of those datasets could be representatives for in-situ sensing and modelling data to be considered in CAMEO.

## 2.2 Static Datasets

This section describes EO datasets that are fixed, rarely need to be updated, or historical.

### 2.2.1 Archived/offline/historic satellite datasets

In addition to the fresh data recently captured by the satellite missions described in Sections 2.1.1 to 2.1.6, CAMEO will retain certain historical data from those missions in the data warehouse. The data types and formats are not different from those described in Section 2.1. A notable difference is that historical data are archived and require time to restore before being accessible on some data access platforms such as the Copernicus Open Access Hub and the ONDA DIAS. Access to historical data is referred to as asynchronous access in the Copernicus Open Access Hub and as offline access in the ONDA DIAS.

### 2.2.2 Information products from Copernicus thematic services

Many of the data products from the Copernicus thematic services are static data. Among them is the Corine Landcover data, which is of interest to some CAMEO SMEs. The CORINE Land Cover inventory was initiated in 1985 (reference year 1990). Updates have been produced in 2000, 2006, 2012, and 2018. It consists of an inventory of land cover in 44 classes. The inventory also includes change layers, which highlight changes in land cover[[59]](#footnote-58). CORINE Land Cover data are derived from analysing data from Landsat, SPOT, Sentinel-2, and other satellites. The data are available free of charge from the Copernicus Land Monitoring Service. In Ireland, the Environment Protection Agency (EPA) maintains a version of the CORINE Land Cover for the country’s territory[[60]](#footnote-59). The EPA version was assessed and verified against national in-situ geospatial information on agriculture, forestry and water bodies from a national team including the Department of Agriculture, Food and the Marine, the Forest Service, Coillte and the EPA. The primary land cover and change in land cover data are produced in vector formats (i.e. ESRI GeoDatabase, SQLite Geopackage, Shapefile) with polygon topology. Derived products in raster formats (i.e. GeoTIFF, ADF) are also available[[61]](#footnote-60).

Another data source relevant to CAMEO SMEs is the Forest data series[[62]](#footnote-61), which is also part of the Copernicus Land Monitoring Service. The status products are available for the 2012 and 2015 reference years. The data include (i) tree cover density, (ii) dominant leaf type, and (iii) a forest type product. In addition to these datasets, there is a tree cover density change product for 2012-2015. The Ireland’s EPA hosts a version of the tree cover density and dominant leaf type for 2015 on its system. That version was assessed and verified by a national team[[63]](#footnote-62). The full data series is accessible on the Copernicus Land Monitoring Service. The data are delivered in GeoTIFF format.

Water and Wetness[[64]](#footnote-63) is another data product from the Copernicus Land Monitoring Service relevant to the CAMEO SMEs. The data show the occurrence of water and wet surfaces over the period from 2009 to 2018. The data were created based on interpretation of optical high-resolution satellite images and radar information. Quality assessment conducted by the EPA in Ireland showed that the overall quality of the Wetness and Water data for 2015 for Ireland is insufficient[[65]](#footnote-64). It is unclear whether there is a validated, national version of the data product for Ireland. The SMEs survey responses indicate uses of the data. The data are available in GeoTIFF format.

Natura 2000 - 2012 capturing important hotspots for nature conservation is another Copernicus thematic dataset relevant to the CAMEO SMEs. The EPA provides data quality assessments and disseminates the data for Ireland[[66]](#footnote-65). The data are delivered in Shapefile format. All EPA versions of the Copernicus data products described in this section can be downloaded by submitting requests through the EPA website[[67]](#footnote-66). Requests are processed automatically, and data are provided as static files by email.

### 2.2.3 Archived weather data from Met Eireann

In addition to the live data feeds described in Section 2.1.7, Met Eireann provides archives of past weather forecast[[68]](#footnote-67),[[69]](#footnote-68),[[70]](#footnote-69),[[71]](#footnote-70), past weather warnings[[72]](#footnote-71), past rainfall records (e.g., [[73]](#footnote-72),[[74]](#footnote-73)), and past observations at weather stations (e.g., [[75]](#footnote-74),[[76]](#footnote-75),[[77]](#footnote-76)). The data are published in ODS, CSV, and TXT formats on the Ireland’s Open Data Portal.

### 2.2.4 Elevation data products

Elevation data is another important source of data considered in CAMEO. There are several types of elevation data, including Digital Terrain Model (DTM), Digital Surface Model (DSM), and three-dimensional (3D) point cloud. A DTM is a bare earth model of the Earth’s surface without trees or buildings. A DSM is a model of the Earth’s surface in its current state and includes trees, buildings and every manmade structure. A point cloud consists of discrete sampling points of the Earth’s surface. Point clouds are often collected using LiDAR (Light Detection and Ranging), a 3D remote sensing technology, or using photogrammetry technologies (e.g. Structure from Motion). Many DEMs and DSMs are derived from point cloud data.

There are multiple sources of elevation data products in Ireland. The Geological Survey Ireland (GSI) made available on its Open Topographic Data Viewer[[78]](#footnote-77) a collection of DEMs and DSMs contributed by the Department of Culture, Heritage and the Gaeltacht, the Discovery Programme, Transport Infrastructure Ireland, New York University, the Office of Public Works, Westmeath County Council, and the GSI. The data are available for download free of charge. Notably, the dataset does not provide a full spatial coverage for the country. In addition to the GSI elevation data, there is available a nation-wide DTM for the entire country as reported by the GSI in a project on groundwater flooding[[79]](#footnote-78). The DTM, available at 5 m resolution, was compiled from best available datasets, incorporating topographic data from IfSAR NEXTMAP (OPW), LiDAR (OPW, GSI, TII), drone-based photogrammetry (GSI) and dGPS datasets (Trinity College Dublin, GSI). That DTM is not available to the public. The Ordnance Survey Ireland (OSI) also has LiDAR data captured between 2006 and 2008 that cover major cities and areas in the country. The OSI data, which have a 2 m resolution for urban areas, and 10 m resolution for rural area, are available for purchase[[80]](#footnote-79).

Most elevation data products in Ireland are available in raster formats (e.g. GeoTIFF, ESRI ASCII Grid). While many datasets are derived from point clouds, the original point clouds are rarely available. End users often have access to the derived DSMs and DTMs only. Rare examples of LiDAR point clouds available in the original format are the open aerial LiDAR datasets captured over a region in Dublin city centre in 2007 and 2015[[81]](#footnote-80),[[82]](#footnote-81). Both datasets have exclusively high resolution (i.e. 300 - 500 points/m2) and can be valuable to test the data warehouse’s capability. In addition to the public aerial LiDAR data, terrestrial LiDAR data (i.e. data collected from a ground-based platform) will be considered. Most aerial and terrestrial LiDAR data are available in LAS format. The SME survey responses indicate interest in DTM, DSM, as well as LiDAR data.

### 2.2.5 Other sources of static data

This section highlights other static datasets indicated in the SMEs survey responses.

**Metocean data**

In addition to the metocean data available from Copernicus thematic services (see 2.1.6), some SMEs indicated interest in NOAA Wave Watch III (NWW3)[[83]](#footnote-82), Ifremer’s wave hindcast database[[84]](#footnote-83), and paid metocean data from Furgo[[85]](#footnote-84) and ABPMER Seastates[[86]](#footnote-85). NWW3 are historic archives of wave hindcast and reanalysis data. The data is publicly available as a set of GRIB files on a web server[[87]](#footnote-86). Wave hindcast database from Ifremer (L'Institut Français de Recherche pour l'Exploitation de la Mer - French National Institute for Ocean Science) can be downloaded in netCDF format though a web application[[88]](#footnote-87). OGC Web Map Service and Web Coverage Service access to the data is also available.

**IFORIS**

The IFORIS (Integrated Forestry Information System) database includes a mix of vector and raster data layers[[89]](#footnote-88) from the EPA, the Forest Service of the Department of Agriculture, Food and the Marine, Ordnance Survey Ireland, Electricity Supply Board (ESB), and other sources. Most of the EPA data is publicly available via Web Map Service (e.g. waterbodies). The forest boundaries and other layers are publicly available by request to the Forest Service. Some data layers are not publicly available.

**OSI Rivers and Lakes & OSI Geometric River Network**

The vector datasets, disseminated by the EPA, consist of information of the water flow network[[90]](#footnote-89) and lake segment polygons[[91]](#footnote-90) in the country. The datasets are available free of charge and are in Shapefile and GDB formats.

**UAV images**

UAV images were mentioned in the SMEs survey responses. UAV images are often available in common image formats such as JPEG. In addition, location and orientation data may be retained to aid data analysis. Those data are often available in text formats, such as CSV, XML, and JSON.

**Field data in vector formats**

Another data source mentioned in the SMEs survey responses is point and polygon vector datasets representing sample plots, harvested data, and forest areas. Data are often in common vector formats such as Shapefile, KML, and GeoJSON.

## 2.3 Data Formats

This section describes the data formats relevant to the data sources mentioned in Sections 2.1 and 2.2. Some information in this section is derived from project AIREO (AI ready EO training datasets)[[92]](#footnote-91).

### 2.3.1 Container formats

**SAFE**

SAFE (Standard Archive Format For Europe) is a container format designed by ESA for archiving and conveying Earth observation data within the ESA archiving facilities and, potentially, within the archiving facilities of cooperating agencies[[93]](#footnote-92). This is a ZIP package containing multiple files, one of which is a manifest.xml file with metadata, others include directories with measurement data in NetCDF or TIFF format, and any associated preview images. This format was designed for the file-level transfer of data granules, ensuring all components are kept together. The use of zip-level compression means it is not possible to index into and retrieve a part of the package without processing the entire package first. That makes the format not optimised for data storage and transfer on the cloud.

**NITF**

NITF (National Imagery Transmission Format)[[94]](#footnote-93) is a file format designed by the United States’ National Geospatial-Intelligence Agency to help simplify the transfer of images and associated image support data (ISD) by reducing the number of independent files associated with a given data product. To meet this requirement, NITF was designed as a container format that can store a variety of data types together, including images, ancillary data such as shapefiles and tables, and associated ISD. By using a container format, all the data used to support a specific data product can be packed into an NITF file container. The United States Department of Defense (DoD) and Intelligence Community (IC) have standardised on NITF to ensure that data products include everything needed to support a variety of exploitation workflows.

**File Geodatabase/GDB**

File Geodatabase[[95]](#footnote-94) is a proprietary file-based format from ESRI. It is a member of ESRI’s GeoDB family of formats for geodatabase, the primary data storage model for ArcGIS software products and services. A file geodatabase is a collection of files in a folder on disk that can store, query, and manage many different types of spatial and nonspatial data.

**SQLite GeoPackage**

A GeoPackage[[96]](#footnote-95) is an extended SQLite database file (\*.gpkg) containing data and metadata tables with specified definitions, integrity assertions, format limitations and content constraints. The GeoPackage standard describes a set of conventions for storing vector features, tile matrix sets of imagery and raster maps at various scales, schema and metadata. A GeoPackage can be extended by using the extension rules as defined in Clause 2.3 of the standard. Additional extensions may be added by following the rules for GeoPackage extensions, however doing so can impact interoperability.

GeoPackage is a relatively new format from the OGC standards consortium. It was designed to be as lightweight as possible and be contained in one ready-to-use single file. This makes it suitable for rapid sharing on cloud storage, USB drives, etc. It is open source and platform independent.

It can contain both vector and raster data, including tile image “pyramids” in JPG and PNG formats. It could in principle replace all existing formats, though most usage to date appears to be as a vector format. Its RDBMs specification is compatible with the OGC Features API, giving a natural “cloud compatibility”. For stacks using GDAL, its Virtual Filesystem layer allows access without download for network-hosted geopackages.

For vector features, descriptions of the features may be defined including free-text and raster/photo “blobs”. No attribute or extensions currently exist for quality information.

**HDF5**

HDF5 (Hierarchical Data Format Version 5)[[97]](#footnote-96) is an extensible data format for self-describing files. The format supports an unlimited variety of data types and is designed to be flexible and efficient for large and complex data. HDF5 is extensible via customising data types, allowing communities and their applications to evolve in the use of HDF5. HDF supports both C and Fortran interfaces, and it has been successfully ported to a wide variety of machine architectures and operating systems. HDF emphasises a single common format for data, on which many interfaces can be built.

### 2.3.2 Imagery/raster data formats

**TIFF, GeoTIFF and COG**

TIFF (Tag Image File Format)[[98]](#footnote-97) files are containers for a number of images. The Image File Directory describes the data and has a set of tags which contain metadata. A GeoTIFF[[99]](#footnote-98) includes a set of specific tags which describe the georeferencing of the data. Cloud Optimised GeoTIFFs (COGs)[[100]](#footnote-99) are GeoTIFFs which are organised in tile format so that any particular tile can be read without reading the entire file, for example using HTTP GET range requests that allow data to be fetched within certain byte ranges. This allows for fast analysis and rendering of specific areas of large images.

**JPEG 2000**

JPEG 2000 is a common image compression standard based on discrete wavelet transform. Imagery can be coded at different levels of quality and information loss. The standard supports very large image size (up to thousands of terapixels) and precisions as high as 38 bits/sample, with or without tiling, and with a variety of interchangeable data progressions and random-access capabilities.

**ASCII Grid and ADF**

ASCII Grid[[101]](#footnote-100) is a textual data format published by ESRI for exchanging cell-based or raster data. An ASCII Grid file begins with header information that defines the properties of the raster such as the cell size, the number of rows and columns, and the coordinates of the origin of the raster. The header information is followed by cell value information specified in space-delimited row-major order, with each row separated by a carriage return. The file format is simple but is not storage efficient.

ADF[[102]](#footnote-101) is a proprietary raster data file format used by ESRI software applications. Spatial data is stored as a binary grid that is native to ESRI. The grid is made up of rows and columns of cells that can be integer as well as floating point. Integer grids in an ADF file represent discrete data and floating-point grids represent continuous data.

### 2.3.3 Vector data formats

**GeoJSON**

GeoJSON[[103]](#footnote-102) is a vector format for encoding geographic information and their nonspatial attributes based on JSON. It includes:

* Geometry object: can be a point, line, or polygon and gives the location information.
* Feature object: is the geometry object and the associated data.
* FeatureCollection: a list of feature objects.

**Shapefile**

Shapefile[[104]](#footnote-103) is a vector format that stores data as primitive geometric shapes such as points, lines, and polygons. These shapes are linked with data attributes to create representation of the geographic data. The term “shapefile” is quite common, but the format consists of a collection of files with a common filename prefix, stored in the same directory. The three mandatory files have filename extensions .shp, .shx, and .dbf. The actual shapefile relates specifically to the .shp file, but alone is incomplete for distribution as the other supporting files are required.

Shapefile is a proprietary format, designed and published by ESRI very early in the development of geospatial tools and formats. Though it is very popular as a transfer format, there are certain key issues to consider with regards to attributes in Shapefile. For example, they cannot store null values, they round up numbers, they have poor support for Unicode character strings, they do not allow field names longer than 10 characters, and they cannot store time in a date field. Additionally, they do not support capabilities found in geodatabases, such as domains and subtypes.

**KML and KMZ**

KML (Keyhole Markup Language)[[105]](#footnote-104) is an XML notation for expressing geographic annotation and visualisation within two-dimensional maps and three-dimensional Earth browsers (e.g. Google Earth). KML files can include both vector and raster data. KMZ provides zip-level lossless compression for KML data.

### 2.3.4 Data formats for array-oriented data

**NetCDF**

NetCDF (Network Common Data Form)[[106]](#footnote-105) is a set of software libraries and self-describing, machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data. “Self-describing” means that there is a header which describes the layout of the rest of the file, in particular the data arrays, as well as arbitrary file metadata in the form of name/value attributes. The format is platform independent, with issues such as endianness being addressed in the software libraries. The data are stored in a fashion that allows efficient subsetting and efficient access in IO-intensive bulk and parallel processing. NetCDF is a common data format among the meteorological and climate science communities.

**ODS, XLS, and CSV/TSV**

ODS (OpenDocument Spreadsheet Document)[[107]](#footnote-106), XLS (Microsoft Excel Binary File format)[[108]](#footnote-107), CSV (Comma-Separated Values)[[109]](#footnote-108), and TSV (Tab-Separated Values)[[110]](#footnote-109) are common formats for storing tabular data.

ODS is an open file format for spreadsheets, charts, presentations and word processing documents using ZIP-compressed XML files. XLS is Microsoft’s file format for its Excel spreadsheet application. An XLS file is a collection of records and structures that specify workbook content, which can include unstructured or semi-structured tables of numbers, text, or both numbers and text, formulas, external data connections, charts, and images. CSV and TSV are text formats that allow rectangular array data to be saved in a table structured format. Array data can contain numeric and textual values.

**GRIB**

GRIdded Binary (GRIB)[[111]](#footnote-110) is a standard file format for the storage and transport of gridded meteorological data, including historical and forecasted data. The file format standard was designed and is maintained by the World Meteorological Organisation.

### 2.3.4 Point cloud data formats

**LAS and LAZ**

LAS (LASer)[[112]](#footnote-111) is an open binary file format intended for point cloud data records. LAS was originally developed for storing point clouds derived from aerial LiDAR data sensors. However, the use of the format has extended beyond the original intended purpose. Today, LAS is a common format for point clouds derived from aerial, terrestrial LiDAR data, as well as photogrammetry point clouds. LAZ[[113]](#footnote-112) is a lossless compression format for LAS data. LAZ offers a high level of compression (3-10 times) without information loss.

**Cloud-Optimised Point Cloud**

Cloud-Optimised Point Cloud (COPC)[[114]](#footnote-113) is a very recent effort on optimising LAZ data for access on the cloud. The current optimisation strategy is available in the form of an octree index structure stored in the Extended Variable Length Record near or at the end of a LAS/LAZ file. The octree contains pointers to the point cloud data organised as variable-length LAZ chunks. COPC allows data in a file to be consumed sequentially or partially.

### 2.3.5 Text-based formats for documenting metadata, annotation and other purposes

**JSON**

JSON (JavaScript Object Notation)[[115]](#footnote-114) is a syntax for storing and exchanging data in text, with JavaScript syntax. It is an open standard file format, and data interchange format, that uses human-readable text to store and transmit data objects consisting of attribute–value pairs and array data types (or any other serializable value).

**XML**

XML (Extended Markup Language)[[116]](#footnote-115) is a markup language and file format for storing, transmitting, and reconstructing arbitrary textual data. It defines a set of rules for encoding documents in a format that is both human-readable and machine-readable. The format is popular for data exchange on the web.

## 2.4 Data Processing Tools

This section presents a summary of some selected data pre-processing tools potentially useful for preparing data for ingestion and for manipulating data inside the data warehouse.

### 2.4.1 GDAL (Geospatial Data Abstraction Library)

GDAL[[117]](#footnote-116) is a translator library for raster and vector geospatial data formats and released under MIT style open-source Licence by open source geospatial foundation. The library presents a single raster and single vector abstract data model to calling applications for all supported formats. It has various useful command line utilities for data translation and processing. GDAL comprises approximately 142 raster drivers and 84 vector drivers to support various file formats. Instead of implementing drivers for every file format, application developers only need to write a GDAL client driver. Therefore, when a new file format is proposed, there is no need to convince many application developers to support it, only a GDAL driver for the new format was required to support as a quick adoption. For big earth observation, GDAL has some limitations[[118]](#footnote-117):

* It has weak semantics because the data model does not include band wavelength i.e. colour and observation time, but addresses bands by dataset name and band number.
* Raster datasets are unable to inform whether pixels refer to points, cells to constant value or cells with an aggregated value, however, most methods silently assume the least likely option points.
* The library cannot do much processing, and clients that do the processing and use GDAL for reading and writing, need to be close to where the data is, and are usually far away from the user's computer.

### 2.4.2 Orfeo Toolbox (OTB)

OTB[[119]](#footnote-118) is an open source software for processing remote sensing images. It can process high resolution optical, radar and multispectral images at a terabyte scale. The software supports multiple image formats, vector formats, metadata formats, DEM, and KMZ image export. A wide variety of applications are available, including image manipulation, data pre-processing (e.g. radiometric and atmospheric corrections, accuracy refinement, registration, orthorectification, etc.), features extraction, image segmentation, image classification, change detection, hyperspectral processing, and SAR processing. A complete list of functionalities and algorithms is available on the project website[[120]](#footnote-119). Orfeo is integrated in several DIAS.

### 2.4.3 SNAP Toolbox

SNAP (Sentinel Application Platform)[[121]](#footnote-120) is an open-source software for Earth Observation data processing and analyses and is designed to be a common architecture for all Sentinel Toolboxes. Notable features include very fast image display and navigation, bitmask definition and overlay, band arithmetic, reprojection and orthorectification, geo-coding and rectification using ground control points. There is multithreading and multi-core processor support, and the SNAP is readily integrated with WorldWind for data visualisation atop a virtual globe environment.

### 2.4.4 GeoTools

GeoTools is an open-source Java library providing a standard compliant approach for visualisation and processing of geospatial data and web services using data structures based on OGC specifications. The library supports: (i) interfaces for spatial concepts and data structures, (ii) JTS Topology Suite Geometry, (iii) attribute, spatial and temporal filters matching the OGC Filter specification, and (iv) decoding technology. It supports a wide range of raster, vector and database formats. Support for additional data formats is possible via plugins. GeoTools implements a range of standard geospatial standards including GeoJSON, GeoTIFF, KML and OGC web services standards such as Web Feature Service, Web Map Service, Web Coverage Service, Web Processing Service, etc.

### 2.4.5 QGIS

QGIS[[122]](#footnote-121) is a popular free and open-source desktop application to create, edit, visualise, analyse and publish geospatial information. QGIS has a rich set of capabilities ranging from searching and downloading public satellite data (e.g. Sentinel and Landsat), downloading base maps, reading and writing data in a multitude of vector and raster formats, manipulating data, etc. There is also a large collection of plugins developed for QGIS by third party developers for different purposes.

### 2.4.6 ArcGIS

ArcGIS[[123]](#footnote-122) is a family of client software, server software, and online GIS services developed and maintained by ESRI. ArcGIS software are common commercial tools for GIS data management and manipulation. The software provides many different functionalities for both vector and raster data. Selected examples include instant viewing of satellite images, processing imagery data (geolocation, radiometric correction, orthorectification, and georeferencing), vector data manipulation, and publishing data on the web.

# 3 Requirements

This chapter contains the requirements for the data warehouse. Both functional and non-functional requirements are listed in this section. Note, we adhere to RFC2119[[124]](#footnote-123) with regards to best practises when we describe the requirements.

## 3.1 Functional Requirements:

* **REQ-DW-01-01:** Data Warehouse should provide APIs for EO platform/users to easily access data and services depending on their expertise level e.g., novice, advanced etc.
* **REQ-DW-01-02:** It should provide a metadata catalogue describing the data that can be accessed and separate metadata catalogue which is in cache.
* **REQ-DW-01-03:** It should provide a catalogue for services/APIs provided.
* **REQ-DW-01-04:** Data warehouse should provide search capabilities with advanced searching through filters with multiple parameters/keywords like cloud cover, location, time period, etc. based on the metadata catalogue.
* **REQ-DW-01-05:** Data warehouse should limit the number of search results.
* **REQ-DW-01-06:** Data warehouse should limitdownloadable size at user/ organisation level and track the pricing for accessing/downloading data at user/ organisation level.
* **REQ-DW-01-07:** It should provide information/metadata before pulling the data so that users can decide whether to pull or visualise, thus, managing the network bandwidth as well as pricing.
* **REQ-DW-01-08:** Data Warehouse should support different data types and formats, structured and unstructured data.
* **REQ-DW-01-09:** It should support access of data in patches as selected by the user to manage network bandwidth.
* **REQ-DW-01-10: I**t should provide a mechanism to update the metadata catalogue when new data is added in the Data Warehouse.
* **REQ-DW-01-11:** It should have a mechanism to harvest metadata from external sources like ESA catalogues and to periodically synchronise with the external source’s catalogues.
* **REQ-DW-01-12:** The system should incorporate models, ontologies and taxonomies for the classification and semantic representation of the accommodated datasets and services in the platform.
* **REQ-DW-01-13:** It should provide documentation and tutorials.
* **REQ-DW-01-14:** It should support batch data handling/ batch processing.
* **REQ-DW-01-15:** It should support real-time stream data handling.
* **REQ-DW-01-16:** It should enforce access control and authorization at user level or business/ organisation level.
* **REQ-DW-01-17:** Data warehouse should supportgrant and revoke access to specific data.
* **REQ-DW-01-18:** Data warehouse should allow or restrict data sharing or sharing of analytics results with other users.
* **REQ-DW-01-19:** It should support storing the results from analytics (at user level or business/organisation level)
* **REQ-DW-01-20:** It should provide a mechanism to track data movement and access.
* **REQ-DW-01-21:** Data Warehouse should be able to integrate other platforms and data sources e.g., government geospatial datasets, OSI, geological survey, open LiDAR datasets, census data etc.
* **REQ-DW-01-22:** It should providea workspace/working directory for users/ organisations.
  + Services to customise dataset or store intermediate result from analytics or store processed data
  + Store customised ML algorithms
  + Allow or restrict access to private data or results from analytics
  + Allow or restrict users to create, read, update, and delete data/services/users.
* **REQ-DW-01-23:** It should store ML algorithms and provide a catalogue of ML algorithms and allow or restrict access to the algorithms.
* **REQ-DW-01-24:** Data warehouse should provide a SPARQL interface to access data.
* **REQ-DW-01-25:** The data should be encrypted according to the encryption policy chosen by the user.
* **REQ-DW-01-26**: Components in the data warehouse should be implemented as docker components and should be configurable via configuration files and Kubernetes will be used as container orchestrater.
* **REQ-DW-01-27**: Data warehouse should have a metadata crawler to synchronise or update the metadata catalogue.
* **REQ-DW-01-28:** It should provide a mechanism to archive or delete data according to the retention policy.

## 3.2 Non-functional requirements:

* **REQ-DW-02-01:** Data warehouse should have security incorporated at all levels.
* **REQ-DW-02-02:** Data warehouse should be reliable with concurrent users accessing the data warehouse.
* **REQ-DW-02-03:** It should be resilient from failures, cyber-attacks etc. and easy to recover from a crash or failure.
* **REQ-DW-02-04:** Data warehouse should be scalable according to the user demands.
* **REQ-DW-02-05:** Data warehouse should be easy to manage.
* **REQ-DW-02-06:** It should be interoperable and flexible with other similar technologies and portable to any similar environment.
* **REQ-DW-02-07:** It should have minimal response time to the requests as permitted.
* **REQ-DW-02-08:** It should be always available or with minimal downtime.
* **REQ-DW-02-09:** It should adhere to the standards and rules.
* **REQ-DW-02-10:** It should support idempotent requests to handle repeated requests without problems.
* **REQ-DW-02-11:** Data warehouse should be long-term cost effective.
* **REQ-DW-02-12:** Data warehouse should be restored after a failure with ease and minimal time i.e., it should be easy to maintain.
* **REQ-DW-02-13:** It should be user-friendly and usable.
* **REQ-DW-02-14:** It should be GDPR compliant.

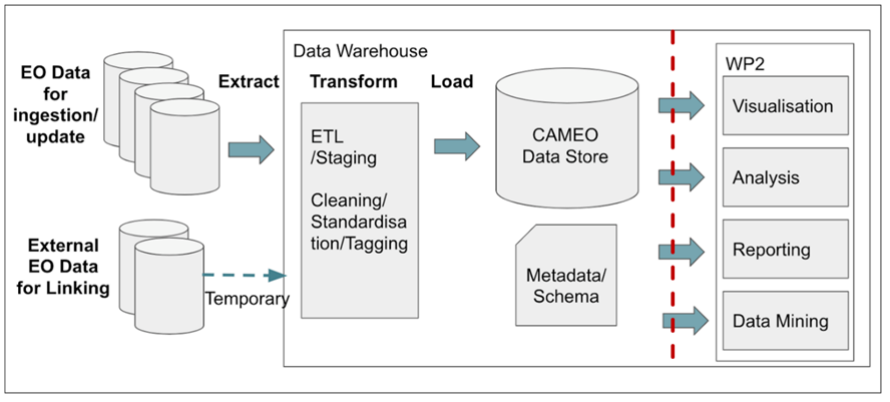
## 3.3 Summary

| Requirement Reference | Description | Priority | Dependencies on other requirements |
| --- | --- | --- | --- |
| Functional requirements | | | |
| REQ-DW-01-01 | The Data Warehouse should provide APIs for EO platform/users to easily access data and services depending on their expertise level e.g., novice, advanced etc. | High |  |
| REQ-DW-01-02 | It should provide a metadata catalogue describing the data that can be accessed and a separate metadata catalogue which is in cache. | High |  |
| REQ-DW-01-03 | It should provide a catalogue for services/APIs provided. | Medium |  |
| REQ-DW-01-04 | The Data warehouse should provide search capabilities with advanced searching through filters with multiple parameters/keywords like cloud cover, location, time period, etc. based on the metadata catalogue | High | REQ-DW-01-02 |
| REQ-DW-01-05 | The Data warehouse should limit the number of search results. | Low | REQ-DW-01-04 |
| REQ-DW-01-06 | The Data warehouse should limit downloadable size at user/organisation level and track the pricing for accessing/downloading data at user/organisation level. | Medium | REQ-DW-01-04 |
| REQ-DW-01-07 | It should provide information/metadata before pulling the data so that users can decide whether to pull or visualise, thus, managing the network bandwidth as well as pricing. | Medium |  |
| REQ-DW-01-08 | The Data Warehouse should support different data types and formats, structured and unstructured data. | High |  |
| REQ-DW-01-09 | It should support access of data in patches as selected by the user to manage network bandwidth. | Medium |  |
| REQ-DW-01-10 | **I**t should provide a mechanism to update the metadata catalogue when new data is added in the Data Warehouse. | Medium |  |
| REQ-DW-01-11 | It should have a mechanism to harvest metadata from external sources like ESA catalogues and to periodically synchronise with the external source’s catalogues. | Medium | REQ-DW-01-10 |
| REQ-DW-01-12 | The system should incorporate models, ontologies and taxonomies for the classification and semantic representation of the accommodated datasets and services in the platform. | Medium |  |
| REQ-DW-01-13 | It should provide documentation and tutorials. | Medium |  |
| REQ-DW-01-14 | It should support batch data handling/ batch processing. | High |  |
| REQ-DW-01-15 | It should support real-time stream data handling. | High |  |
| REQ-DW-01-16 | It should enforce access control and authorization at user level or business/organisation level. | High |  |
| REQ-DW-01-17 | The Data warehouse should supportgrant and revoke access to specific data. | High | REQ-DW-01-16 |
| REQ-DW-01-18 | The Data warehouse should allow or restrict data sharing or sharing of analytics results with other users. | High | REQ-DW-01-16  REQ-DW-01-17 |
| REQ-DW-01-19 | It should support storing the results from analytics (at user level or business/organisation level) | High | REQ-DW-01-16 |
| REQ-DW-01-20 | It should provide a mechanism to track data movement and access. | High |  |
| REQ-DW-01-21 | The Data Warehouse should be able to integrate other platforms and data sources e.g., government geospatial datasets, OSI, geological survey, open LiDAR datasets, census data etc. | High |  |
| REQ-DW-01-22 | It should provide a workspace/working directory for users/organisations.   * Services to customise dataset or store intermediate result from analytics or store processed data * Store customised ML algorithms * Allow or restrict access to private data or results from analytics * Allow or restrict users to create, read, update, and delete data/services/users. | Medium | REQ-DW-01-16  REQ-DW-01-17  REQ-DW-01-18  REQ-DW-01-19 |
| REQ-DW-01-23 | It should store ML algorithms and provide a catalogue of ML algorithms and allow or restrict access to the algorithms. | Medium |  |
| REQ-DW-01-24 | Data warehouse should provide a SPARQL interface to access data. | High |  |
| REQ-DW-01-25 | The data should be encrypted according to the encryption policy chosen by the user. | High |  |
| REQ-DW-01-26 | Components in the data warehouse should be implemented as docker components and should be configurable via configuration files and Kubernetes will be used as container orchestrater. | High |  |
| REQ-DW-01-27 | Data warehouse should have a metadata crawler to synchronise or update the metadata catalogue. | High |  |
| REQ-DW-01-28 | It should provide a mechanism to archive or delete data according to the retention policy. | High | REQ-DW-01-20 |
| Non-functional requirements | | | |
| REQ-DW-02-01 | Data warehouse should have security incorporated at all levels. | High | REQ-DW-01-25 |
| REQ-DW-02-02 | Data warehouse should be reliable with concurrent users accessing the data warehouse. | High |  |
| REQ-DW-02-03 | It should be resilient from failures, cyber-attacks etc. and easy to recover from a crash or failure. | High |  |
| REQ-DW-02-04 | Data warehouse should be scalable according to the user's demands. | High |  |
| REQ-DW-02-05 | Data warehouse should be easy to manage. | Medium |  |
| REQ-DW-02-06 | It should be interoperable and flexible with other similar technologies and portable to any similar environment. | High | REQ-DW-01-26  REQ-DW-01-21 |
| REQ-DW-02-07 | It should have minimal response time to the requests as permitted. | High |  |
| REQ-DW-02-08 | It should be always available or with minimal downtime. | High |  |
| REQ-DW-02-09 | It should adhere to the standards and rules. | High |  |
| REQ-DW-02-10 | It should support idempotent requests to handle repeated requests without problems. | High |  |
| REQ-DW-02-11 | Data warehouse should be long-term cost effective. | Medium |  |
| REQ-DW-02-12 | Data warehouse should be restored after a failure with ease and minimal time i.e., it should be easy to maintain. | Medium |  |
| REQ-DW-02-13 | It should be user-friendly and usable. | Medium |  |
| REQ-DW-02-14 | It should be GDPR compliant. | High | REQ-DW-01-09 |

# 4 Data Warehouse Architecture

This chapter summarises the structure and design of the data warehouse and provides references to detailed design documentation.

The following subsections describe the main design features, architecture of the application, and the technology architecture of the data warehouse. Furthermore, we also present the web application of the data warehouse which facilitates the monitoring and configuration of the data warehouse. Also covered in this chapter are the various abstraction layers used to illustrate the architecture.



*Figure 1 CAMEO Architecture Overview*

## 4.1 Main Design Features

This section provides information on the main design features of the data warehouse. These features have been derived from the requirements presented in the previous chapter and we describe these features in an implementation agnostic fashion such as not to constrain the implementers of the architecture. Figure 1 CAMEO Architecture Overview presents a high-level illustration of the CAMEO architecture. To the left-hand side of the red broken line is the Data Warehouse. The Main Design features will be described briefly using the terms in this diagram as well as references to the requirements.

### 4.1.1 EO data for ingestion

The EO data presented to the CAMEO platform is expected to be ingested from external repositories. The data formats have already been discussed in Chapter 2 "Data Sources”.

### 4.1.2 Extract

The “Extract” phase is the first phase of a pattern known as Extract-Transform-Load (ETL). An example of how this pattern can be applied to EO data is discussed here[[125]](#footnote-124). This data is expected to be ingested via the APIs provided by the EO data owner.

### 4.1.3 Transform

At this point the raw data undergoes data processing. Here, the data is transformed and consolidated for its intended analytical use case. This phase can involve the following tasks:

* Filtering, cleansing, de-duplicating, validating, and authenticating the data.
* Performing calculations, translations, or summarizations based on the raw data. This can include changing row and column headers for consistency, converting currencies or other units of measurement, editing text strings, and more.
* Conducting audits to ensure data quality and compliance
* Removing, encrypting, or protecting data governed by industry or governmental regulators
* Formatting the data into tables or joined tables to match the schema of the target data warehouse.

### 4.1.4 Load

In this last step, the transformed data is moved from the staging area into a target data warehouse. Typically, this involves an initial loading of all data, followed by periodic loading of incremental data changes and, less often, full refreshes to erase and replace data in the warehouse. For most organisations that use ETL, the process is automated, well-defined, continuous and batch-driven.

### 4.1.5 CAMEO Data Store

The data store is, conceptually, the storage technology used to store the EO data post-ETL. At this stage the specific software application which will be used as the data store has not been decided however it is expected that an open-source data store, e.g., Dell Elastic Cloud Storage (ECS) will be used. Also, of importance is Amazon S3 protocol compliance. Finally, of note, is the requirement that the data store must be agnostic.

### 4.1.6 Metadata/Schema

An appropriate schema for the data warehouse will be defined based on input from T1.1 . Given the complexity and varying nature of the data, a flexible schema is required. To aid with the development of the schema, semantic enrichment through tagging of datasets will be considered. This will add meaning and context to the data to make it more usable and searchable which is a novel aspect of CAMEO.

## 4.2 Data Warehouse Layers

To gain a greater understanding of the data warehouse architecture, it is beneficial to view the warehouse from the logical layers required to deliver the requirements. These layers are presented in Figure 2 Data Warehouse Layers and they are discussed in greater detail in this section.

Graphical user interface

Description automatically generated with medium confidence

*Figure 2 Data Warehouse Layers*

### 4.2.1 Data Warehouse Access Layer

The Data Warehouse Access (DWAL) Layer conceptually provides the primary interface to the Analytics Engine (AE). This layer also concerns itself with security issues such as ensuring the access from the AE.

### 4.2.2. Storage Abstraction Layer

The Storage Abstraction Layer (SAL) provides an abstraction to the underlying storage and supports the access from the DWAL. It facilitates the concept of storage agnostic technologies being used within the DW.

### 4.2.3. Extract Transform Load Layer

The Extract Transform Load Layer (ETLL) facilitates the ingestion of EO data from the external EO repository. It also allows for the semantification of data such that it is amenable to storage within graph-based databases if they are desired.

## 4.3 Data Warehouse Components

In this section we provide information regarding the individual logical blocks of the DW. We do not prescribe what technologies to use or what protocols should be used to communicate between the blocks. Figure 3 Data Warehouse Block Diagram illustrates the blocks required to construct the DW. We briefly describe each one in the following subsections.

*Diagram

Description automatically generated*

*Figure 3 Data Warehouse Block Diagram*

### 4.3.1 CAMEO Webpage

This component interacts with the end user of the CAMEO system. Development of this interface is outside of the scope of WP1.

### 4.3.2 Login Authorization

Control of access to EO data analytics as well as data protection is a core requirement for the CAMEO platform. This component provides this authorisation and also provides access control.

### 4.3.3 Security Layer

Data in flight is protected at this layer via encryption as well as digital cert verification between the CAMEO webpage and the CAMEO search results.

### 4.3.4 Data Quality Monitoring

This component ensures that the data being ingested meets the quality required to satisfy the user queries. Some key data dimensions[[126]](#footnote-125) are: Completeness: Is data missing or not user friendly? Timeliness: Is data available for use in the time frame in which it is expected? Conformity: Is the data conforming to expected format? Uniqueness: Is the data duplicated within the available data set? Integrity: Ensure integrity of data and its relationships along with source or lineage of the data. Is the integrity ensured? Consistency: Is there a single source of truth or are different versions for the same data entity available across multiple environments? Accuracy: Is the data accurately representing the business data as expected?

### 4.3.5 API Gateway

An API gateway is an API management tool that sits between a client and a collection of backend services. An API gateway acts as a reverse proxy to accept all application programming interface (API) calls, aggregate the various services required to fulfil them, and return the appropriate result. There are several services required by the API gateway to provide and we discuss them briefly in this subsection. Also provided by the API gateway is routing of the queries to the appropriate service API.

There are three crucial parts of a request: Type, endpoint, and Parameters.

1. The type of request (or action) will depend on what you want to do with the information (e.g., retrieve, create, update, or delete). It is done using the HTTP request methods mentioned above.
2. The URL or endpoint is available in the documentation, and this will be based on what information you want to retrieve or update. For example, the endpoint for retrieving EO data might differ from the endpoint for retrieving metadata.
3. The parameters are additional constraints or requirements that the API needs to serve your request. One of the parameters can be the API key (when authentication is required). Other parameters may include filters and additional information that you need to provide.

#### 4.3.5.1 Data Retrieval API

The data retrieval API facilitates the fetching of specific datasets based on an ID. This ID is crucial to the operation of the DW and allows for asynchronous operations and also the reuse of datasets.

#### 4.3.5.2 Metadata Retrieval API

Similar to the data retrieval API, this metadata retrieval API allows for retrieval of EO metadata. Metadata can be generated from data within the DW such that the data retrieval can be enhanced, or the search query can be improved upon.

#### 4.3.5.3 Search API

The Search API triggers the generation of a Dataset ID based on search query terms provided by the end user. Rather than return a dataset itself, it will provide a dataset ID which allows the user to retrieve the dataset when it has been generated.

### 4.3.6 Storage Abstraction Layer

The Storage Abstraction Layer (SAL) is an abstraction to the underlying storage of the DW. In a practical sense, the users of the CAMEO platform have no requirement or need to concern themselves with the practicalities of the specific storage technologies used in the DW. This allows us to provide an abstraction of the storage and determine, at a more appropriate time, the exact storage tools and software to use such as the most efficient and rich data services can be provided. In this subsection we detail the core subcomponents of the SAL.

#### 4.3.6.1 Metadata abstraction & ID allocation

A metadata repository will be developed which compiles information about available data sources. The information stored will include: the data sources available; the data contained in each source, including relevant properties; how each source can be accessed. The repository will be dynamically populated, and the software platform will provide functionality to add metadata related to newly added data sources. This repository will form the backbone of an EO data source discovery and ranking service. This repository will also support abstraction of metadata such that the metadata can be represented regardless of the underlying repository technology.

Regarding ID allocation, at this moment a specific scheme has not been decided on, it will be either a URI or UUID.[[127]](#footnote-126)

#### 4.3.6.2 Event recognition & notification

As the datasets are generated based on search queries in an asynchronous fashion there needs to exist a mechanism to notify the users when the dataset is ready for retrieval. Such a method is supported in the SAL by the event recognition & notification subcomponent. This subcomponent also supports receiving of events from other subcomponents within the SAL.

#### 4.3.6.3 Event Emitter/Receiver

Similar to the Event recognition & notification subcomponent, this subcomponent supports the registration of event listeners/receivers and also the emitting of events when they meet a specific condition, e.g., dataset ready.

#### 4.3.6.4 Storage Service

The Storage Service abstraction subcomponent defines the functions to read/write/update the data in the specific data store. This subcomponent is a modular one as it needs to support swapping in and out depending on the data store being used at the specific time.

### 4.3.7 Object Store

An object store is a method of storing data as objects. Other methods of storage use other technologies, e.g., filesystems, however in an object store the data is stored as an object which contains additional data. This data might include file size, type of file, and potentially a unique identifier. The DW will support the use of object stores however the intention is to remain agnostic as to which object store is used.

#### 4.3.7.1 Namespace

We will create a unique prefix for identifying data within the object store. This will ensure the names are unique and will also facilitate the interaction with a semantic knowledge graph.

### 4.3.8 Pre-processing

The data we receive from EO repositories and from users typically requires some manipulation to best interact with the DW components. The process of manipulating this data is referred to as pre-processing. We require a number of software systems to pre-process this data and we discuss them in this subsection.

#### 4.3.8.1 Pipeline Engine

Our pipeline engine allows us to construct complex processing pipelines in a modular fashion. This modular complex pipeline allows us to modify the pipeline with minimum impact. An example of a pipeline engine is Apache NiFi although many others exist.

#### 4.3.8.2 Additional Processes

Additional plugin modules are predicted to be required outside of the pipeline engine. Examples of these include file format detection and statistics generation.

### 4.3.9 Metadata Index

The envisaged metadata index is based on a knowledge graph. As discussed earlier, the metadata is integral to both gaining additional insights into the data contained within the DW but also it allows us to gain semantic meaning of the EO data stored in the DW. The concept of using a knowledge graph for this purpose has been previously explored[[128]](#footnote-127) and exploited in other projects and these can serve as a guide how to integrate the metadata index with the DW.

### 4.3.10 Stream/Data Ingestion Mechanism

A number of DIAS systems provide APIs to EO data. A number of questions arise as to how CAMEO should ingest this data. Should the DW poll for new data or should it wait until notified of new data? Should the DW only download relevant data, or should it download all data? Another consideration is how to access the data. For example, CREODIAS allows for CLI access as illustrated in Figure 4 CREODIAS CLI Data Access Example.



*Figure 4 CREODIAS CLI Data Access Example*

The Stream/Data Ingestion Mechanism stored the API credentials for the DW therefore the actual data download should be limited to the minimum. Fortunately, product metadata is also accessible from the DIAS which enables more precise datasets to be requested.

### 4.3.11 Metadata Crawler

The DIAS systems also provide metadata regarding the EO data which they store. To enhance the CAMEO DW it is deemed unnecessary to fetch all EO data, just that which is needed just in time. To facilitate that will be the generation of the DW metadata index which will in turn be supported by the metadata crawler component. The decision to regularly crawl the EO source repositories metadata or to do so on demand requires further investigation.

### 4.4 Tools Used

There exists literature[[129]](#footnote-128) which can be consulted for tools suitable for use within the DW. Furthermore, there are many open-source tools which can be used for the construction[[130]](#footnote-129). A sample of tools and categories are provided in Table 1 Available DW Tools.

*Table 5 Available DW Tools*

| **Tool Classification** | **Tool Name** |
| --- | --- |
| Data Type | Rasterio, GDAL, Geopandas, netCDF4 |
| Schema | IsaVix, myontology, Rej, SWOOP |
| ETL | Apache NiFi, Talend, Apache Camel, KETL |
| Data Storage | LakeFS, JuiceFS, SeaweedFS, MinIO |
| Metadata Storage | Neo4J, Apache Jena |
| Metadata Crawling | Squirrel[[131]](#footnote-130), LDSpider |
| Data Quality | Apache Griffin, OpenRefine, Talend Open Studio |

### 4.5 Standards

A number of standards bodies provide standards which are relevant to the DW and should be adhered to in the construction of the DW. The W3C provides vocabularies and ontologies which can be used to support the metadata. The Quality Assurance Framework for Earth Observation (QA4EO) was established to meet a requirement to enable interoperability and quality assessment of earth observation data. The Open Geospatial Consortium (OGC) provides royalty free, publicly available, open geospatial standards. The Research Data Alliance (RDA) provides ongoing collection and analyses on state-of-the-art data cubes and array databases in particular, including technical reports, standards, and implementations. The appropriate standards will be decided at the design phase of the DW.

### 4.6 Database design

It is expected a number of databases will exist within the DW with each one addressing a specific need for EO data and metadata storage. Examples of those include knowledge graphs for metadata storage and potentially RDBMS for data storage purposes. The precise schemas have not been finalised at this time.

### 4.7 Files

There are many different file types required to be stored within the DW, e.g., SAFE files. A number of these have been discussed in Chapter 2 Data Sources.

### 4.8 Reports

The DW is expected to provide reports on various activities and states within the DW. Examples of reports are metric reports, transaction reports, operating performance reports, and other reports as deemed necessary for the maintenance, operation, and future prediction of the requirements of the DW.

### 4.9 Error Handling

Errors are broadly classified into two types:

* Recoverable Errors - Recoverable errors are the errors that client programs can recover from to take appropriate alternate execution paths. Such errors are the result of failure to meet a particular business rule.
* Non-Recoverable errors - These are the errors that client programs cannot recover from. These kinds of errors are the result of some unexpected errors during runtime such as programming errors such null pointers, resources not available etc.

An escalation plan for non-recoverable errors needs to be put in place such that data loss or data unavailability is avoided. Recoverable errors must be logged and a report of them should be distributed to predict non-recoverable errors and to prevent performance issues with the DW.

### 4.10 Interfaces

If there are any mistakes, a description of what went wrong will be shown. Anything that deviates from the standard and expected use will be described as an error. The application's log files can be used to reduce the number of errors caused by the application.

### 4.11 Help

The “Help” component of the DW will liaise closely with WP5 to develop help techniques for the users of the DW.

### 4.12 Performance

This project's success will be determined by how well it performs. The backend must be able to process the information of all the users in a timely manner for this system to work smoothly. The database server would have to keep up with all of the requests. The connection between the frontend and the backend should be fast so that the system does not lag.

### 4.13 Security

The security component of the DW will liaise closely with WP4 which is the Security work package.

### 4.14 Reliability

Since the data of the user is the cornerstone of the project, the system must be dependable in terms of cleanly handling errors and retaining user’s data.

### 4.15 Maintainability

A number of aspects need to be addressed to ensure the DW is maintainable. These include:

* Ensuring all artefacts of the DW are version controlled.
* Ensuring all developed artefacts follow a coding standard.
* Ensuring all development flows are documented.
* Developing a test harness to ensure developed changes can be verified.
* Version controlling of all tools where possible.
* Maintenance of a software bill of materials (SBOM)

### 4.16 Portability

To ensure portability, the use of containers is strongly recommended. Furthermore, the deployment of these containers onto a container orchestrator is strongly recommended. Any containers or scripts developed should also be placed under version control.

### 4.17 Reusability

The code that was written and the components that were used should be able to be reused without difficulty. All the components would be totally reusable to everyone if time allows and clear instructions about how to make this project are written.

### 4.18 Application compatibility

Application compatibility is the task of a software lifecycle in managing the overall software portfolio used in the DW. The introduction of application incapability is a known factor in increasing cost and time in developing software and has the potential of disrupting the development of the DW. An application compatibility tool or toolkit should be employed to avoid this risk.

# 5 Conclusion

In this document, we describe the design for the data warehouse which will support the CAMEO Earth Observation platform. Furthermore, we describe the EO data formats which the data warehouse will be expected to store and provide services upon. The realisation of the data warehouse will be carried out in Task 1.4 which in turn relies on the development of the metadata repository in Task 1.3. D1.3, which is spread across three releases (D1.3.1, D1.3.2, D1.3.3) will provide the implementation.

1. <https://www.earthobservations.org/g_faq.html> [↑](#footnote-ref-0)
2. <https://www.copernicus.eu/en/about-copernicus/infrastructure-overview> [↑](#footnote-ref-1)
3. <https://sentinels.copernicus.eu/documents/247904/685154/Sentinel+Products+List-Issue1-Rev1.pdf> [↑](#footnote-ref-2)
4. <https://scihub.copernicus.eu/userguide/OpenSearchAPI> [↑](#footnote-ref-3)
5. <https://scihub.copernicus.eu/userguide/ODataAPI> [↑](#footnote-ref-4)
6. https://www.eumetsat.int/eumetsat-data-store [↑](#footnote-ref-5)
7. <https://navigator.eumetsat.int/start> [↑](#footnote-ref-6)
8. <https://www.eumetsat.int/formats> [↑](#footnote-ref-7)
9. https://www.eumetsat.int/eumetcast [↑](#footnote-ref-8)
10. https://earthexplorer.usgs.gov/ [↑](#footnote-ref-9)
11. https://www.usgs.gov/landsat-missions/landsat-data-access [↑](#footnote-ref-10)
12. https://eebulk.cr.usgs.gov/ [↑](#footnote-ref-11)
13. <https://modis.gsfc.nasa.gov/data/> [↑](#footnote-ref-12)
14. <https://earth.esa.int/eogateway/missions/worldview> [↑](#footnote-ref-13)
15. <https://earth.esa.int/eogateway/catalog/worldview-3-full-archive-and-tasking> [↑](#footnote-ref-14)
16. <https://earth.esa.int/eogateway/catalog/worldview-4-full-archive-and-tasking> [↑](#footnote-ref-15)
17. <https://docs.sentinel-hub.com/api/latest/data/maxar/world-view/> [↑](#footnote-ref-16)
18. <https://earth.esa.int/eogateway/missions/iceye> [↑](#footnote-ref-17)
19. <https://www.iceye.com/hubfs/Downloadables/ICEYE-Level-1-Product-Specs-2019.pdf> [↑](#footnote-ref-18)
20. <https://directory.eoportal.org/web/eoportal/satellite-missions/content/-/article/capella-x-sar> [↑](#footnote-ref-19)
21. <https://www.capellaspace.com/community/capella-open-data/> [↑](#footnote-ref-20)
22. <https://www.capellaspace.com/data/console-api/> [↑](#footnote-ref-21)
23. <https://earth.esa.int/eogateway/missions/planetscope> [↑](#footnote-ref-22)
24. <https://earth.esa.int/eogateway/catalog/skysat-full-archive-and-new-tasking> [↑](#footnote-ref-23)
25. <https://earth.esa.int/eogateway/missions/pleiades> [↑](#footnote-ref-24)
26. <https://earth.esa.int/eogateway/catalog/spot-6-and-7-esa-archive> [↑](#footnote-ref-25)
27. <https://earth.esa.int/eogateway/catalog/pleiades-full-archive-and-tasking> [↑](#footnote-ref-26)
28. <https://earth.esa.int/eogateway/documents/20142/37627/Pleiades-Imagery-User-Guide.pdf/a7082657-3d73-02d2-9c9a-d7b84e44dd61> [↑](#footnote-ref-27)
29. <https://www.copernicus.eu/en/copernicus-services> [↑](#footnote-ref-28)
30. <https://www.copernicus.eu/en/access-data/copernicus-services-catalogue/active-fires-european-forest-fire-information-system> [↑](#footnote-ref-29)
31. <https://resources.marine.copernicus.eu/product-detail/NORTHWESTSHELF_ANALYSIS_FORECAST_PHY_004_013/INFORMATION> [↑](#footnote-ref-30)
32. <https://www.copernicus.eu/en/accessing-data-where-and-how/copernicus-services-catalogue> [↑](#footnote-ref-31)
33. https://www.copernicus.eu/en/accessing-data-where-and-how/copernicus-services-catalogue?combine=wave&cc\_source\_service\_target\_id%5B2769%5D=2769 [↑](#footnote-ref-32)
34. <https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5> [↑](#footnote-ref-33)
35. <https://data.gov.ie/dataset/met-eireann-weather-forecast-api> [↑](#footnote-ref-34)
36. <https://data.gov.ie/dataset/met-eireann-live-text-forecast-data> [↑](#footnote-ref-35)
37. [https://data.gov.ie/dataset/weather-warnings](https://data.gov.ie/dataset/weather-warnings?package_type=dataset) [↑](#footnote-ref-36)
38. [https://data.gov.ie/dataset/todays-weather-phoenix-park](https://data.gov.ie/dataset/todays-weather-phoenix-park?package_type=dataset) [↑](#footnote-ref-37)
39. <https://data.gov.ie/dataset/yesterdays-weather-dublin-airport> [↑](#footnote-ref-38)
40. <https://data.gov.ie/dataset/rainfall-radar> [↑](#footnote-ref-39)
41. <https://data.gov.ie/dataset/weather-buoy-network-real-time-data> [↑](#footnote-ref-40)
42. <https://data.gov.ie/dataset/wave-buoy-network-real-time-data-30-minute> [↑](#footnote-ref-41)
43. <https://data.gov.ie/dataset/marine-institute-buoy-wave-forecast> [↑](#footnote-ref-42)
44. <https://data.gov.ie/dataset/smartbay-ireland-galway-bay-buoy-wave> [↑](#footnote-ref-43)
45. <https://data.gov.ie/dataset/smartbay-ireland-cork-buoy-wave> [↑](#footnote-ref-44)
46. <https://data.gov.ie/dataset/weather-buoy-network> [↑](#footnote-ref-45)
47. <https://erddap.marine.ie/erddap/tabledap/IWBNetwork.html> [↑](#footnote-ref-46)
48. <https://data.gov.ie/dataset/wave-buoy-network> [↑](#footnote-ref-47)
49. <http://data.marine.ie/geonetwork/srv/eng/catalog.search#/metadata/ie.marine.data:dataset.2783> [↑](#footnote-ref-48)
50. <https://www.smartbay.ie/> [↑](#footnote-ref-49)
51. <https://data.gov.ie/dataset/galway-bay-observatory-hydrophone-processed-data> [↑](#footnote-ref-50)
52. <https://data.gov.ie/dataset/smartbay-observatory-video-camera-data-raw> [↑](#footnote-ref-51)
53. <https://data.gov.ie/dataset/irish-national-tide-gauge-network> [↑](#footnote-ref-52)
54. <https://data.gov.ie/dataset/north-east-atlantic-real-time-bottom-current> [↑](#footnote-ref-53)
55. <https://data.gov.ie/dataset/north-east-atlantic-real-time-sea-surface-temperature> [↑](#footnote-ref-54)
56. <https://data.gov.ie/dataset/north-east-atlantic-real-time-sea-surface-salinity> [↑](#footnote-ref-55)
57. <https://data.gov.ie/dataset/opw-hydrometric-network-real-time-water-level-data> [↑](#footnote-ref-56)
58. <https://data.gov.ie/dataset/air-pollution-monitoring-data-dublin-city> [↑](#footnote-ref-57)
59. <https://land.copernicus.eu/pan-european/corine-land-cover> [↑](#footnote-ref-58)
60. <https://gis.epa.ie/geonetwork/srv/api/records/fb5d2fa9-95fe-4d3f-8aed-e548348a40ea> [↑](#footnote-ref-59)
61. <https://land.copernicus.eu/user-corner/technical-library/clc-product-user-manual> [↑](#footnote-ref-60)
62. <https://land.copernicus.eu/pan-european/high-resolution-layers/forests> [↑](#footnote-ref-61)
63. <https://gis.epa.ie/geonetwork/srv/api/records/fbe8afe3-77e1-48f1-ac77-218f568f6f39> [↑](#footnote-ref-62)
64. <https://land.copernicus.eu/pan-european/high-resolution-layers/water-wetness> [↑](#footnote-ref-63)
65. <https://gis.epa.ie/geonetwork/srv/api/records/26ae46e4-a1d1-47b2-a5b0-52da004396fa> [↑](#footnote-ref-64)
66. <https://gis.epa.ie/geonetwork/srv/api/records/d2ba6ddb-bf00-4d23-aed5-17854ae7aa61> [↑](#footnote-ref-65)
67. <https://gis.epa.ie/GetData/Download> [↑](#footnote-ref-66)
68. [https://data.gov.ie/dataset/archived-national-forecast](https://data.gov.ie/dataset/archived-national-forecast?package_type=dataset) [↑](#footnote-ref-67)
69. [https://data.gov.ie/dataset/archived-regional-forecasts](https://data.gov.ie/dataset/archived-regional-forecasts?package_type=dataset) [↑](#footnote-ref-68)
70. <https://data.gov.ie/dataset/archived-sea-area-forecasts> [↑](#footnote-ref-69)
71. [https://data.gov.ie/dataset/archived-inland-lakes-forecasts](https://data.gov.ie/dataset/archived-inland-lakes-forecasts?package_type=dataset) [↑](#footnote-ref-70)
72. <https://data.gov.ie/dataset/archived-weather-warnings> [↑](#footnote-ref-71)
73. <https://data.gov.ie/dataset/granard-ballymore-rainfall-data> [↑](#footnote-ref-72)
74. <https://data.gov.ie/dataset/newport-coole-rainfall-data> [↑](#footnote-ref-73)
75. [https://data.gov.ie/dataset/dublin-airport-monthly-data](https://data.gov.ie/dataset/dublin-airport-monthly-data?package_type=dataset) [↑](#footnote-ref-74)
76. <https://data.gov.ie/dataset/dublin-airport-hourly-data> [↑](#footnote-ref-75)
77. <https://data.gov.ie/dataset/dublin-airport-daily-data> [↑](#footnote-ref-76)
78. <https://dcenr.maps.arcgis.com/apps/webappviewer/index.html?id=b7c4b0e763964070ad69bf8c1572c9f5> [↑](#footnote-ref-77)
79. <https://www.gsi.ie/en-ie/publications/Pages/GWFlood-Project-Monitoring-Modelling-and-Mapping-Karst-Groundwater-Flooding-in-Ireland.aspx> [↑](#footnote-ref-78)
80. <https://osi.ie/wp-content/uploads/2015/05/Lidar_prod_overview.pdf> [↑](#footnote-ref-79)
81. <https://digital.ucd.ie/view/ucdlib:30462> [↑](#footnote-ref-80)
82. <https://geo.nyu.edu/catalog/nyu_2451_38684> [↑](#footnote-ref-81)
83. <https://polar.ncep.noaa.gov/waves/hindcasts/> [↑](#footnote-ref-82)
84. <https://sextant.ifremer.fr/geonetwork/srv/api/records/c87f6f24-63b4-46ec-b40e-f185a61dc672> [↑](#footnote-ref-83)
85. <https://www.fugro.com/our-services/marine-asset-integrity/monitoring-and-forecasting/metocean-modelling-and-analysis> [↑](#footnote-ref-84)
86. <https://www.seastates.net/explore-data/> [↑](#footnote-ref-85)
87. <https://polar.ncep.noaa.gov/waves/hindcasts/nww3/> [↑](#footnote-ref-86)
88. <https://tds3.ifremer.fr/thredds/ncss/IOWAGA-GLOBAL_ECMWF-WW3-HINDCAST_FULL_TIME_SERIE/dataset.html> [↑](#footnote-ref-87)
89. <https://irishriverproject.com/2022/03/16/detailed-schema-for-iforis-database-forestry-service> [↑](#footnote-ref-88)
90. <https://gis.epa.ie/geonetwork/srv/eng/catalog.search#/metadata/c4043e19-38ec-4120-a588-8cd01ac94a9c> [↑](#footnote-ref-89)
91. <https://gis.epa.ie/geonetwork/srv/eng/catalog.search#/metadata/c4040e19-38ec-4120-a588-8cd01ac94a9c> [↑](#footnote-ref-90)
92. <https://www.aireo.net/> [↑](#footnote-ref-91)
93. <https://earth.esa.int/eogateway/activities/safe-the-standard-archive-format-for-europe> [↑](#footnote-ref-92)
94. <https://nsgreg.nga.mil/doc/view?i=2063> [↑](#footnote-ref-93)
95. <https://pro.arcgis.com/en/pro-app/latest/help/data/geodatabases/manage-file-gdb/file-geodatabases.htm> [↑](#footnote-ref-94)
96. <https://www.geopackage.org/spec/> [↑](#footnote-ref-95)
97. <http://docs.opengeospatial.org/is/18-043r3/18-043r3.html> [↑](#footnote-ref-96)
98. <https://www.awaresystems.be/imaging/tiff/specification/TIFF6.pdf> [↑](#footnote-ref-97)
99. <http://docs.opengeospatial.org/is/19-008r4/19-008r4.html> [↑](#footnote-ref-98)
100. <https://github.com/cogeotiff/cog-spec/blob/master/spec.md> [↑](#footnote-ref-99)
101. <https://desktop.arcgis.com/en/arcmap/latest/manage-data/raster-and-images/esri-ascii-raster-format.htm> [↑](#footnote-ref-100)
102. <https://docs.fileformat.com/gis/adf/> [↑](#footnote-ref-101)
103. <https://tools.ietf.org/html/rfc7946> [↑](#footnote-ref-102)
104. <https://www.esri.com/library/whitepapers/pdfs/shapefile.pdf> [↑](#footnote-ref-103)
105. <https://www.ogc.org/standards/kml> [↑](#footnote-ref-104)
106. <https://www.ogc.org/standards/netcdf> [↑](#footnote-ref-105)
107. <https://www.iso.org/standard/66363.html> [↑](#footnote-ref-106)
108. <https://docs.microsoft.com/en-us/openspecs/office_file_formats/ms-xls/2aa787a8-0af8-4279-b1fd-9f210970bb55> [↑](#footnote-ref-107)
109. <https://www.loc.gov/preservation/digital/formats/fdd/fdd000323.shtml> [↑](#footnote-ref-108)
110. <https://www.loc.gov/preservation/digital/formats/fdd/fdd000533.shtml> [↑](#footnote-ref-109)
111. <https://weather.gc.ca/grib/what_is_GRIB_e.html> [↑](#footnote-ref-110)
112. <https://www.asprs.org/wp-content/uploads/2010/12/LAS_1_4_r13.pdf> [↑](#footnote-ref-111)
113. <https://laszip.org/> [↑](#footnote-ref-112)
114. <https://copc.io/> [↑](#footnote-ref-113)
115. <http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf> [↑](#footnote-ref-114)
116. <https://www.w3.org/XML/> [↑](#footnote-ref-115)
117. <https://gdal.org/> [↑](#footnote-ref-116)
118. <https://r-spatial.org/2016/11/29/openeo.html> [↑](#footnote-ref-117)
119. <https://www.orfeo-toolbox.org/> [↑](#footnote-ref-118)
120. <https://www.orfeo-toolbox.org/features-2/> [↑](#footnote-ref-119)
121. <https://step.esa.int/main/toolboxes/snap/> [↑](#footnote-ref-120)
122. [https://www.qgis.org](https://www.qgis.org/en/site/) [↑](#footnote-ref-121)
123. <https://www.esri.com/en-us/arcgis/products/index> [↑](#footnote-ref-122)
124. https://datatracker.ietf.org/doc/html/rfc2119 [↑](#footnote-ref-123)
125. Semlali, B.E.B., El Amrani, C. and Ortiz, G., 2020. SAT-ETL-Integrator: an extract-transform-load software for satellite big data ingestion. *Journal of Applied Remote Sensing*, *14*(1), p.018501. [↑](#footnote-ref-124)
126. Juneja, A. and Das, N.N., 2019, February. Big data quality framework: Pre-processing data in weather monitoring application. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)* (pp. 559-563). IEEE. [↑](#footnote-ref-125)
127. Kindel, Charlie. "The uuid: URI scheme." *URL Working Group* 5 (1997). [↑](#footnote-ref-126)
128. Aidan O Mahony, Alan Barnett, & Merry Globin. (2021). Using automotive property graph-based data models in a knowledge graph. *The 1st workshop on Squaring the circle on graphs (SCG2021),* Amsterdam. Zenodo. https://doi.org/10.5281/zenodo.5593219 [↑](#footnote-ref-127)
129. Yang, Qishan, Mouzhi Ge, and Markus Helfert. "Analysis of data warehouse architectures: modeling and classification." (2019). [↑](#footnote-ref-128)
130. Raghunath, A., Zou, Y. and Chagam, A., 2020, December. On Evolving Software Defined Storage Architecture. In *2020 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)* (pp. 57-64). IEEE. [↑](#footnote-ref-129)
131. https://github.com/dice-group/squirrel [↑](#footnote-ref-130)