This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101059548.

OPEN EARTH MONITOR

D2.6 Report FAIR operation, validation and analysis guidelines

1st version

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101059548.

Document control page

Project	Open-Earth-Monitor (OEMC)
Project, full title	A cyberinfrastructure to accelerate uptake of environmental information and help build user communities at European and global levels
Project number	101059548
Project start	June 1st 2022
Deliverable	Report "FAIR operation, validation and analysis guidelines" 1st version
Work Package	WP2 – User-driven system design and FAIR workflows
Document title	D2.6 FAIR operation, validation and analysis guidelines - 1st version
Version	0.1
Responsible author	Joan Masó (CREAF)
Contributors	Joan Masó (CREAF), Imma Serra (CREAF), Didac Pardell Sangrà (CREAF)
Date of delivery	31th July 2023
Due date of deliverable	31th July 2023
Type	Report
Language	English
Rights	CC-BY
Status	( ) In progress (x) For review ( ) Approved
Dissemination level	Public

Version history			
Version	**Implemented by**	**Date**	**Description**
0.1	CREAF	18 July 2023	Review version
1.0	CREAF	31 July 2023	Final version

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101059548.

Table of contents

Figures

Tables

Acronyms

CSW	Catalogue Service for the Web
COG	Cloud Optimised GeoTiff
DMP	Data Management Plan
EV	Essential Variable
GEO	Group for Earth Observation
GEOSS	Global Earth Observation System of Systems
FAIR	Findable, Accessible, Interoperable and Reusable
HTTPS	Hypertext Transfer Protocol Secure
STAC	SpatioTemporal Asset Catalog
TRUST	Transparency, Responsibility, User focus, Sustainability and Technology

Executive summary

The first version of Open-Earth-Monitor FAIR Operation, Validation and Analysis Guidelines gives guidelines defining data formats, metadata standards, communication protocols and implementation tools aligned with the D1.3 Data Management Plan (DMP). For each of the previous subjects this document focuses on some main representative examples aiming to check their capability to be FAIR and contribute to decide which is the best to apply in the project. After analysing four formats the document concludes that the COG is the format that provides better support to FAIR principles complying with 7/15 sub-principles. From the catalogue services reviewed, Zenodo complies 8½/15 sub-principles. From communication protocols studied the OGC API Records has the largest compliance with 9½/15 sub-principles.
A summary of this analysis is shown in Table 1: Summary table of formats, catalogues and communication protocols.

The implementation tool provided in the e-shape[1] (EuroGEO Showcases: Applications Powered by Europe) project, called Data Management Self-assessment Tool, Menard, L., & Fichaux, N. (2022)[2] is considered appropriate to be applied to the datasets provided by this project to ensure the implementation of FAIR principles [1].

A second and final version of this deliverable will be published in month 41 as an updated version including the description of the work completed in Task 2.2 and Task 2.7 and according to the project partners and stakeholders' feedback.

1. Introduction

The aim of this document is to assess the diverse subjects proposed in D 1.3 Data Management Plan according to its FAIR capabilities.
The second section describes some data formats selected by this project, in respect of how they meet with FAIR requirements. These requirements are described in the Deliverable 2.2 the Report "Status and prospect for European environmental data" 1st version [2].
The third section provides catalogues and repositories, there is an example of each one that is considered to be a good tool to help manage the project.
The fourth section describes communications protocols which are introduced to explain how the new OGC APIs apply to the FAIR principles and semantic interoperability.
The last part of the document provides some implementation tools to define the validation and analysis strategies to ensure the implementation of FAIR principles in the project.

[1] https://e-shape.eu/
[2] Data management self-assessment tool https://gkhub.earthobservations.org/records/0ksgt-7v316

2. Data Formats

2.1 COG

The Cloud Optimised GeoTIFF[3] (COG) is a particular way of using the popular image TIFF format or the extension for long files BigTIFF.
COG conforms to a subset of TIFF v6.0 or BigTIFF formats specifications. In COG, data needs to be structured in tiles and not in stripes. COG also conforms to HTTP (or HTTPS) protocol and exploits the GET range request capacity that adds the capability to request fragment of a file instead of the full dataset to maximise efficiency in the web. In addition COG makes use of the GeoTIFF extension to store some metadata, mainly about the Coordinate Reference System of the data.

Because of these functional features, COG respects FAIR principles described as follows:

a. “F2. Data is described with rich metadata”: COG uses metadata from GeoTIFF for sorting relevant geospatial metadata related to tiled original resolution and tiled reduced-resolution subfiles; also called overviews.
b. “F3. Metadata clearly and explicitly include the identifier of the data they describe”: Each TIFF file contains at same time data and metadata. Then there are no needs to have identifiers because both metadata and data are in the same file.
c. “A1. (Meta)data are retrievable by their identifier using a standardised communication protocol” and “A1.1. The protocol is open, free and universally implementable”: HTTP protocol is used without doing modification on it.
d. "A1.2 The protocol allows for an authentication and authorization procedure, where necessary": The possibility of using HTTPS protocol allows for authentication and authorization in a secure environment.
e. “I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation”: Because it is based in GeoTIFF format respects interoperability principle.
f. "F4. (Meta)data are registered or indexed in a searchable resource": The combination of COG and STAC allows for the discoverability of COGs in a web environment.

In that respect, the project has worked with the Open Geospatial Consortium to standardise the COG practices in an OGC international standard. The standard was approved and now, since very recently, it has been approved for public release on the OGC website.

[3] https://github.com/cogeotiff/cog-spec/blob/master/spec.md

2.2 Zarr

The Zarr format was inspired by the HDF5 format that is used to store chunks of data in N-dimensional arrays (also known as "cubes"). It is characterised by allowing storing data from distributed objects, high performance I/O and parallel computing. Is it open source and open developed in GitHub and received Chan-Zuckerberg Essential Open Source Software grant. The core format is based on a serialisation of the x-Array Python data structure.

The current format specification is v.2 and it is worth noting that there is no reference to geospatial concepts. Currently the community is working on version 3, which is going to accept specification extensions and it will be submitted to OGC for approval as an international standard.

GeoZarr, a variety of Zarr format, is going to formalise geospatial related concepts and specify georeferenced earth observation data stored into a Zarr cube. OGC has accepted Zarr as a community standard because of its fast acceptance in the geospatial field. This good acceptance is because Zarr can represent a huge amount of data in a simple, scalable and cloud object storage compatible.

Zarr internal organisation is formed by stores that are made of zgroups, zarrays, zattributs and data chunks structured in different paths directories:

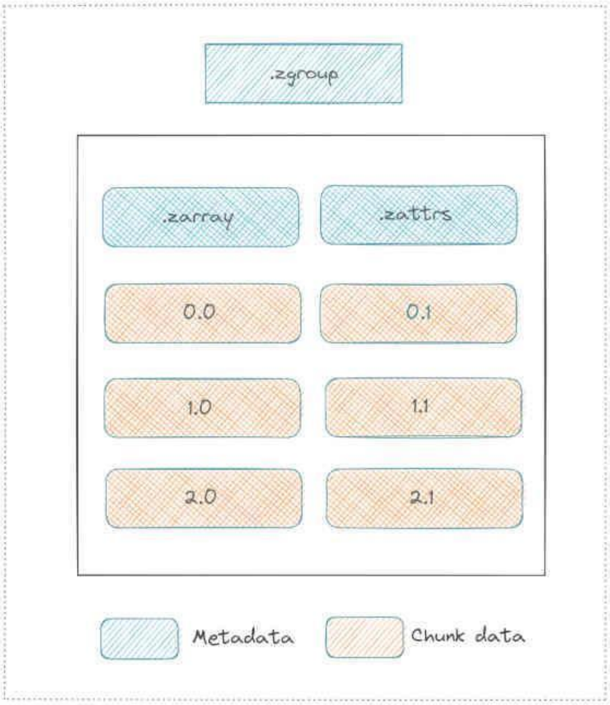

Figure 1: Files and directories scheme from Zarr format

The metadata is encoded in JSON and saved as a value of “.zarray” key into the array store.

Chunks are compressed with no additions. Each of them receives a key base on the grid index it occupies. The index is converted to strings and separated with “.”.
If an array dimension is different from chunk dimensions data falling apart is undefined.

Store locations: Each chunk once stored in an array store is bound with a key based on the logical path of the store array. Example: “foo/bar/.zarray” for metadata, “foo/bar/.zattrs” for user-defined attributes and “foo/bar/0.0”, “foo/bar/0.1”, etc for each chunk.

Zarr conforms to FAIR principles in the following way:

a. “F2. Data are described with rich metadata (defined by R1 below)”: There are some mandatory and optional metadata names to describe data exhaustively.
b. “F3. Metadata clearly and explicitly include the identifier of the data they describe”. Metadata file and the data (chunks) files are separate files; its hierarchy specifies the storage of metadata file and chunks in the same root directory. Therefore, we assume that metadata and data are mutually related, so identified.
c. “A1. (Meta)data are retrievable by their identifier using a standardised communication protocol”. Zarr benefits from the HTTP range request protocol to retrieve only a part of the information. Because of the use of HTTP protocol, then A1.1 is also complying.

2.3 Geoparquet

The origins of GeoParquet can be found in the Apache Parquet[4], a columnar data format designed for efficient data storage and adding geospatial schema specification. Currently, it is in development to become an OGC standard as an Encoding Standard for “cloud-native vector” data (point, lines and polygons). The documentation[5] provides specifications for GeoParquet tools and libraries that are implemented.

The format supports structured data columns which makes it possible to store data and metadata. The key components of GeoParquet are the following:

- Geometry columns
- Metadata
- File metadata

[4] https://parquet.apache.org
[5] https://geoparquet.org/

The specification, the version of the GeoParquet file, at the time of writing this document, is 1.0.0-beta. It describes how geospatial data should be stored in Parquet format, and the representation of geometries and the required additional metadata.

The GeoParquet structure enables interoperability between any two systems that read or write spatial data in Parquet format. This means that Geometry columns are stored in Well-Known Binary (WKB) format that can store metadata joined to the data. It contributes to promoting reproducibility in addition. With the components Metadata and File metadata, the format ensures compliance with the FAIR sub-principle "F2. Data is described with rich metadata". This format complies with the sub-principle “A1.1. The protocol is open, free and universally implementable”. As the column metadata includes a Coordinate Reference System (CRS) as an optional parameter, represented by PROJJSON[6], a JSON encoding of a Well Known Text (WKT). This will ensure conformity with the FAIR sub-principle "I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation". It supports multiple spatial reference systems.

2.4 GeoJSON

GeoJSON is an open standard format designed for representing simple geographical features, along with their non-spatial attributes. The features include points (therefore addresses and locations), line strings (therefore streets, highways and boundaries), polygons (countries, provinces, tracts of land), and multi-part collections of these types. The GeoJSON is a formal standard in the IETF. The standard specifies the way a collection of geospatial features are encoded in the JSON format. While it was not originally designed for tiles, some implementations use it to encode feature based tile data. The standard does not define how to include metadata or semantics in the same file.

Due to the limitations of the format, the only FAIR sub-principle that the format complies with "A1. (Meta)data are retrievable by their identifier using a standardised communications protocol". That is because GeoJSON can be retrieved from the web using HTTP and HTTPS using the media type application/geo+json. Indirectly, it complies with the sub-principle "A1.1 The protocol is open, free, and universally implementable". When using HTTPS, the authentication and authorization is possible and then GeoJSON could comply with the sub-principle "A1.2 The protocol allows for an authentication and authorisation procedure, where necessary".

[6] https://proj.org/en/9.2/specifications/projjson.html

3. Catalogues and repositories

3.1 STAC

Spatial Temporal Asset Catalog (STAC) is a service specification for catalogues of geospatial data (mainly used for remote sensing imagery), enabling it to be easily searchable and queryable. The STAC is composed of four components: items, catalogues, collections and the STAC API. Each one can be used independently. However, the four components work better in combination.

- STAC Item is the core unit that represents a single spatiotemporal asset as a GeoJSON feature with datetime and reference links.
- STAC Catalog provides a structure and organises the metadata like STAC items, collections and other catalogues, defined in a JSON file.
- STAC Collection extends the STAC Catalog adding information (licence, keywords, providers...) in order to link the STAC items with the Collection.
- STAC API provides a RESTful endpoint enabling search of STAC Items, specified in OpenAPI, following OGC's WFS 3.

With this specification, the aim is to gather data across multiple satellite providers in a standardised way, avoiding access through each API.

STAC through its components, conforms with the following sub-principles:

According to the specification[7], a STAC item provides a Identifier that should be unique within the Collection[8] that contains the Item and in the same way, the Collection identifier is unique globally, so, this ensures compliance with the FAIR sub-principle "F1. (Meta)data is assigned a globally unique and persistent identifier". The STAC Item can include additional fields to describe rich metadata, such as the time the asset represents, a thumbnail for quick browsing, assets links and relationship links to connect with other related STAC items. This makes it compliant with the "F2. Data is described with rich metadata".

In particular, STAC Collection is a subtype of the catalogue that has additional properties which provides the possibility to search queries for discovering, it conforms with the FAIR sub-principle "F4. (Meta)data are registered or indexed in a searchable resource" and it can also contain additional fields and JSON structures to describe rich metadata, as well conforming with the I1. "(Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation."

[7] https://stacspec.org/en/about/stac-spec/
[8] https://github.com/radiantearth/stac-spec/blob/master/collection-spec/collection-spec.md

It contributes to making data more accessible in a simple and reliable way. The STAC API defines a RESTful service interface for search. It dynamically generates a GeoJSON FeatureCollection of STAC Items in response to a user query conforming with the sub-principle “A1. (Meta)data are retrievable by their identifier using a standardised communications protocol”. And also makes it compliant to its sub-principles: A1.1 The protocol is open, free, and universally implementable" and "A1.2 The protocol allows for an authentication and authorisation procedure, where necessary".

3.2 Zenodo

Zenodo[9] is a catalogue and preservation facility for the wide academic content. It aims to share scientific data among the research community. Collaborating with European Commission, it promotes the FAIR principles because:

a. F1 "(Meta)data are assigned a globally unique and persistent identifier": Assigns a unique identifier to all the resources it stores.
b. F2 “Data are described with rich metadata“: The metadata are meant to allow an accurate and consistent identification of the product they belong to. They are compliant with DataCite Metadata Schema[10] and so compliant with the F2 principle.
c. F3 “Metadata clearly and explicitly include the identifier of the data they describe”: A DOI is mandatory in metadata of each record.
d. F4 “(Meta)data are registered or indexed in a searchable resource”: Metadata is indexed and searchable in Zenodo platform.
e. A1 "(Meta)data are retrievable by their identifier using a standardised communications protocol": It adopts the protocol OAI-PHM[11] that helps to spread and harvest contents (metadata) efficiently.
f. A1.1 "The protocol is open, free, and universally implementable": OIM-PHM requests are expressed as HTTP requests an open and free protocol.
g. A1.2 "The protocol allows for an authentication and authorization procedure, where necessary": Being FAIR does not mean to be compulsorily free and open. Zenodo encourages sharing as open and freeing as possible but allows uploading content under different and wide variety of access levels.
h. I1 "(Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation": All metadata is stored in a JSON-schema to ensure its understanding afterwards and allows export it to other popular formats.
i. R1 “(Meta)data are richly described with a plurality of accurate and relevant attributes”: The records have the minimum mandatory metadata set specified in DataCite and other recommended terms and other enrichments depending on each publication they could be more or less exhaustives.

[9] https://about.zenodo.org/principles/
[10] https://schema.datacite.org/
[11] https://www.openarchives.org/OAI/openarchivesprotocol.html

4. Communication Protocols

Communication standards:

One of the most successful types of standards in the OGC is the OGC web services. With the services that implement the OGC web services standard, you can give access to different kinds of data on the web. Most OGC web services provide instructions on how to build a URI that give access to the data behind the service. The URL contains an action to do and some parameters to modify the result. While perfectly functional, The OGC web services do not follow modern practices on the web. In particular they do not focus on resources but on operations. To correct that issue, the OGC has recently started to transform the OGC web services into OGC web APIs that define resources and use the HTTP methods to retrieve them.

Communication standards for finding the data:

The Catalogue Service for the Web (CSW) is an OGC web service that provides the capacity to query a collection of metadata and find the data or the services that the user requires. Deploy a CSW (e.g. a GeoNetwork instance) is a way to comply with the FAIR Findable sub-principle "F4. (Meta)data are registered or indexed in a searchable resource". CSW is compatible with Dublin Core and ISO 19115 metadata documents, compiling with principle F2 “Data are described with rich metadata” and also R1 “(Meta)data are richly described with a plurality of accurate and relevant attributes” and R1.2 “(Meta)data are associated with detailed provenance”. An interesting characteristic of the GeoNetwork is its capability to store attachments to the metadata. This provides a way to store the actual data as an attachment and link it to the distribution section of an ISO 19115, so principle F3 “Metadata clearly and explicitly include the identifier of the data they describe” is partially compliant. This ensures not only findability of the metadata but also findability of the data. This communication protocol is binded to HTTP[12] web protocol so complies with A1”(Meta)data are retrievable by their identifier using a standardised communications protocol” and A1.1 “The protocol is open, free, and universally implementable”. In the OEMC project, CSW can be used to store metadata about the in-situ data and some of the results of the pilots. The original Remote Sensing data is offered through a stack catalogue. The OGC API Records is an alternative to CSW that uses the resource oriented architecture and its two goals are, first to provide modern API patterns and encodings to facilitate further lowering the barrier to finding the existence of spatial resources on the Web, and second goal, provide metadata discovery retrieval functionality that is comparable to that of the OGC CSW standard. These two goals comply with F2, F3, F4 and R1 principles. OGC API is defined with OpenAPI[13] commonly written in YAML or JSON languages, then it also complies with the principles A.1 and A1.1. It gives a URL to each and every

[12] OGC Catalogue Services 3.0 Spec - HTTP Protocol Binding https://docs.ogc.org/is/12-176r7/12-176r7.html
[13] OpenAPI definition: https://www.openapis.org/what-is-openapi

metadata/data record stored in the catalogue making it compliant with the FAIR Findable sub-principle "F1. (Meta)data are assigned a globally unique and persistent identifier". The OGC API Records is still on draft phase and the authors are making efforts to make it compatible with STAC. Even if it could be desirable in many cases, in the CSW or in OGC API records there is no obligation to have a metadata record associated to dataset. This can be useful for preservation purposes, when the dataset may not be available, so that this ensures compatibility with the FAIR Accessible sub-principle "A2. Metadata are accessible, even when the data are no longer available". Because all OCG API Records search cases are using standard HTTP also OAuth could be used to authenticate and secure them, so is compliant with the sub-principle “A1.2 .The protocol allows for an authentication and authorisation procedure, where necessary”.

Communication standards for accessing data:

The OGC Web Feature Service (WFS) and the Web Coverage Service (WCS) give access to the feature data or coverage data independently of the data model and data schema of the data. Implementations of these services are based on Open Standards that can be implemented for free. This complies with the FAIR Accessible sub-principle "A1.1 The protocol is open, free, and universally implementable". It is possible to get the whole resource or a subset of it based on spatial or thematic queries. This precise queries can be achieved thanks to constraints defined by the client on features properties, so here we found compliant with F4 and R1 principles. However, these services are based on the service oriented architecture and do not necessarily provide a URI for each resource. The new OGC API Features and OGC API Coverages provide similar functionality but with a resource oriented architecture. They provide a URI for each resource they expose. This makes the new OGC web APIs, as well as the Sensor Things API, compliant to the FAIR sub-principle: “F1. (Meta)data are assigned a globally unique and persistent identifier”. Both OGC Web services and OGC APIs use the HTTP protocols over the Internet and can make use of the current standards and practices for authentication and authorization such as OpenID Connect. Being more modern the OGC APIs are better positioned for adopting practices for authentication and authorization as they are making use of the web using the resource orientation. In this paradigm, authorization on geospatial resources can be fine tuned for each resource URI in the same way as any other resource on the Web. That means that OGC API features, OGC API coverages and The Sensor Things API comply with the FAIR Accessible sub-principles: "A1. (Meta)data are retrievable by their identifier using a standardised communications protocol", "A1.1: The protocol is open, free and universally implementable" and "A1.2 The protocol allows for an authentication and authorisation procedure, where necessary". The OGC API features as described in WFS complies with the sub-principle "F4. (Meta)data are registered or indexed in a searchable resource" and the sub-principle “R1. (Meta)data are richly described with a plurality of accurate and relevant attributes”.

Semantic interoperability:

The focus of the FAIR principle “Interoperable” is on semantic interoperability. The OGC is working on implementing the OGC RAINBOW as a Web accessible source of information about concepts and vocabularies that the OGC defines or that communities ask the OGC to host on their behalf. The RAINBOW follows the FAIR principles too. OGC RAINBOW is implemented using Linked Data principles providing enhanced findability, currently making it compliant with the principles “F1. (Meta)data are assigned a globally unique and persistent identifier” and “F4: (Meta)data are registered or indexed in a searchable resource”. The set of concepts stored in the RAINBOW or in other vocabularies should be used by the data and metadata to comply with the FAIR Interoperable sub-principle “I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation” and the “I2. (Meta)data use vocabularies that follow FAIR principles”. This new standard uses the HTTP protocols over the Internet, this makes it compliant with the “A1. (Meta)data are retrievable by their identifier using a standardised communication protocol" and also compliant with the sub-principle “A1.1 The protocol is open, free, and universally implementable".

The OGC Sensor Things API is an open and free standard that complies to “A1.1 The protocol is open, free, and universally implementable". It incorporates a data model that includes two properties that allow for linking to URLs for “units of measurement” and “observed properties” (e.g. references to variable definitions) that makes it compliant with the sub-principle I2. “(Meta)data use vocabularies that follow FAIR principles”. However, other services and APIs such as the OGC API features and the OGC API coverages do not specify how this could be done into practice and more work needs to be done in that respect. On the other hand, the new OGC APIs use link mechanisms to connect datasets, resource collections and resources to other resources for different purposes making them compliant to the FAIR sub-principle "I3 (Meta)data include qualified references to other (meta)data”. In that sense, the new OGC Sensor Things API plus (STAplus) includes an additional element called “relation” that allows for relating an observation to other internal or external observations. In addition, it also adds an element called “licence” associated to the datastream or to the observation group that complies to the sub-principle R1.1. (Meta)data are released with a clear and accessible data usage licence”. The STA data model can be extended to domain specific areas by subclassing some of the entities such as Thing and Observation making it possible to meet the sub-principle R1.3. (Meta)data meet domain-relevant community standards. STAplus includes many considerations for secure CRUD operations and can support authentication and authorization through the implementation of business logic making it compliant with the sub-principle A1.2. The protocol allows for an authentication and authorisation procedure, where necessary.

In addition to this, the OGC curates standards that define thematic data models and knowledge representations. For example, WaterML is an information model for the representation of water observations data. In addition, PipelineML defines concepts supporting the interoperable interchange of data pertaining to oil and gas pipeline systems. The PipelineML Core addresses two critical business use cases that are specific to the pipeline industry: new construction surveys and pipeline rehabilitation. Another example is the Land and Infrastructure Conceptual Model for land and civil engineering infrastructure facilities. Subject areas include facilities, projects, alignment, road, railway, survey, land features, land division, and “wet” infrastructure (storm drainage, wastewater, and water distribution systems). Finally, GeoSciML is a model of geological features commonly described and portrayed in geological maps, cross sections, geological reports and databases. This standard describes a logical model for the exchange of geological map data, geological time scales, boreholes, and metadata for laboratory analyses. The existence of these standards can help each thematic sector to comply with the FAIR Interoperability sub-principle "I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation". As well as these standards, connecting their vocabularies to information systems or databases would significantly increase their usefulness encouraging the principle of Reusability “R1.(Meta)data are richly described with a plurality of accurate and relevant attributes” and the sub-principle "R1.3 (Meta)data meet domain-relevant community standards".

This table summarises the data formats, catalogues and communication protocols described, according to the compliance of the FAIR principles:

Table 1: Summary table of formats, catalogues and communication protocols

	COG	Zarr	GeoParquet (draft standard)	GeoJSON	STAC	Zenodo	CSW	OGC API Records	WFS	WCS	OGC API Features	OGC API Coverages	OGC RAINBOW	OGC Sensor Things Plus API	OGC curate (WaterML , PipelineML , GeoSciML)
F1. (Meta)data are assigned a globally unique and persistent identifier					✔	✔	✔	✔			✔	✔	✔		
F2. Data are described with rich metadata	✔	✔	✔		✔	✔	✔	✔							
F3. Metadata clearly and explicitly include the identifier of the data they describe	✔	✔				✔ (DOI)	1/2	1/2							
F4. (Meta)data are registered or indexed in a searchable resource	✔ (STAC)				✔	✔	✔	✔	✔	✔	✔	✔	✔		
A1. (Meta)data are retrievable by their identifier using a standardised communications protocol	✔	✔		✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	
A1.1. The protocol is open, free, and universally implementable	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	
A1.2. The protocol allows for an authentication and authorisation procedure, where necessary	✔			✔	✔	✔	1/2	✔	1/2	1/2	✔	✔		✔	
A2. Metadata are accessible, even when the data are no longer available							✔	✔							
I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	✔		✔ (PROJJSON)			✔ (via publications)							✔		✔
I2. (Meta)data use vocabularies that follow FAIR principles													✔	✔	
I3. (Meta)data include qualified references to other (meta)data								✔			✔	✔		✔	

	COG	Zarr	GeoParquet	GeoJSON	STAC	Zenodo	CSW	OGC API Records	WFS	WCS	OGC API Features	OGC API Coverages	OGC RAINBOW	OGC Sensor Things Plus API	OGC curate (WaterML , PipelineML , GeoSciML)
R1. (Meta)data are richly described with a plurality of accurate and relevant attributes						1/2	✔	✔	✔		✔			✔	✔
R1.1. (Meta)data are released with a clear and accessible data usage licence															
R1.2. (Meta)data are associated with detailed provenance							✔ ISO 19115								
R1.3. (Meta)data meet domain-relevant community standards														✔	✔

5. Implementation tools, validation and analysis strategies

As mentioned in the executive summary, this document shows some implementation tools to define the validation and analysis strategies to ensure the implementation of FAIR principles in the project.

In order to do this, in the e-shape (EuroGEO Showcases: Applications Powered by Europe) project which has received funding from the European Union's Horizon 2020 research and innovation programme grant agreement 820852, a tool was developed called Data Management Self-assessment Tool, Menard, L., & Fichaux, N. (2022)[14], to help users to assess whether or not their Data Management Plans have met the requirements of GEO Data Management Principles and the FAIR Principles. In the deliverable 2.2 Report "Status and Prospect for European Environmental Data" 1st version chapter 1, a description can be found about the mapping between the FAIR principles and the GEO Data Management Principles [4].

Although mostly accepted as model/template by the GEO Community, the GEO principles are sometimes confused with the FAIR principles and not always fully utilised. In order to rectify this problem, a questionnaire enables users to check their project compliance with GEO principles. The overlapping of GEO and FAIR principles and their complementarity is shown on a matrix table.

The tool consists of an Excel file which uses macros. Windows users should first ensure that the macros are enabled. The Excel file contains several sheets according to each Principle with its proposed assessment matrix. After the questionnaire is finished, on the final tab called Your Data Management Plan, the application allows the user to generate a report in Word, so that the results of the compliance can be obtained (see an example below in Table 2). On the Introduction sheet there are some instructions how to check If the application provides the Microsoft Word object library to link with Excel.

On The Data Summary sheet, the user can describe general information about the purpose and origins of the data. As the tool is designed from the e-shape project, there is a drop-down list with the name of the Pilots in the Project title, but in the other cells, users can enter their own answers. It will pop up a form when the cell is selected. In order to save the answer in the cell, the "Close and save answer" button should be pressed otherwise it will not be saved.

The tool allows to analyse the progress of level of compliance from the beginning to the last outcome.

As the GEO-DMP sheets are populated, in the following sheets of FAIR, there is a reminder of the responses that GEO assessment overlaps.

[14] Data management self-assessment tool https://gkhub.earthobservations.org/records/0ksgt-7v316

In next page an example is shown, with the Data Management Plan, related to the dataset: Global (T6.04 Development of the world-land degradation neutrality monitor) urbanisation at 500m (2000-2022) based on the light-night images available at: https://doi.org/10.5281/zenodo.7750174.

Data Management Plan

Table 2: Structure of Data Management Plan created by a user as a report, generated by the Data Management Self-assessment Tool

GEO Data Management Principle	Start: Level of compliance (select)	Finish: Level of compliance (select)	Details included (mandatory)	Exceptions	Reasons for exceptions
DMP-1: METADATA FOR DISCOVERY **Data and all associated metadata will be discoverable, through catalogues and search engines, and data access and use conditions, including licences, will be clearly indicated.**	0 - Not applicable to my Pilot	3 - Fully compliant	License Creative Commons Attribution (CC-BY) Indexed in OpenAIRE		Exception due to commercial restrictions
DMP-2: ONLINE ACCESS **Data will be accessible via online services, including, at a minimum, direct download but preferably user-customizable services for access, visualisation and analysis.**	1 - Applicable but not started	3 - Fully compliant	Data is available in Cloud Optimised GeoTIFF (COG) format. Direct download		Exception due to security risks

DMP-3: DATA ENCODING **Data should be structured using encodings that are widely accepted in the target user community and aligned with organisational needs and observing methods, with preference given to non-proprietary international standards.**	0 - Not applicable to my Pilot	0 - Not applicable to my Pilot	0		(select value in dropdown menu)
DMP-4: DATA DOCUMENTATION **Data will be comprehensively documented, including all elements necessary to access, use, understand, and process, preferably via formal structured metadata based on international or community-approved standards. To the extent possible, data will also be described in peer-reviewed publications referenced in the metadata record.**	(select value in dropdown menu)	3 - Fully compliant	Dublin Core		(select value in dropdown menu)

DMP-5: DATA TRACEABILITY **Automatic metadata creation: Tools that create and manipulate the data also should produce provenance documentation automatically to avoid losing steps or incorrectly documenting Tools need to inherit the provenance from previous sources. References to algorithms and versions need to be added.**	0 - Not applicable to my Pilot	0 - Not applicable to my Pilot	0		Exception due to research embargo
DMP-6: DATA QUALITY-CONTROL **Data will be quality-controlled and the results of quality control shall be indicated in metadata; data made available in advance of quality control will be flagged in metadata as unchecked.**	0 - Not applicable to my Pilot	0 - Not applicable to my Pilot	0		Exception due to research embargo
DMP-7: DATA PRESERVATION **Data will be protected from loss and preserved for future use; preservation planning will be for the long term and include guidelines for loss prevention, retention schedules, and disposal or transfer procedures.**	0 - Not applicable to my Pilot	0 - Not applicable to my Pilot	Zenodo repository DOI: 10.5281/zenodo.7750175		Exception due to security risks

DMP-8: DATA AND METADATA VERIFICATION **Data and associated metadata held in data management systems will be periodically verified to ensure integrity, authenticity and readability.**	2 - Partly implemented / ongoing	2 - Partly implemented / ongoing	0		(select value in dropdown menu)
DMP-9: DATA REVIEW AND REPROCESSING **Data will be managed to perform corrections and updates in accordance with reviews, and to enable reprocessing as appropriate; where applicable this shall follow established and agreed procedures.**	(select value in dropdown menu)	(select value in dropdown menu)	0		Exception due to security risks
DMP-10: PERSISTENT AND RESOLVABLE IDENTIFIERS **Data will be assigned appropriate persistent, unique and resolvable identifiers to enable documents to cite the data on which they are based and to enable data providers to receive acknowledgement for use of their data.**	1 - Applicable but not started	3 - Fully compliant	DOI: 10.5281/zenodo.7750175		(select value in dropdown menu)

FAIR - Data Management Principle	Start: Level of compliance (select)	Finish: Level of compliance (select)	Please provide details (Mandatory)
FAIR 1 - MAKING DATA FINDABLE, INCLUDING PROVISIONS FOR METADATA			
Outline the discoverability of data (metadata provision)	2 - Partly implemented / ongoing	1 - Applicable but not started	Dublinc Core, DCAT
Outline the identifiability of data and refer to standard identification mechanism. **Do you make use of persistent and unique identifiers such as Digital Object Identifiers?**	2 - Partly implemented / ongoing	2 - Partly implemented / ongoing	Yes, it is used a DOI reference
Outline naming conventions used	2 - Partly implemented / ongoing	2 - Partly implemented / ongoing	Title of the dataset_ sensor_units_spatial resolution_temporal_extent_reference_system(EPSG code)_publication date_product format Example: nightlights.average_viirs.v21_m_500m_s_20200101_20201231_go_epsg4326_v20230318.tif
Outline the approach towards search keyword	3 - Fully compliant	3 - Fully compliant	Sensor used and the kind of data obtained

Outline the approach for clear versioning	(select value in dropdown menu)	2 - Partly implemented / ongoing	Version v0.1 10.5281/zenodo.7750175 and Date of the current version
Specify standards for metadata creation (if any). **If there are no standards in your discipline describe what metadata will be created and how.**	0 - Not applicable to my Pilot	0 - Not applicable to my Pilot	0
FAIR 2 - MAKING DATA OPENLY ACCESSIBLE			
Specify which data will be made openly available? **If some data is kept closed provide rationale for doing so**	(select value in dropdown menu)	(select value in dropdown menu)	Yes. open access
Specify how the data will be made available	(select value in dropdown menu)	(select value in dropdown menu)	Zenodo repository

Specify what methods or software tools are needed to access the data? **Is documentation about the software needed to access the data included?** **Is it possible to include the relevant software (e.g. in open source code)?**	(select value in dropdown menu)	(select value in dropdown menu)	Direct Download.
Specify where the data and associated metadata, documentation and code are deposited	(select value in dropdown menu)	(select value in dropdown menu)	Zenodo https://zenodo.org/record/7750175
Specify how access will be provided in case there are any restrictions	(select value in dropdown menu)	(select value in dropdown menu)	Licence CC BY 4.0
FAIR 3 - MAKING DATA INTEROPERABLE			
Foreseen efforts towards interoperability, for each GEO DMP. **Your responses from the GEO section are reminded here.**	🟨 GEO - DMP - 1	GEO - DMP - 6	Color code: NONE - N/A or unknown RED - Not significant effort to ensure interoperability ORANGE - Interoperable but with restrictions GREEN - Fully interoperable in line with GEO recommendations
	🟩 GEO - DMP - 2	🟩 GEO - DMP - 7	
	GEO - DMP - 3	GEO - DMP - 8	
	🟨 GEO - DMP - 4	GEO - DMP - 9	

	GEO - DMP - 5	GEO - DMP - 10	
Assess the interoperability of your data. **Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.**	(select value in dropdown menu)	(select value in dropdown menu)	TIF and COG formats
Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability?	(select value in dropdown menu)	(select value in dropdown menu)	Partially
If not, will you provide mapping to more commonly used ontologies?	(select value in dropdown menu)	(select value in dropdown menu)	0
FAIR 4 - INCREASE DATA RE-USE (THROUGH CLARIFYING LICENCES)			
Specify how the data will be licenced to permit the widest reuse possible	(select value in dropdown menu)	(select value in dropdown menu)	CC BY 4.0

Specify when the data will be made available for re-use. **If applicable, specify why and for what period a data embargo is needed**	(select value in dropdown menu)	(select value in dropdown menu)	It is published.
Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why	(select value in dropdown menu)	(select value in dropdown menu)	Yes
Describe data quality assurance processes	(select value in dropdown menu)	(select value in dropdown menu)	
Specify the length of time for which the data will remain re-usable	(select value in dropdown menu)	(select value in dropdown menu)	
Allocation of resources			
Describe the volume of resources allocated to make your Data FAIR-compliant (% of development costs)	5 %		

In addition, as the data in this example is located in the Zenodo repository, the OpenAIRE Validator service[15] is available, in order to check if the repository complies with OpenAIRE guidelines.

Other tools connected to FAIRness assessment were studied and are cited here:

- FAIR-Enough tool[16] is developed and hosted by the Institute of Data Science at Maastricht University. It allows the evaluation of how much online resources follow the FAIR principles .
- MQA tool[17] is developed by the consortium of data.europa.eu to study the quality of metadata harvested by data.europa.eu. Although the application is limited exclusively to the metadata that data.europa.eu collects during the harvesting process, the dimensions utilised are derived from the FAIR principles.

Next steps

There will be a second version of this document with the title D2.11 Report "FAIR operation, validation and analysis guidelines" final version which should be delivered in month 41 of the OEMC project with the validation.

Related tasks and outputs

This deliverable gathers the work in Task 2.2: Analysis of FAIR data status and workflows in WP2 – User-driven system design and FAIR workflows of the project OEMC.

Next version will also include a description of the work done in Task 2.7. Develop a robust operation, validation and analysis workflow for checking products.

[15] OpenAIRE Validator service https://www.openaire.eu/validator-registration-guide
[16] FAIR-Enough tool https://fair-enough.semanticscience.org/
[17] MQA tool https://data.europa.eu/mqa/methodology?locale=en#inline-nav-2

References

1. Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18 and https://www.nature.com/articles/sdata201618
2. OEMC Deliverable 2.2 - Report "Status and prospect for European environmental data" 1st version