



**Project title:** All Data 4 Green Deal - An Integrated, FAIR Approach for the Common European Data Space

**Project number:** 101061001

**Project Acronym:** AD4GD

**Type:** HORIZON-AG - HORIZON Action Grant Budget-Based

**Work program topics addressed:** HORIZON-CL6-2021-GOVERNANCE-01

**DELIVERABLE NO: D2.2**

**ROADMAP FOR FAIR IN-SITU OBSERVATIONS FOR THE GREEN DEAL (FINAL)**

**Due date of deliverable:** 30/04/2024

**Actual submission date:** 15/07/2024

**Version:** 1.0

**Main Authors:** Joan Masó, Alba Brobia, Lucy Bastin and Victoria Lush

## DOCUMENT METADATA

<b>Project number</b>	101061001
<b>Project title</b>	All Data 4 Green Deal - An Integrated, FAIR Approach for the Common European Data Space

<b>Deliverable title</b>	Roadmap for FAIR in-situ observations for the green deal (final)
<b>Deliverable number</b>	D2.2
<b>Deliverable version</b>	1.0
<b>Contractual date of delivery</b>	29/04/2024
<b>Actual date of delivery</b>	15/07/2024
<b>Document status</b>	Final
<b>Document version</b>	1.0
<b>Online access</b>	URL
<b>Dissemination</b>	Public
<b>Work package</b>	WP2: In-situ networks, CitSci and Socioeconomic Data
<b>Partner responsible</b>	Ecological and Forestry Applications Research Centre (CREAF)
<b>Author(s)</b>	Joan Masó, Alba Brobia, Lucy Bastin, Victoria Lush, Ivette Serral
<b>Editor(s)</b>	Joan Masó, Alba Brobia
<b>Reviewer(s)</b>	Victoria Lush
<b>EC Project Officer</b>	Lara Congiu

<b>Abstract</b>	<p>The common Green Deal Data Space will interconnect currently fragmented and dispersed data from various sources (including in-situ, statistical, cartographical and remote sensing), both for/from the private and public sectors, to support the objectives of the European Green Deal. It will offer an interoperable, trusted IT environment for data processing, and a set of rules of legislative, administrative and contractual nature that determine the rights of access to and use of the data.</p> <p>This document is the Deliverable 2.2 of the AD4GD project. It presents the results achieved in the context of Work Package 2 "In-situ networks, CitSci and Socioeconomic Data". The reader should consider that the document builds on the text of the previous Deliverable 2.1 (of the same title) and extends and modifies it.</p> <p>The document discusses how AD4GD applies the FAIR principles to the GDDS via the European Network of Earth Observation Networks (ENEON) graph, the list of identified in-situ data networks, their relations with the Essential Variables (EVs), the recommendations for aligning GEO DMPs to the FAIR principles and the 12 technical components identified in AD4GD to make the data in the GDDS more compliant to FAIR. The main networks of Earth Observation (e.g. research infrastructures) are recorded in the ENEON. In order to be useful for AD4GD, the original graph has been updated and extended beyond research infrastructures. A detailed report of changes, applied to the original graph, is presented. The role of the critical in-situ data sources for the GDDS is also addressed, including RI,</p>
-----------------	--

	<p>CitSci, INSPIRE and the HVD, and IoT. In addition, the high priority datasets and services as identified by the GREAT project have been analysed with in-situ data on focus and a summary of this analysis is presented. An overview of the Essential Variables as a common framework to semantically tag in-situ data is presented and the use of I-ADOPT ontology framework, designed to facilitate interoperability between existing variable description models across domains by re-using FAIR vocabulary terms, is proposed. AD4GD has expressed the Essential Biodiversity Variables using the I-ADOPT ontology and included the 84 products defined by EuropaBON in the OGC Rainbow. A set of recommendations to align the GEO data management principles with the FAIR principles is presented. In this deliverable we demonstrate that with minimum changes the GEO DMP can be mapped to the FAIR principles formulating 15 concrete recommendations. We also demonstrate that the three pilots (Water quality, Biodiversity, Air Quality) use components of the AD4GD architecture that contribute to make the pilot data FAIR. Indeed, the pilot needs are covered by a set of 12 components that, once combined together, form the full AD4GD architecture. A summary of the contributions to FAIR of each component is provided.</p>
<b>Keywords</b>	Essential Variables, Web Services, Trust, APIs, Architecture, In-situ, Citizen Science.
<b>Disclaimer</b>	Views and opinions expressed in this deliverable are those of the author(s) only and do not necessarily reflect those of the European Union, the United Kingdom or Switzerland. Neither the European Union nor United Kingdom nor Switzerland can be held responsible for them.

## DOCUMENT VERSION HISTORY

<b>Version history</b>			
<b>Version</b>	<b>Date</b>	<b>Modification reason</b>	<b>Modified by</b>
D2.1 0.1	20/03/2023	Initial version of the document.	Joan Masó and Lucy Bastin
D2.1 0.2	28/03/2023	FAIR principles and DMP.	Alba Brobia
D2.1 0.4	15/04/2023	OGC web services and semantic considerations.	Joan Masó
D2.1 0.5	20/04/2023	Citizen science study.	Victoria Lush
D2.1 0.6	25/04/2023	Introduction, Conclusion and Executive Summary	Joan Masó
D2.1 0.9	28/04/2023	Final review.	Victoria Lush
D2.1 1.0	29/04/2023	Formatting.	Alba Brobia
D2.2 0.1	18/06/2024	Adding content on EVs and I-ADOPT.	Ivette Serral
D2.2 0.2	18/06/2024	Section 2: updated ENEON graph. Section 3.4: additional text on IoT.	Victoria Lush
D2.2 0.3	19/06/2024	The role of INSPIRE and environmental data. Standards and components for FAIR data in GDDS. High priority datasets and services.	Alba Brobia
D2.2 0.4	28/06/2024	New Introduction. I-ADOPT Framework ontology applied to the EBV. Relation of FAIR with the GEO DMP. New Conclusions. Final review.	Joan Masó
D2.2 0.8	08/07/2024	Final version for internal review.	Alba Brobia
D2.2 0.9	11/07/2024	Review.	Francesca Noardo
D2.2 1.0	15/07/2024	Final version.	Joan Masó

## ABBREVIATIONS

Abbreviation	Definition
API	Application Programming Interface
CAMS	Copernicus Atmosphere Monitoring Service
CitSci	Citizen Science
CKAN	Comprehensive Knowledge Archive Network
CSV	Comma-Separated Values
CSVW	CSV on the Web
DCAT	Data Catalog Vocabulary
DMP	Data Management Principles
DSSC	Data Spaces Support Centre
EBV	Essential Biodiversity Variable
EC	European Commission
EDC	Eclipse Dataspace Components
ENEON	European Network of Earth Observation Networks
ENVRI	Environmental Research Infrastructures
EO	Earth Observation
EPOS	European Plate Observing System
ESDAC	European Soil Data Centre
ESFRI	European Strategy Forum on Research Infrastructures
EV	Essential Variable
FAIR	Findable, Accessible, Interoperable, Reusable
FRED	Freshwater Research and Environmental Database
GBF	Global Biodiversity Framework
GCOS	Global Climate Observing System
GDDS	Green Deal Data Space
GDIM	Green Deal Information Model
GEO	Group on Earth Observations
GeoDCAT	Geospatial Data Catalog Vocabulary
GEO BON	Group on Earth Observations Biodiversity Observation Network
GUF	Geospatial User Feedback
HVD	High Value Datasets
I-ADOPT	Interoperable Descriptions of Observable Property Terminology
ICT	Information and Communication Technologies

INSPIRE	Infrastructure for Spatial Information in Europe
IDSA	International Data Spaces Association
IoT	Internet of Things
JRC	Joint Research Centre
JSON	JavaScript Object Notation
JSON-LD	JavaScript Object Notation for Linked Data
ML	Machine Learning
OGC	Open Geospatial Consortium
OMS	Observations, Measurements, and Samples
PM	Particulate Matter
PID	Permanent IDentifier
RDA	Research Data Alliance
RDF	Resource Description Framework
RIs	Research Infrastructures
R&I	Research and Innovation
SDG	Sustainable Development Goals
SDI	Spatial Data Infrastructure
STA	Sensor Thing API
STApplus	Sensor Things API plus (extension of STA)
TAPIS	Tables from OGC APIs for Sensors
TRUST	Transparency, Responsibility, User focus, Sustainability, and Technology
TVOCs	Total Volatile Organic Compounds

## Table of Contents

1	INTRODUCTION .....	13
2	MAIN NETWORKS OF IN-SITU DATA IN EUROPE .....	14
3	IN-SITU DATA SOURCE TYPES IN THE GDDS .....	29
3.1	THE ROLE OF THE RESEARCH INFRASTRUCTURES .....	29
3.2	THE ROLE OF CITIZEN SCIENCE .....	31
3.3	THE ROLE OF INSPIRE AND ENVIRONMENTAL DATA .....	33
3.4	THE ROLE OF IoT .....	35
3.5	High Priority Datasets and Services with in-situ on focus as identified by the GREAT project blueprint for the GDDS .....	36
4	ESSENTIAL VARIABLES AS A COMMON FRAMEWORK TO SEMANTICALLY TAG IN-SITU DATA .....	39
4.1	USING EV VOCABULARIES TO IDENTIFY COMPATIBLE DATASETS .....	41
4.1.1	I-ADOPT Framework ontology applied to the EBV .....	42
4.2	USING EV VOCABULARIES IN OGC STANDARDS .....	48
5	RECOMMENDATIONS FOR IN-SITU DATA PROVIDERS .....	50
5.1	FAIR PRINCIPLES .....	50
5.2	RELATION OF FAIR WITH THE GEO DMP .....	52
5.2.1	Findable .....	53
5.2.2	Accessible.....	54
5.2.3	Interoperable .....	55
5.2.4	Reusable.....	56
5.2.5	About the other DMPs.....	56
5.3	IMPLEMENTING FAIR PRINCIPLES .....	60
■	Priority Recommendations .....	60
■	Supporting Recommendations.....	60
6	STANDARDS AND COMPONENTS FOR FAIR DATA IN GDDS.....	61
6.1	STANDARDS AND COMPONENTS FOR FINDING IN-SITU DATA IN THE GDDS.....	63
6.2	STANDARD AND COMPONENTS FOR IN-SITU DATA ACCESS IN THE GDDS.....	64
6.3	STANDARD FOR IN-SITU DATA (SEMANTIC) INTEROPERABILITY IN THE GDDS.....	66
6.4	AUTHENTICATION AND PRIVACY ASPECTS .....	68
7	CONCLUSIONS .....	69

## Table of Figures

Figure 1: ENEON initial data model (the EV product was later added to the model).....	14
Figure 2: LinkRef class of ENEON data model.....	15
Figure 3: Initial ENEON graph.....	15
Figure 4: Detailed view of the Earth Observation Network object modelled in the ENEON graph.....	16

Figure 5: ENEON graph air quality networks, air quality observable properties and connections to SDGs. ... 19

Figure 6: ENEON graph JSON - PM2.5 observable property code with a link to RAINBOW server definition and networks that observe PM2.5 property. .... 19

Figure 7: ENEON graph air quality networks - PM2.5 observable property example. .... 20

Figure 8: RAINBOW server - PM2.5 observable property definition. .... 20

Figure 9: ENEON graph JSON - SDG 3.9.1 with corresponding air quality observable properties. .... 20

Figure 10: ENEON graph air quality networks - OpenAQ network example. .... 21

Figure 11: ENEON graph water quality networks, water quality observable properties and connections to SDGs. .... 22

Figure 12: ENEON graph water quality networks - Wasserportal network example. .... 22

Figure 13: ENEON graph water quality networks - FRED network example. .... 23

Figure 14: ENEON graph JSON - SDG with corresponding water quality observable properties. .... 23

Figure 15: ENEON graph Kunming-Montreal GBF 2030 targets. .... 24

Figure 16: ENEON graph Kunming-Montreal GBF 2030 targets and connections to SDGs. .... 24

Figure 17: ENEON graph EuropaBON and GEOBON Essential Biodiversity Variables (EBVs). .... 26

Figure 18: ENEON graph EuropaBON EBVs. .... 26

Figure 19: ENEON graph GEOBON EBVs. .... 27

Figure 20: ENEON graph EuropaBON EBVs - Community Abundance EBV and EBV product example. .... 27

Figure 21: ENEON graph JSON - Community Abundance EBV and EBV product example. .... 28

Figure 22: ENEON graph citizen science biodiversity networks - iNaturalist network example. .... 28

Figure 23: Environmental research infrastructures included in the ESFRI roadmap 2021. .... 31

Figure 24: EU-Citizen.Science platform projects status. .... 32

Figure 25: EU-Citizen.Science platform participation tasks. .... 33

Figure 26: EU-Citizen.Science platform project topics. .... 33

Figure 27: INSPIRE data themes. .... 34

Figure 28: GCOS Essential Climate Variables overview. .... 40

Figure 29: Data model used in ENEON for EVs and EV products. .... 40

Figure 30: I-ADOPT Framework schema (extracted from <https://i-adopt.github.io/>) .... 43

Figure 31: I-ADOPT UML schema expressing the AD4GD proposal of extension of I-ADOPT current version. .... 44

Figure 32: Extracted from <https://github.com/EuropaBON/EBV-Descriptions/wiki/Freshwater-Species-distributions-of-freshwater-mammals>. .... 45

Figure 33: EBVs csv encoded following the I-ADOPT schema. .... 45

Figure 34: I-ADOPT schema adapted by AD4GD expressing the Species populations | Species distributions EV in the case of Freshwater Mammals. .... 46

Figure 35: Example of the Community Abundance EVs described following the I-ADOPT schema and encoded in json format for the ENEON graph. .... 47

Figure 36: ENEON graph showing the EV of Community Abundance ..... 48

Figure 37: STA entity data model includes the concept of observedProperty. .... 48



Figure 38: Correspondence between the FAIR principles and the GEO DMP. .... 58

Figure 39: Building blocks to support the implementation of the AD4GD proposal of the GDDS, including a mapping of the planned pilots’ components. Credits: AD4GD D6.1 (L: AU)..... 61

Figure 40: AD4GD Building blocks and Components in contribution to FAIR. .... 63

Figure 41: Detailed view of the ISO19110 Feature Catalogue model. .... 66

Figure 42: How to extend CharacterString to include a URI. .... 66

Figure 43: Description of the Range types included as values in a coverage following the ISO 19115-1. .... 67

**Table of Tables**

Table 1: ENEON graph network types definitions. .... 18

Table 2: Updated list of 24 EBVs recorded in the ENEON graph. .... 26

Table 3: High priority data-services defined by the GREAT project classified according to the type of data provided (whether they offer in-situ data services or remote sensing data services), and the specific type of in-situ data source (identifying if the in-situ data originates from RI, CitSci, INSPIRE datasets, or IoT). .... 38

Table 4: Full version of the FAIR principles: Findable (F1, F2, F3, F4), Accessible (A1, A1.1, A1.2, A2), Interoperable (I1, I2, I3), and Reusable (R1, R1.1, R1.2, R1.3)..... 51

Table 5: Summary table of the current text of the GEO DMP. .... 53

Table 6: Summary table of the proposed revision of the GEO DMPs. .... 59

## EXECUTIVE SUMMARY

*NOTE: This section is designed a self-contained condensed version of the complete document and includes its essential information. To that purpose, it repeats some enumerations from the main document. If you are planning to read the whole document, you may consider skipping this section.*

The common European Green Deal data space will interconnect currently fragmented and dispersed data from various sources (including in-situ, statistical, cartographical and remote sensing), both for/from the private and public sectors, to support the objectives of the European Green Deal. It will offer an interoperable, trusted IT environment for data processing, and a set of rules of legislative, administrative and contractual nature that determine the rights of access to and use of the data.

The main networks of Earth Observation (e.g. research infrastructures) are recorded in the ENEON graph focused on the essential component of the in-situ measurements and networks, evolving thus to a convenient platform to analyse the Essential Variables in Earth Observation, their relations with the in-situ networks and their role to monitor SDGs. The ENEON data model defines mainly 5 classes of resources, 3 of them relevant for this project: an essential variable, an essential variable product (defined by its requirements), and the Earth Observation network (and its supporting projects). In order to be useful for AD4GD, the original graph has been updated and extended beyond research infrastructures. This is a summary of changes applied to the ENEON graph:

1. The list of ENEON themes has been extended to include additional terms that were required to describe new networks.
2. ENEON graph entities have been additionally labelled with the EU Green Deal thematic areas using the "GDThematicArea" key.
3. Societal Benefit Areas (SBA) tags in the original ENEON graph were duplicated in both EV and EO network entities. It was decided to use SBA tags only in EVs
4. The list of geographic extent terms has been updated to the coverage vocabulary from Giuliani et al., 2020<sup>1</sup>.
5. Network types have been revised and updated to include terms with clear definitions
6. The graph has been revised to include network types (SDG, EV, EONetwork) and subtypes (e.g., Federation, Data Centre) to label EONetwork classifications
7. SGDs recorded in the original ENEON graph have been revised to include the latest SDG indicators (e.g., SDG 13.2.2).
8. The original ENEON graph included 21 GEOBon-defined Essential Biodiversity Variables (EBVs) (<https://geobon.org/ebvs/what-are-ebvs/>). The graph has been updated to include 24 EBVs based on the EuropaBON EBV descriptions (<https://github.com/EuropaBON/EBV-Descriptions>).
9. Kunming-Montreal Global Biodiversity Framework (GBF) 2030 targets (<https://www.cbd.int/gbf/targets>) have been added to the ENEON graph and connected to the corresponding SDGs.
10. New atmospheric and air quality observable properties (e.g., Air Temperature, Ozone, PM10) have been added to the ENEON graph and connected to the corresponding SDGs.
11. New air quality networks including citizen science networks (e.g., Sensor.Community) have been added to the ENEON graph and connected to the observable properties that they measure.
12. New water level and water quality observable properties (e.g., Groundwater Water Level, Water Flow, pH) have been added to the ENEON graph and connected to the corresponding SDGs.
13. New water quality networks have been added to the ENEON graph and connected to the observable properties that they measure.
14. New citizen science biodiversity networks have been added to the ENEON graph and connected to the EBVs.

<sup>1</sup> <https://www.mdpi.com/2306-5729/5/4/100>

Essential Variables are an abstract concept that is commonly associated with a set of measurements (e.g. for the Climatic set there are the Essential Climate Variables). Each set of EVs is classified in themes (e.g. Atmosphere, Land, Ocean) and has several individual variables in it. The EV framework and its EV products provide an excellent starting point to define a set of common concepts that can be used both in remote sensing and in-situ data (including Citizen Science). However, the EV are too generic to describe product requirements. GCOS has released a set of requirements for EV products. EuropaBON has also reformulated the EBV in terms of products that vary depending on the realm and other factors. One way to express this as an ontology is the use of I-ADOPT, an ontology framework designed to facilitate interoperability between existing variable description models across research domains. It provides a common set of core components and relations to define machine-interpretable variable descriptions that re-use FAIR vocabulary terms. Key features of I-ADOPT include four main classes: Variable, Property, Entity, and Constraint and 6 relations: hasProperty, hasObjectOfInterest, hasContextObject, hasMatrix, hasConstraint, constrains. AD4GD has expressed the EBV using the I-ADOPT ontology and included the 84 products defined by EuropaBON in the OGC Rainbow. This way, any biodiversity dataset in the GDDS could be semantically tagged using this vocabulary.

In-situ Earth observation data can be produced by several actors, including the scientific community via the Research Infrastructures (RIs), the public administration geoportals of Member States offering socio-economic and INSPIRE datasets, the citizen science (CitSci) initiatives generating crowdsourced information, and the more recent Internet of Things (IoT).

The implementation of FAIR principles, complemented by other relevant principles (e.g., GEO DMPs) should be achieved in the GDDS. The principles emphasise machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data. It is expected that the GDDS will have one or more nodes that will be dedicated to data discovery and finding information. The FAIR principles should inspire the architecture of the GDDS as well as the DMP in GEO. In this deliverable we demonstrate that with minimum changes the GEO DMP can be mapped to the FAIR principles. In that respect, 15 concrete recommendations are formulated.

- **Recommendation 1:** Include in DMP-10 that "Metadata will be assigned appropriate unique, persistent, resolvable identifiers".
- **Recommendation 2:** Modify DMP-10 to explicitly request that the PID is included in the metadata.
- **Recommendation 3:** In DMP-2, include the need of "resolvable" PID (i.e. the capacity to access the data using the PID). This is not current common practice in the geospatial world.
- **Recommendation 4:** Add to DMP-2 the use of services or Web APIs (to consider the migration of the OGC web services to APIs) (While this is not suggested by the FAIR-DMP comparison it is a necessary update)
- **Recommendation 5:** Add to DMP-2 the word "remotely" in front of "computation" to emphasize that for big data moving code close to the data might be more efficient than a simple download.
- **Recommendation 6:** In DMP-2 include a reference to "open and universally implementable protocol" after mention of "web service and APIs".
- **Recommendation 7:** In DMP-3 replace "non-proprietary international standards." By "open and universally implementable standards".
- **Recommendation 8:** Add a reference to possible authentication and authorisation procedure to DMP2.
- **Recommendation 9:** Add " metadata will remain accessible, even when data have been disposed and transferred to an archive" to the end of DMP-7.
- **Recommendation 10:** Rephrase the DMP-3 by adding references to data models. "Data will be structured using encodings **described by knowledge representation models (i.e. data models or data schemas)** that are widely accepted in the target user community and aligned with

organizational needs and observing methods, with preference given to non-proprietary international standards.

- **Recommendation 11:** Include a reference to vocabularies in DMP-4 in the following way: "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. **Data should be annotated with references to external and accepted vocabularies preferable in machine readable formats.** To the extent possible, data will also be described in peer reviewed publications referenced in the metadata record.
- **Recommendation 12:** Extend DMP-5 to other kinds of links between datasets by adding: "Datasets will also include other relations to other datasets when appropriate, such as being a part, a new version or a different resolution of another dataset.
- **Recommendation 13:** Move the reference to "use conditions, including licenses" from DMP-1 to DMP-4.
- **Recommendation 14:** Move DMP-10 to the first position to align it with the FAIR principles Also consider DMP10 as Discoverability.
- **Recommendation 15:** Switch DMP-6 and DMP-7 to group then 3 principles that go beyond the scope of FAIR.

In addition the AD4GD three pilots have been designed in WP6. This deliverable demonstrates that the pilots use components of the AD4GD architecture that contribute to make the pilot data FAIR. Indeed, the pilot needs are covered by a set of 12 components that once combined together form the AD4GD architecture. This is the summary of the contributions to FAIR of each component:

- **AD4GD Components contributing to data Findability**
  - Component 9 – Data catalogue and metadata
  - Component 10 – Data catalogue and metadata with Semantic Uplift
- **AD4GD Components contributing to data Accessibility**
  - Component 2 – Evaluation of Connector Solutions
  - Component 4 – Mobilising Data to STAplus
  - Component 5 – Mobilising sensor data with STA
  - Component 6 – Data Cubes for Consuming, Publishing and Processing Multidimensional Data
  - Component 7 – Open workflows for habitat connectivity computation
  - Component 8 – Water modelling and prediction
- **AD4GD Components contributing to data Interoperability**
  - Component 1 - Automated Ingestion of Data from Diverse Low-cost Sensors
  - Component 3 – Semantic uplift: Sparql/JSON-LD For Data Exchange
  - Component 12 – GDDS Semantic Terms - RAINBOW Server
- **AD4GD Components contributing to data Reusability**
  - Component 11 – Data Trustworthiness Framework
  - Component 9 – Data catalogue and metadata for data provenance (quality)
  - Component 2 – Evaluation of Connector Solutions for data trust are also considered as main contributions to the data reusability in the AD4GD proposal of the GDDS.

This deliverable discusses how AD4GD contributes to the FAIR principles with the ENEON graph, the recommendations for making the GEO DMPs more FAIR and with 12 components to make the data in the GDDS more FAIR.

## 1 INTRODUCTION

Environmental pressures play an important role in the most critical global challenges that humanity faces today (including those related to sustainable energy and food production, water supply, human health and well-being). The mitigation and adaptation to climate change, prevention of environmental pollution, conservation and sustainable use of key natural resources and ecosystem services are vital. Modern society is progressively vulnerable to the increased frequency of natural hazards (such as extreme weather, earthquakes, floods, hunger due to failed harvests or pandemic disease outbreaks) causing loss of life and having an enormous impact on society, and environmental catastrophes can shutter societal security and cause migration with related security problems. In the context of all these problematics, the European Union has adopted the European Green Deal as a way to transform the EU into a modern, resource-efficient and competitive economy, while respecting the boundaries of sustainability and ensuring: no net emissions of greenhouse gases by 2050, economic growth decoupled from resource use, no person and no place left behind. The European Green Deal will improve the well-being and health of citizens and future generations by providing: fresh air, clean water, healthy soil and biodiversity, renovated, energy efficient buildings, healthy and affordable food, more public transport, and cleaner energy<sup>2</sup>.

The common European Green Deal Data Space will interconnect currently fragmented and dispersed data from various sources (including in-situ, statistical, cartographical and remote sensing), both for/from the private and public sectors, to support the objectives of the European Green Deal. It will offer an interoperable, trusted IT environment for data processing, and a set of rules of legislative, administrative and contractual nature that determine the rights of access to and use of the data. This deliverable focuses on the in-situ data sources and its necessary components for the Green Deal Data Space.

This document is the Deliverable 2.2 for the AD4GD project. It presents results achieved in the context of the Work Package 2 "In-situ networks, CitSci and Socioeconomic Data".

The section 2 of this document provides a description of the work done to improve the European Network of Earth Observation Networks (ENEON)<sup>3</sup> graph produced in the ConnectinGEO. To make it useful for the GDDS, EU Green Deal thematic areas were included, the connections to the Essential Variables were updated, other sets of global targets beyond the SDGs (such as the Kunming-Montreal Global Biodiversity Framework - GBF, 2030 targets) were added and some relevant citizen science initiatives were included in the updated graph.

In-situ Earth observation data can be produced by several actors, including the scientific community via the Research Infrastructures (RIs), the public administration geoportals of Member States offering socio-economic and INSPIRE datasets, the citizen science (CitSci) initiatives generating crowdsourced information, and the more recent Internet of Things (IoT). Section 3, classifies the different in-situ data sources contributing to the GDDS.

Section 4 elaborates on the evolution of the concept of the Essential Variables and the Essential Variable Products. In the Essential Biodiversity Variables (EBV), the I-ADOPT ontology has been used to define the EBV products defined by EuropaBON. This specialisation will help the ENEON graph to better integrate the EBV. It also describes how the different sets of EVs are included in the OGC Rainbow in collaboration with WPI.

Section 5 discusses how the FAIR principles should inspire the architecture of the GDDS as well as the DMP in GEO. Recommendations are provided to do a set of minimum changes in the GEO DMP to make them a profile of the FAIR principles for GEO. In addition, this deliverable characterises the 12 components of the AD4GD architecture and describes how they contribute to make the data of the pilots FAIR.

---

<sup>2</sup> [https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal\\_en](https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal_en)

<sup>3</sup> <https://www.eneon.net/>

## 2 MAIN NETWORKS OF IN-SITU DATA IN EUROPE

ENEON is the European Observatory of Earth Observation Networks, formerly European Network of Earth Observation Networks, funded by the European Union under the H2020 ConnectinGEO project and continued under the ERA-PLANET GEOEssential. ENEON mainly works in involving non-space EO in-situ networks into GEOSS with the aim of providing a better approach to Essential Variables and UN SDG indicators to resolve interdisciplinary problems, and helping to improve the European EO in-situ participation to GEO and EuroGEO.

The main networks of Earth Observation are recorded in the ENEON graph. Within the GEOEssential H2020 project, ENEON graph became more focused on the essential component of the in-situ measurements and networks, evolving thus into a convenient platform to analyse the Essential Variables in Earth observation, their relations with the in-situ networks and their role to monitor SDGs. The ENEON data model (Figure 1) defines mainly 5 classes of resources:

- A Essential Variable (EV, a variable defined as having a high impact, high feasibility and relatively low cost of implementation for observing and monitoring the Earth system across the various Societal Benefit Areas. This is an abstract construction.)
- A Essential Variable product (A product that measures a Essential Variable defined by its requirements.)
- A Earth Observation network (A network that is responsible for acquiring measurement about a particular topic that can constitute EV products) and its supporting projects
- An SDG (A Sustainable Development Goal As defined by the UN)
- An SDG indicator (an indicator to measure progress towards the targets defined in each SDG, as defined by the UN. Many SDG indicators can be derived from EO and EVs)

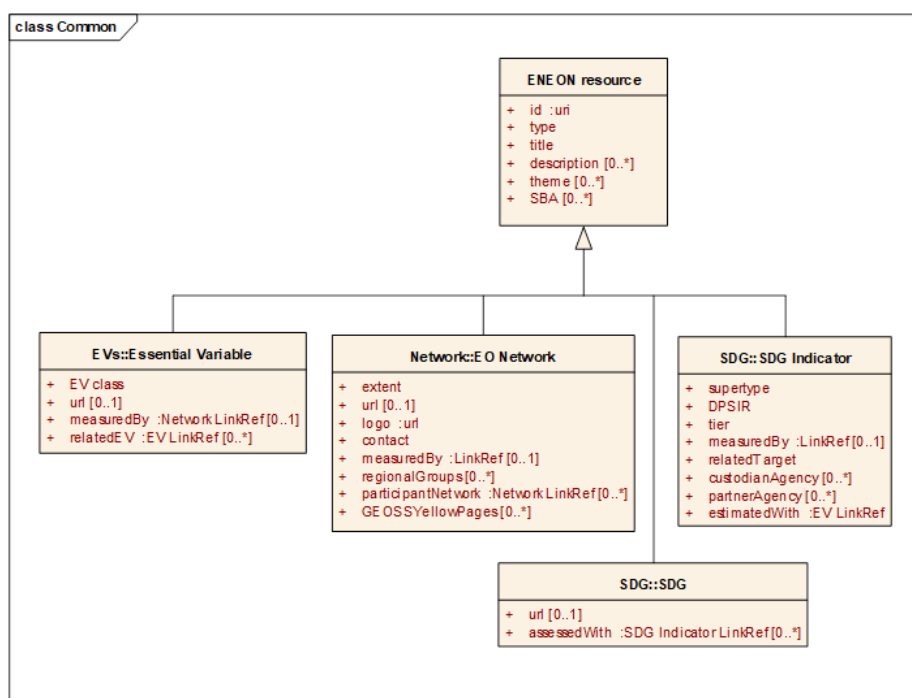


Figure 1: ENEON initial data model (the EV product was later added to the model)

All 5 types of resources have several common characteristics including a unique identifier. Some resources have links to other objects by referencing their identifiers. In the UML, we model a generic LinkRef class



that specializes classes (see Figure 2). In practice, the name of the subclass indicates the type of a target resource.

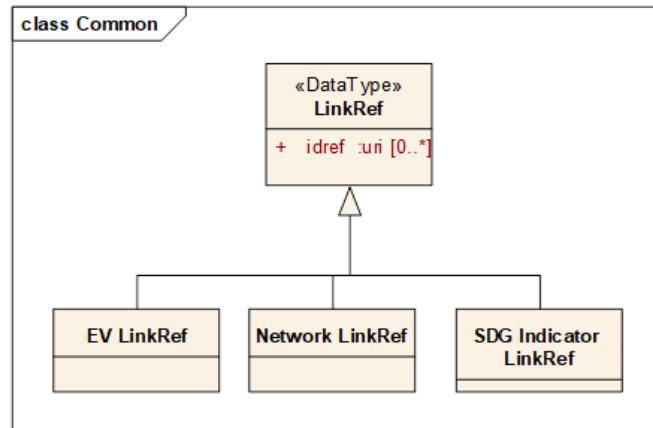


Figure 2: LinkRef class of ENEON data model.

This linking mechanism allows for the creation of relations in the “RDF style” and to represent all these resources as a graph (Figure 3).

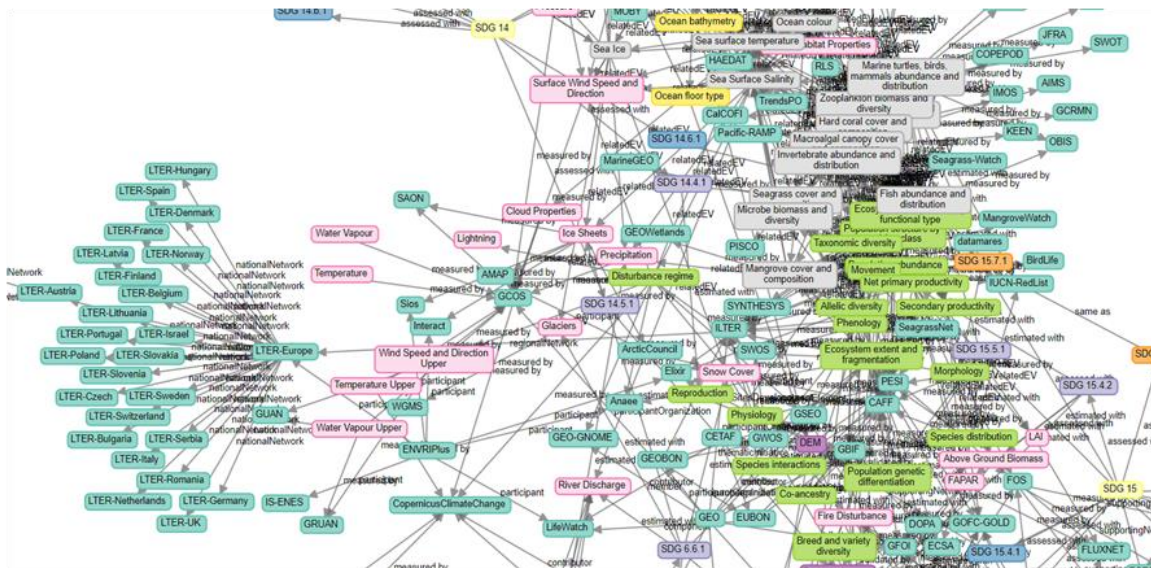


Figure 3: Initial ENEON graph.

In particular, an Earth Observation Network is one of the objects modelled in the ENEON graph (see Figure 4).

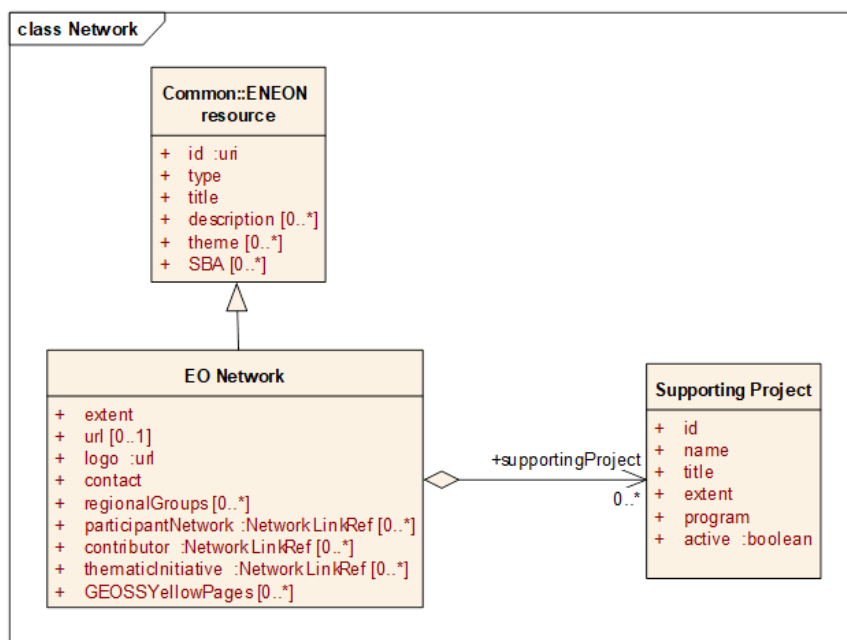


Figure 4: Detailed view of the Earth Observation Network object modelled in the ENEON graph.

A network has relations to other networks. The relation between the EO network and the EV is not visible in the UML diagram because it is a backlink between the EV and the EO Network (see the section about the EVs).

In order to be useful for AD4GD, the original graph was updated and extended beyond research infrastructures.

Below is a summary of changes applied to the ENEON graph:

1. The list of ENEON themes has been extended to include additional terms that were required to describe new networks. The themes are not currently based on any controlled vocabularies and act as tags to label and categorise the ENEON entities.
2. The list of ENEON themes has been extended to include additional terms that were required to describe new networks. The themes are not currently based on any controlled vocabularies and act as tags to label and categorise the ENEON entities.
3. ENEON graph entities have been additionally labelled with the EU Green Deal thematic areas using the "GDThematicArea" key. The tags are based on the eight themes defined by the EU Commission (1-Increasing climate ambition; 2-Clean, affordable and secure energy; 3-Industry for a clean and circular economy; 4-Energy and resource efficient buildings; 5-Sustainable and smart mobility; 6-Farm to fork; 7-Biodiversity and ecosystems; 8-Zero-pollution, toxic-free environments)<sup>4</sup>.
4. Societal Benefit Areas (SBA) tags in the original ENEON graph were duplicated in both EV and EO network entities. It was decided to use SBA tags only in EVs. If an observation network does not connect to any EVs and does not consequently contribute to any SBAs, that would mean that new EVs should be defined to create this connection. Potentially, some EVs could be empty nodes acting as placeholders to be defined by domain experts.

<sup>4</sup> [https://ec.europa.eu/commission/presscorner/detail/en/ip\\_20\\_1669](https://ec.europa.eu/commission/presscorner/detail/en/ip_20_1669)



5. The list of geographic extent terms has been updated to the coverage vocabulary from Giuliani et al., 2020 [1]<sup>5</sup>. Author’s classifications (Local, National, Regional, Global) are based on literature search in scientific libraries such as Science Direct, Scopus, Web of Science and Google Scholar.
6. Network types have been revised and updated to include terms with clear definitions as listed in Table 1.
7. The graph has been revised to include network types (SDG, EV, EONetwork) and subtypes (e.g., Federation, Data Centre) to label EONetwork classifications as defined in Table 1.
8. SDGs recorded in the original ENEON graph have been revised to include the latest SDG indicators (e.g., SDG 13.2.2).
9. The original ENEON graph included 21 GEOBon-defined Essential Biodiversity Variables (EBVs) (<https://geobon.org/ebvs/what-are-ebvs/>). The graph has been updated to include 24 EBVs based on the EuropaBON EBV descriptions (<https://github.com/EuropaBON/EBV-Descriptions>). These EuropaBON EBV descriptions have been revised to elicit high-level EBVs and corresponding EBV products.
10. Kunming-Montreal Global Biodiversity Framework (GBF) 2030 targets (<https://www.cbd.int/gbf/targets>) have been added to the ENEON graph and connected to the corresponding SDGs (<https://www.cbd.int/sbstta/sbstta-24/post-2020-sdg-linkages-en.pdf>).
11. New atmospheric and air quality observable properties (e.g., Air Temperature, Ozone, PM10) have been added to the ENEON graph and connected to the corresponding SDGs.
12. New air quality networks including citizen science networks (e.g., Sensor.Community) have been added to the ENEON graph and connected to the observable properties that they measure.
13. New water level and water quality observable properties (e.g., Groundwater Water Level, Water Flow, pH) have been added to the ENEON graph and connected to the corresponding SDGs.
14. New water quality networks have been added to the ENEON graph and connected to the observable properties that they measure.
15. New citizen science biodiversity networks have been added to the ENEON graph and connected to the EBVs.

Network Name	Definition
Federation	Includes SystemOfSystems, NetworkOfNetworks and Data Space networks.
EO Agency	A government or non-governmental organisation that specialises in EO activities, which involve the acquisition, processing, analysis, and dissemination of data and information about the Earth's environment using remote sensing technologies.
Data Centre	Serves as a centralised location for the storage, management, processing, and dissemination of digital data. The primary purpose of a data centre is to store and process vast amounts of data and make it available to users or other systems.
EO Infrastructure	Constitutes the comprehensive framework of systems, technologies, and resources necessary to support the entire life cycle of EO data – from data acquisition to processing, analysis, management, and dissemination.

<sup>5</sup> <https://www.mdpi.com/2306-5729/5/4/100>

EO Network	Refers to a collection of individual sensors, instruments, or platforms distributed across various geographic locations and operated by different organisations or entities. These sensors or instruments are part of a network that collaboratively collects data from different regions or specific monitoring sites. EO Networks are often established to address specific research objectives or monitoring needs. For example, a network of weather stations deployed across a country, sharing real-time weather data.
EO System	Refers to a more comprehensive and integrated framework that encompasses the entire suite of components, instruments, platforms, data management systems, and data processing capabilities necessary for remote sensing and Earth monitoring. An EO System is designed to acquire, process, analyse, and disseminate Earth observation data on a broader scale, covering larger areas or even the entire planet. An EO System typically includes satellites, ground-based stations, data centres, data processing algorithms, and other infrastructure elements. It involves the coordination and integration of multiple sensors, platforms, and data collection methods to provide a comprehensive view of the Earth's environment.
EO Program	Refers to a coordinated and systematic set of activities and initiatives aimed at acquiring, processing, analysing, and utilising EO data for specific purposes. Earth Observation Programs are typically organised by governmental bodies, international organisations, or research institutions to address specific scientific, environmental, or societal objectives. An EO Program should have a start and end date.

**Table 1: ENEON graph network types definitions.**

Seven air quality in situ networks have been added to and described in the ENEON graph: AirGradient, IQAir, Copernicus Atmosphere Monitoring Service (CAMS), Sensor.Community, European Environment Agency (EEA) air pollution, OpenAQ and PurpleAir (see Figure 5). Of these, AirGradient, Sensor.Community, and PurpleAir are in situ networks that aggregate data from sensors installed by community scientists, i.e., collect and share citizen science data.

The following atmospheric and air quality observable properties have been added to the ENEON graph: air temperature, air humidity, air pressure, PM1, PM2.5, PM10, O3, NO, NOx, NO2, CO, CO2, SO2, TVOCs (see Figure 7).

At present, there are no essential variables defined for air quality, therefore the observable properties have been directly linked with the in situ networks that they are measured by (see Figure 5, 6, 7).

The observable properties are linked to the corresponding SDG indicators using the “estimated with” key of SDG indicators in the JSON file. (see Figure 9). In the case of air quality, there are six related SDG indicators: SDG 3.d.1, SDG 3.9.1, SDG 9.4.1, SDG 11.6.2, SDG 13.2.1, and SDG 13.2.2 (see Figure 5).

Each observable property is linked to the corresponding RAINBOW server definition using the “url” key (Figure 6). This ensures that each property has a clear definition based on controlled vocabularies to support data interoperability.



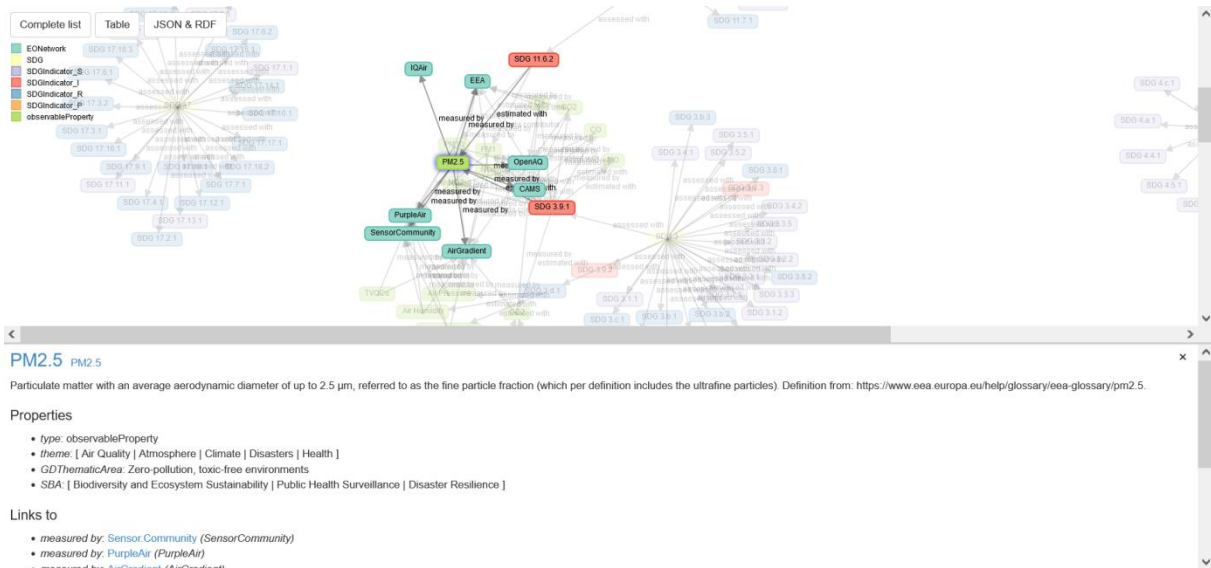


Figure 7: ENEON graph air quality networks - PM2.5 observable property example.

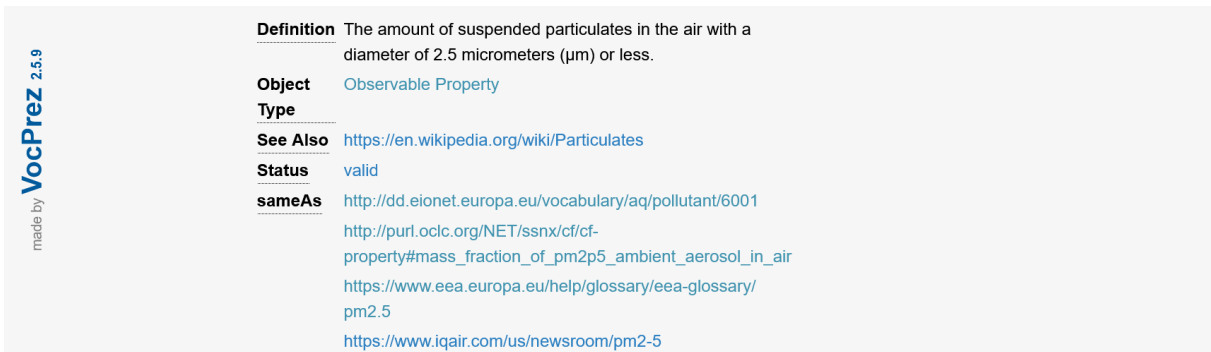


Figure 8: RAINBOW server - PM2.5 observable property definition.

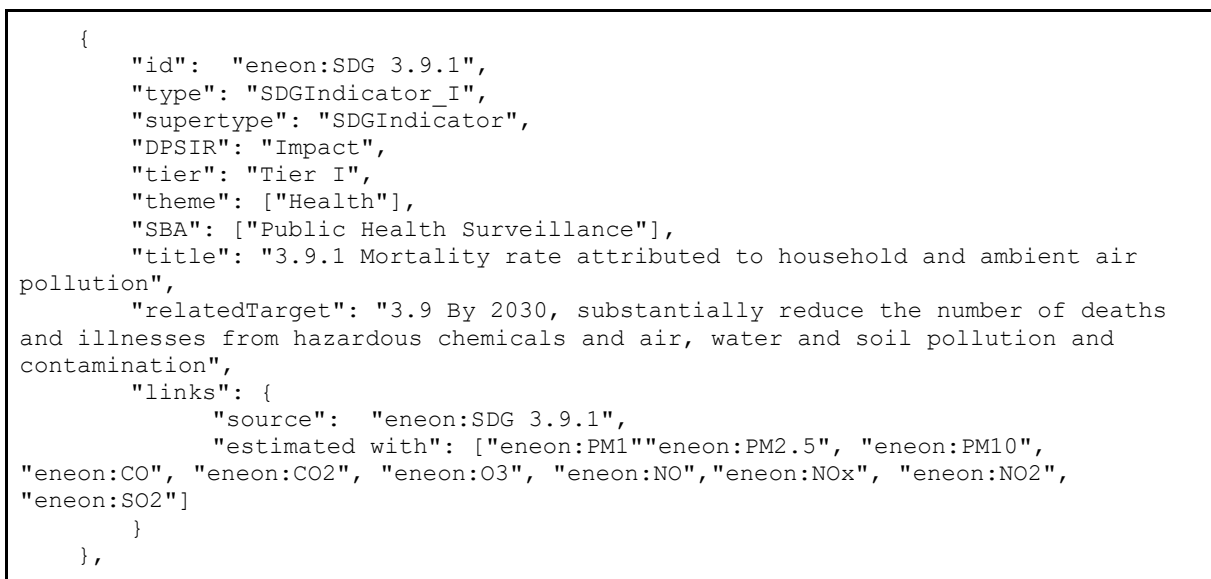


Figure 9: ENEON graph JSON - SDG 3.9.1 with corresponding air quality observable properties.

Figure 10 shows an example of OpenAQ air quality network that aggregates data from EEA and AirGradient in situ networks where EEA provides official air quality data and AirGradient provides citizen science data. OpenAQ aggregates data for six air pollutants: PM2.5, PM10, NO2, O3, SO2, and CO.

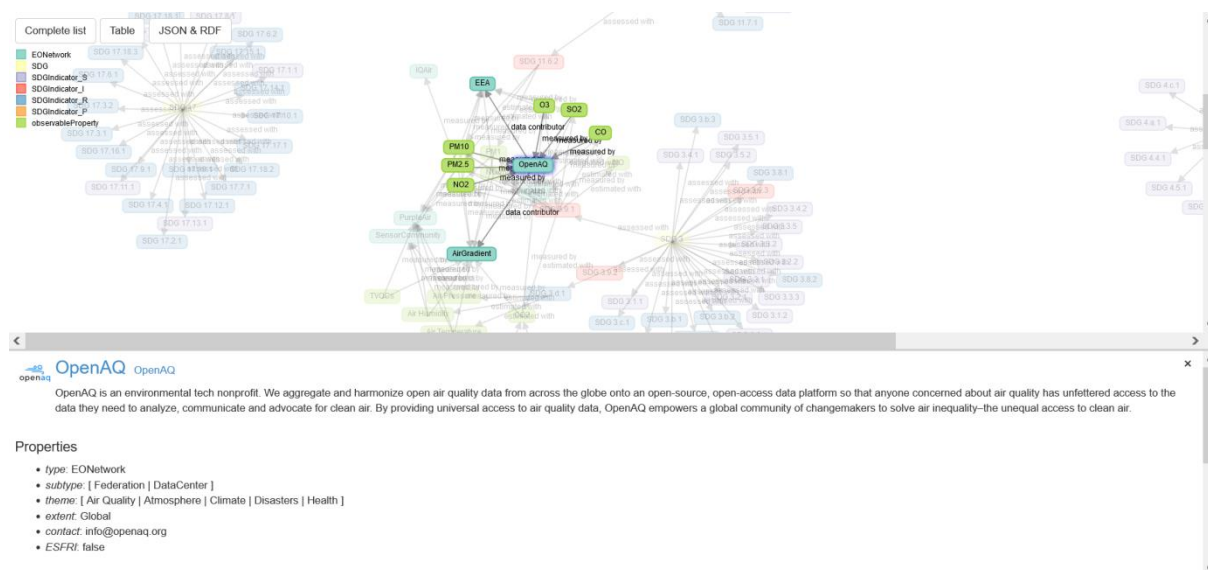


Figure 10: ENEON graph air quality networks - OpenAQ network example.

Two water quality in situ networks have been added to and described in the ENEON graph: WasserportalBerlin<sup>6</sup> and IGB Freshwater Research and Environmental Database (FRED) (see Figure 12 and Figure 13). Both networks provide official water level and water quality data.

The following water level and water quality observable properties have been added to the ENEON graph: Surface Water Level, Water Flow, Water Temperature, Water Electrical Conductivity pH, Water Oxygen Level, Water Oxygen Saturation, Water Oxygen Concentration, Soil Moisture, Soil Temperature, Groundwater Water Level, Water Turbidity, Chlorophyll a, and Phycocyanin (see Figure 11, Figure 12 and Figure 13).

At present, there are no essential variables defined for water quality, therefore the observable properties have been directly linked with Wasserportal and FRED in situ networks that measure these properties (see Figure 12 and Figure 13).

The observable properties are linked to the corresponding SDG indicators using the “estimated with” key of SDG indicators in the JSON file. (see Figure 14). In the case of water quality, there are eight related SDG indicators: SDG 3.9.2, SDG 6.1.1, SDG 6.3.2, SDG 6.4.1, SDG 6.4.2, SDG 6.5.1, SDG 6.5.2, and SDG 6.6.1 (see Figure 11).

<sup>6</sup> <https://wasserportal.berlin.de/start.php>

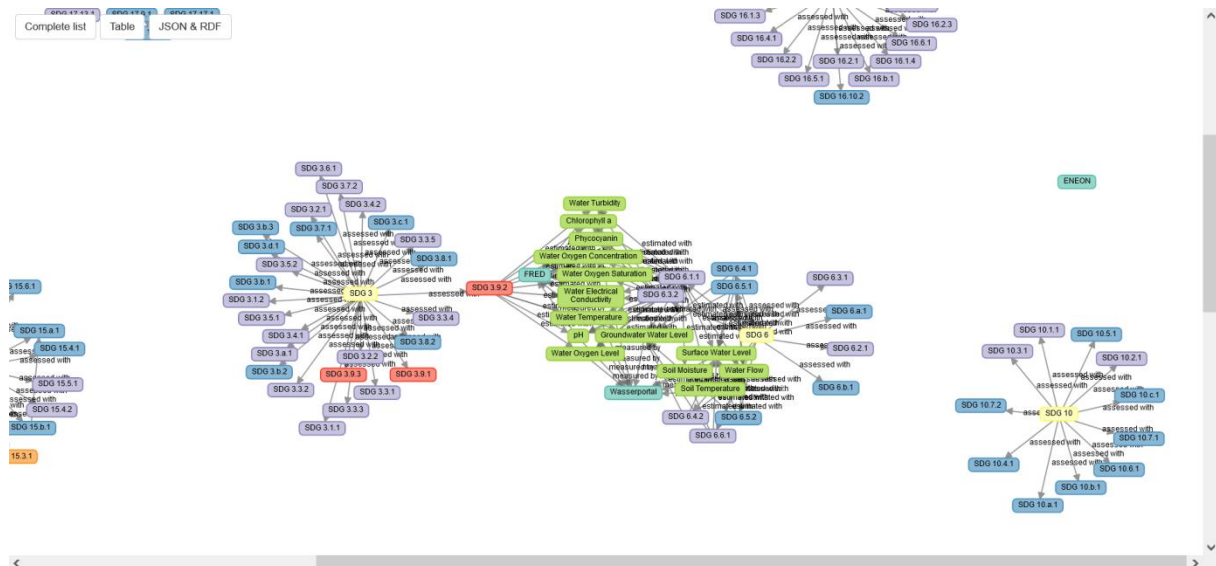


Figure 11: ENEON graph water quality networks, water quality observable properties and connections to SDGs.

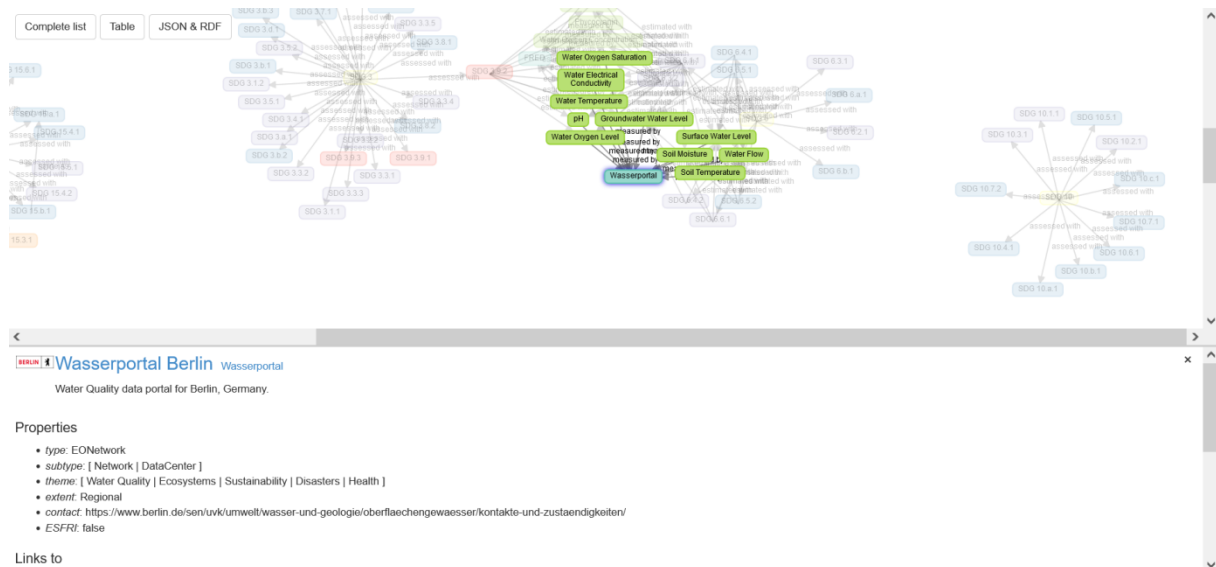


Figure 12: ENEON graph water quality networks - Wasserportal network example.



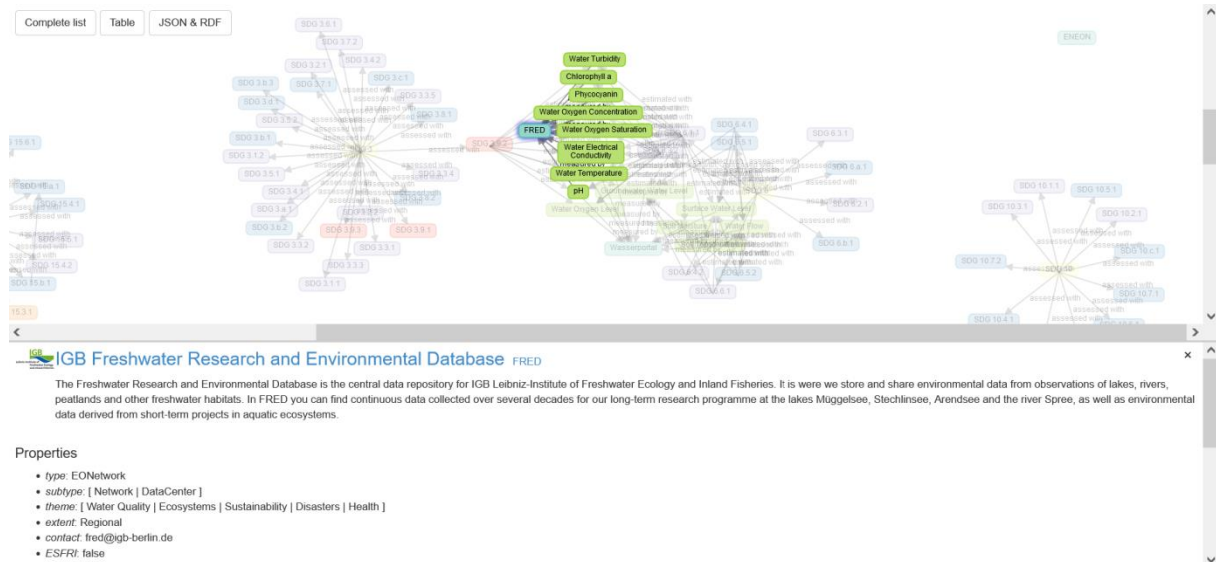


Figure 13: ENEON graph water quality networks - FRED network example.

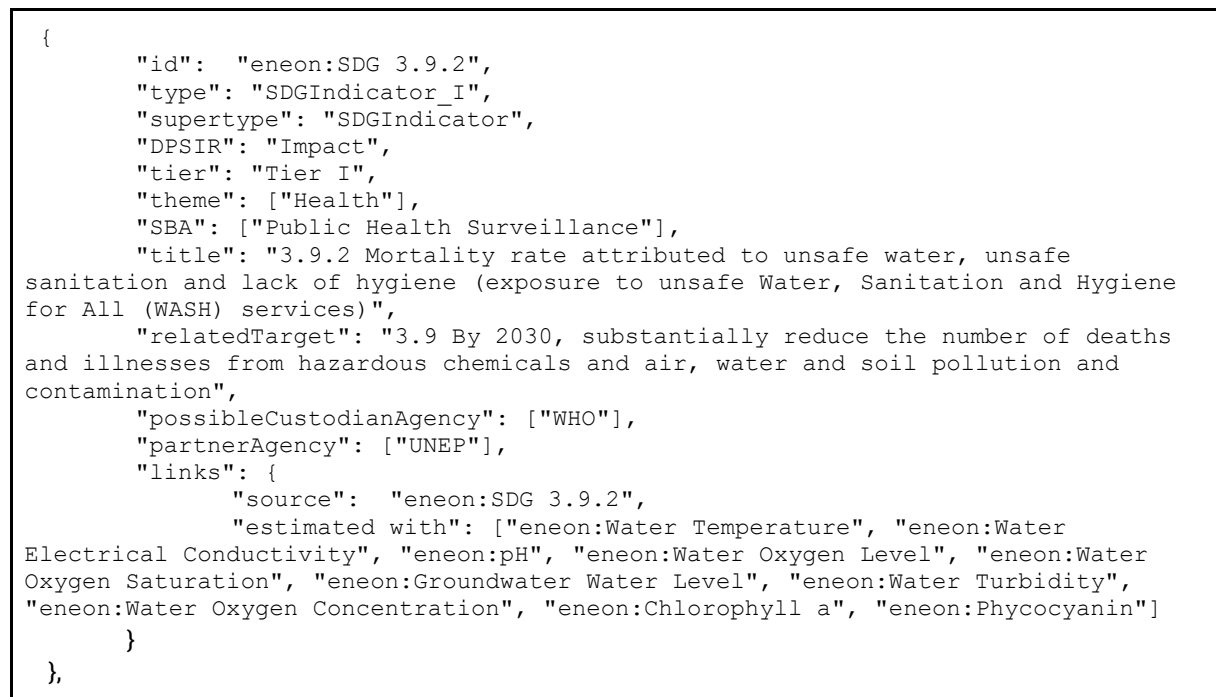


Figure 14: ENEON graph JSON - SDG with corresponding water quality observable properties.

Figure 15 and Figure 16 show Kunming-Montreal Global Biodiversity Framework (GBF) 2030 targets and target connections to SDG indicators. At present, GBF 2030 indicators are still under community development and review. Via connections to SDG indicators, it can be automatically identified which networks and EVs already contribute to estimating the GBF targets.

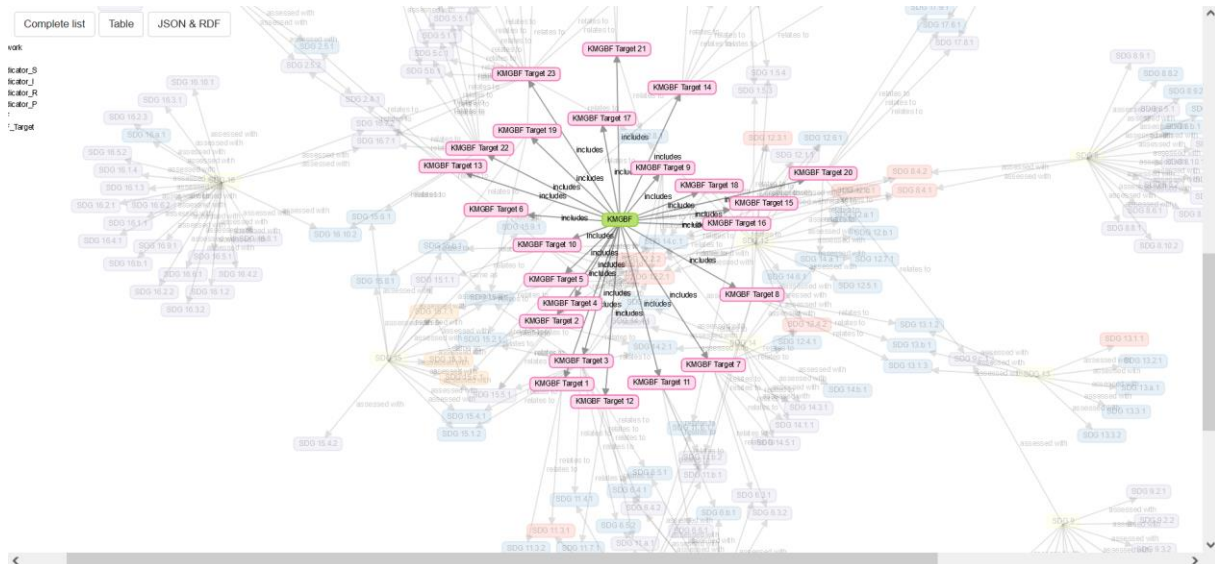


Figure 15: ENEON graph Kunming-Montreal GBF 2030 targets.

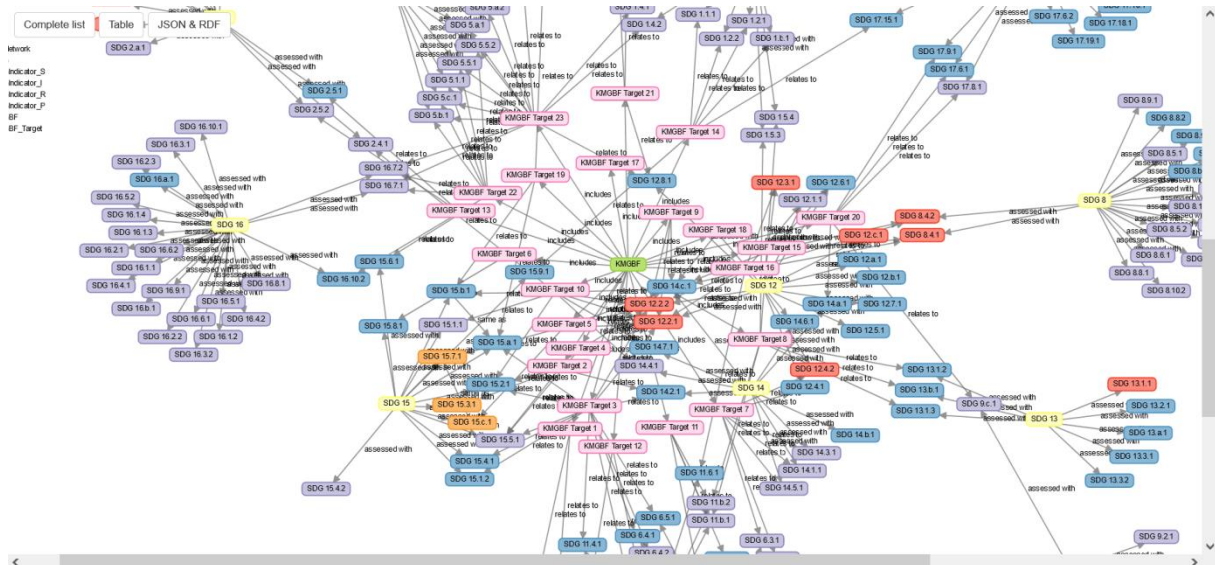


Figure 16: ENEON graph Kunming-Montreal GBF 2030 targets and connections to SDGs.

Table 2 and Figure 17 show the updated 24 EBVs based on GEOBON EBV definitions and the latest EuropaBON EBV descriptions (<https://github.com/EuropaBON/EBV-Descriptions>). Section 4 provides full details on the harmonisation process applied to elicit high-level EBVs and corresponding EBV products.

EuropaBON descriptions include 14 high-level EBVs (see Figure 18) and GEOBON definitions include 21 EBVs (see Figure 19).

EBV Name	EuropaBON	GEOBON	EBV Description
Functional composition	Yes	No	–
River Connectivity/Free river flow	Yes	No	–
Structural complexity	Yes	No	–



Community abundance	Yes	Yes	The abundance of organisms in ecological assemblages.
Ecosystem distribution	Yes	Yes	The horizontal distribution of discrete ecosystem units.
Ecosystem disturbances	Yes	Yes	Abrupt deviances in the functioning of the ecosystem from its regular dynamics.
Ecosystem phenology	Yes	Yes	Duration and magnitude of cyclic processes observed at the ecosystem level, such as in vegetation activity, phytoplankton blooms, etc.
Ecosystem Vertical Profile	Yes	Yes	The vertical distribution of biomass in ecosystems, above and below the land surface.
Genetic diversity	Yes	Yes	The variation in DNA sequences among individuals of the same species.
Phenology	Yes	Yes	Presence, absence, abundance or duration of seasonal activities of organisms.
Primary productivity	Yes	Yes	The rate at which energy is transformed into organic matter primarily through photosynthesis.
Species abundances	Yes	Yes	Predicted count of individuals over contiguous spatial and temporal units addressing the global extent of a species group.
Species distributions	Yes	Yes	The species occurrence probability over contiguous spatial and temporal units addressing the global extent of a species group.
Taxonomic/phylogenetic diversity	Yes	Yes	The diversity of species identities, and/or phylogenetic positions, of organisms in ecological assemblages.
Effective population size	No	Yes	The number of individuals in an idealized population that will exhibit the same amount of genetic diversity loss as the population under consideration.
Genetic differentiation (number of genetic units and genetic distance)	No	Yes	Divergence in genetic composition (identity and frequencies of alleles) among multiple populations.
Inbreeding	No	Yes	Mating between related individuals.
Interaction diversity	No	Yes	The diversity and structure of multi-trophic interactions between organisms in ecological assemblages.
Live cover fraction	No	Yes	The horizontal (or projected) fraction of area covered by living organisms, such as vegetation, macroalgae or live hard coral.
Morphology	No	Yes	The variation in physical attributes of organisms of the same species.
Movement	No	Yes	Behaviors related to the spatial mobility of organisms such as dispersal and migration routes.
Physiology	No	Yes	Chemical or physical functions promoting organism fitness and responses to environment.
Reproduction	No	Yes	Sexual or asexual production of new individual organisms ('offspring') from parents. Examples: Age at maturity, number of offspring, lifetime reproductive output.

Trait diversity	No	Yes	The diversity of functional traits of organisms in ecological assemblages.
-----------------	----	-----	--

Table 2: Updated list of 24 EBVs recorded in the ENEON graph.

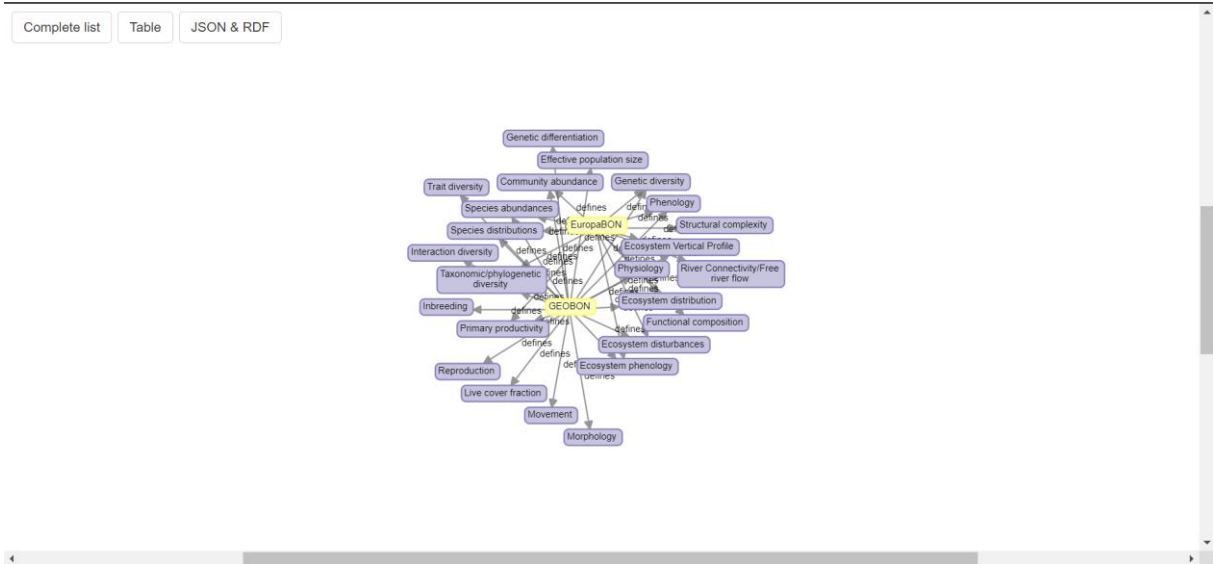


Figure 17: ENEON graph EuropaBON and GEOBON Essential Biodiversity Variables (EBVs).

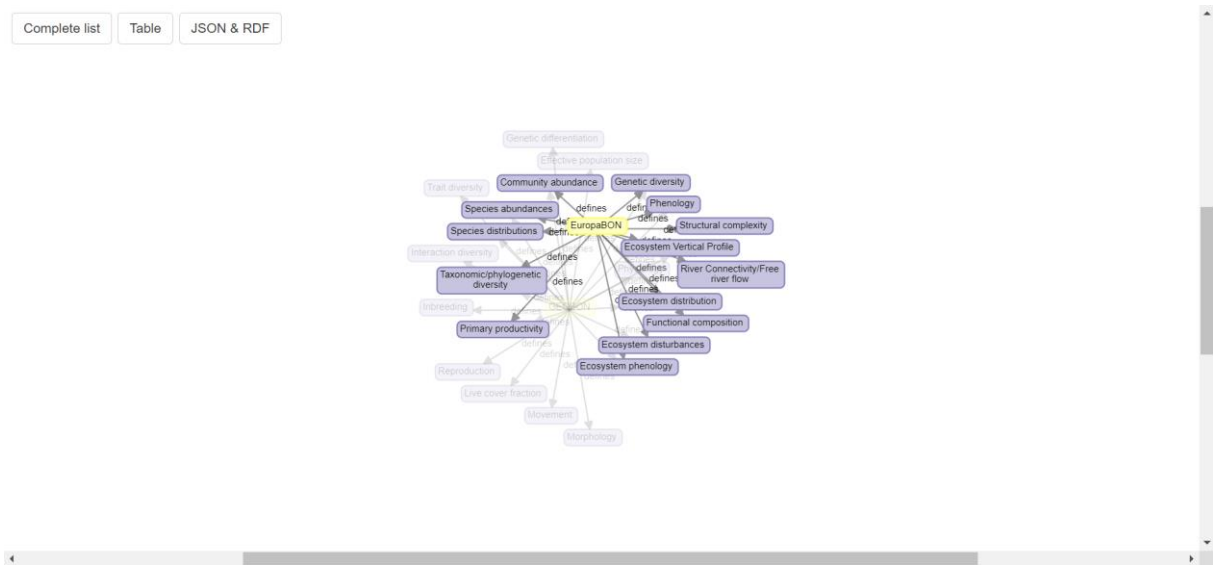


Figure 18: ENEON graph EuropaBON EBVs.

Complete list Table JSON & RDF

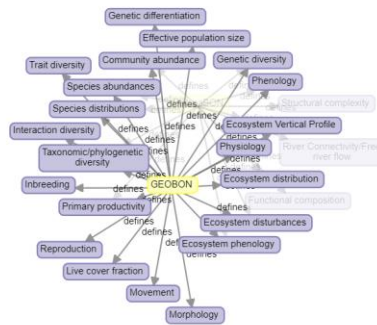
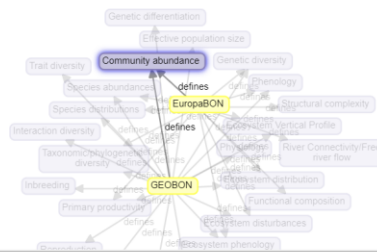


Figure 19: ENEON graph GEOBON EBVs.

Figure 20 and Figure 21 show an example of Community Abundance EBV and corresponding EBV product as a graph and JSON code representations.

Complete list Table JSON & RDF



```

• EV_class: Community composition
• product[0]:
  ◦ Product Name: Community abundance Freshwater phytoplankton
  ◦ Realm (Constraint 1): Freshwater
  ◦ Product Definition: Composition of phytoplankton in the European catchments and rivers network system (ECRINS) as measured by the Ecological Quality Ratio (EQR) or based on total abundance (biolume), taxonomic composition of indicator species, or bloom intensity (e.g. maximum biomass of cyanobacteria or percentage of cyanobacteria of the total biomass for all taxa)
  ◦ Observed Property Reference: https://github.com/EuropaBON/EBV-Descriptions/wiki/Freshwater-Community-composition-of-phytoplankton
  ◦ Entity: phytoplankton
  ◦ Constraint 2: in lakes
  ◦ Constraint 3:
  ◦ EBV Metric: Relative abundance
  ◦ Spatial Resolution Unit: https://qudt.org/vocab/quantitykind/Area
  ◦ Temporal Resolution Unit: 1 year
product[1]:
    
```

Figure 20: ENEON graph EuropaBON EBVs - Community Abundance EBV and EBV product example.

```

...
{
  "id": "eneon:Community abundance",
  "type": "EBV",
  "supertype": "EV",
  "theme": ["Biodiversity", "Biosphere", "Ecosystems", "Environment"],
  "GDThematicArea": "Biodiversity and ecosystems",
  "SBA": ["Biodiversity and Ecosystem Sustainability"],
  "EV_class": "Community composition",
  "title": "Community abundance",
  "description": "The abundance of organisms in ecological assemblages.",
  "url": "",
  "product": [{
    
```

```

        "Product Name": "Community abundance Freshwater phytoplankton",
        "Realm (Constraint 1)": "Freshwater",
        "Product Definition": "Composition of phytoplankton in the European
catchments and rivers network system (ECRINS) as measured by the Ecological
Quality Ratio (EQR) or based on total abundance (biovolume), taxonomic composition
of indicator species, or bloom intensity (e.g. maximum biomass of cyanobacteria or
percentage of cyanobacteria of the total biomass for all taxa)",
        "Observed Property Reference": "https://github.com/EuropaBON/EBV-
Descriptions/wiki/Freshwater-Community-composition-of-phytoplankton",
        "Entity": "phytoplankton",
        "Constraint 2": "in lakes",
        "Constraint 3": "",
        "EBV Metric": "Relative abundance",
        "Spatial Resolution Unit":
"https://qudt.org/vocab/quantitykind/Area",
        "Temporal Resolution Unit": "1 year"
    },
    {
        "Product Name": "Community abundance Freshwater phytoplankton",
        "Realm (Constraint 1)": "Freshwater",
        "Product Definition": "Composition of phytoplankton in the European
catchments and rivers network system (ECRINS) as measured by the Ecological
Quality Ratio (EQR) or based on total abundance (biovolume), taxonomic composition
of indicator species, or bloom intensity (e.g. maximum biomass of cyanobacteria or
percentage of cyanobacteria of the total biomass for all taxa)",
        "Observed Property Reference":
"https://github.com/EuropaBON/EBV-Descriptions/wiki/Freshwater-Community-
composition-of-phytoplankton",
        "Entity": "phytoplankton",
        "Constraint 2": "in lakes",
        "Constraint 3": "",
        "EBV Metric": "Ecological Quality Ratio (EQR)",
        "Spatial Resolution Unit":
"https://qudt.org/vocab/quantitykind/Area",
        "Temporal Resolution Unit": "weekly-monthly during growing season"
    },
    ...
    
```

Figure 21: ENEON graph JSON - Community Abundance EBV and EBV product example.

Two key biodiversity citizen science in situ networks have been added to and described in the ENEON graph: iNaturalist and eBird (see Figure 22). Both networks contribute data to GBIF and support FAIR guiding principles.

Figure 22: ENEON graph citizen science biodiversity networks - iNaturalist network example.

### 3 IN-SITU DATA SOURCE TYPES IN THE GDSS

In-situ Earth observation data can be produced by several actors, including the scientific community via the Research Infrastructures (RIs), the public administration geoportals of Member States offering socio-economic and INSPIRE datasets, the citizen science (CitSci) initiatives generating crowdsourced information, and the more recent Internet of Things (IoT). The following sections describe the role of the aforementioned in-situ data providers in the context of the GDSS.

#### 3.1 THE ROLE OF THE RESEARCH INFRASTRUCTURES

As defined by the Directorate-General for research and innovation of the European Commission<sup>7</sup>, the term Research Infrastructures refers to “facilities that provide resources and services for research communities to conduct research and foster innovation in their fields, including [...] knowledge-related facilities such as collections, archives or scientific data infrastructures; computing systems, communication networks and any other infrastructure of a unique nature and open to external users, essential to achieve excellence in R&I; they may, where relevant, be used beyond research, for example for education or public services and they may be 'single sited', 'virtual' or 'distributed'.”

Environmental Research Infrastructures are key to provide systematic and coherent datasets needed for research addressing many aspects relevant for the European Green Deal such as: climate, natural resources, health, food security, biodiversity, and sustainable use of the marine, freshwater and soils. They target both the scientific community as well as support the environmental monitoring activities conducted by agencies across Europe. In particular, they are well positioned to give hard facts on the efficiency of the European Union and its Member States mitigation and adaptation actions.

Environmental research as a scientific domain focuses on understanding how the Earth system works at various spatial and temporal scales. Environmental research requires comprehensive observations integrated with relevant experimental and modelling approaches which are essential for understanding and predicting the Earth’s environmental system functions. A federated approach to IT resources and e-science facilities is also necessary together with liable data policies compliant with the FAIR principle.

The environmental RIs already play an important role in supporting the scientific community and the society at large by<sup>8</sup>:

- generating coherent, comparable, and sustained time-series of key environmental variables;
- providing accurate large datasets and new solutions to share these data for increased scientific and technical knowledge that underpin the construction of tools supporting decision making and development of efficient regulations and policies;
- delivering essential data for more reliable communication to the public on events such as volcanic eruptions, earthquakes, poor air quality and extreme weather as well as information on biodiversity impacts;
- opening access to environmental big data from space-based and in-situ observations as a key driver for the development of new services and for promoting activities in the private sector.

However, creative research beyond the traditional silos is needed to develop innovative solutions for protective and preventive measures and to identify the optimal mechanisms for their implementation.

In the geosphere, the ESFRI Landmark EPOS ERIC integrates several hundreds of individual RIs in the Solid Earth domain covering seismology, near-fault observatories, geodetic data and products, volcano

<sup>7</sup> REGULATION (EU) 2021/695 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 28 April 2021 establishing Horizon Europe – the Framework Programme for Research and Innovation, laying down its rules for participation and dissemination, and repealing Regulations (EU) No 1290/2013 and (EU) No 1291/2013 <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32021R0695&from=EN>

<sup>8</sup> <https://roadmap2021.esfri.eu/landscape-analysis/section-1/environment/>

observations, satellite data, geomagnetic observations, anthropogenic hazards, geological information and modelling, multi-scale laboratories, and geo-energy test-beds for low-carbon energy. Other geosciences RIs and projects are operated globally; on-going work is conducted to ensure the required coordination and integration. These include: SoWa RI for comprehensive research and understanding of soil and water ecosystems in context of sustainable landscape use and the Joint Research Centre – European Soil Data Centre (ESDAC) – Soil Atlas of Europe.

In the atmosphere, RIs includes: Long-term atmospheric observation platforms (Figure 23): the ESFRI Landmark ACTRIS; the ESFRI Landmark IAGOS (Airborne, lower atmosphere); the ESFRI Landmark ICOS ERIC; the ESFRI Landmark EISCAT\_3D (upper atmosphere). International atmospheric monitoring networks in support of the European and international policies such as European Monitoring and Evaluation programme (EMEP) established in support of the Long-Range Transboundary Air Pollution (UNECE LRTAP) Convention, the Global Air Passive Sampling (GAPS) and the MONitoring NETwork (MONET) passive air monitoring programmes supporting the effectiveness evaluation of the UN Stockholm Convention on Persistent Organic Pollutants (POPs), the Global Mercury Observation System (EU project GMOS) in support of the UN Minamata Convention, Arctic Monitoring and Assessment Programme (AMAP), etc. Most of these networks were not formally established as RIs such as the IGOSP (Integrated Global Observing Systems for Persistent Pollutants) project focused on the integration of real-time monitoring data from various platforms, development of modelling tools and advanced global cyber-infrastructure for data sharing and interoperability, the Global Observation System for Mercury (GOS4M) and the Global Observation System for POPs (GOS4POP).

In the hydrosphere, much of the current science is done relying on access to existing water bodies, i.e. without specific and dedicated large-scale Research Infrastructures. The ESFRI Project DANUBIUS-RI supports interdisciplinary research in river-sea systems. It is the only physical pan-European Research Infrastructure devoted to support research on transitional zones between coastal marine and freshwater areas. The ESFRI Landmark LifeWatch ERIC as the only e-RI, extends its area of interest to the whole freshwater environment. The ESFRI Landmark AnaEE (H&F); also offers access to experimental facilities in freshwater environments, applying an ecosystem services approach to key sectors including food security, human welfare and the wider bio-economy.

In the biosphere, there are Observatories and Monitoring Facilities such as the ESFRI Landmark ICOS ERIC and the ESFRI Landmark EMBRC ERIC (H&F), related IA ASSEMBLE Plus, the ESFRI Projects DANUBIUS-RI and eLTER RI, the IAs INTERACT and JERICO-S3, SIOS (Integrating all observations, terrestrial, marine and atmosphere at Svalbard). In addition, we can find facilities for in-situ and in vivo experimentation: the ESFRI Landmark AnaEE (H&F), the IAs AQUACOSM-plus and HYDRALAB+. Some biological collections, data infrastructures and reference data exist: the ESFRI Project DiSSCo (linked IA Synthesis PLUS), the ESFRI Landmarks ELIXIR and MIRRI (H&F), and the IA BiCIKL. Finally, there are two e-Infrastructures for data, analysis and modelling: the ESFRI Landmark LifeWatch ERIC, and the IAs IS-ENES3 and SeaDataCloud. The ESFRI Project eLTER RI is tackling a broad spectrum of ecological challenges, based on observations that enable understanding ecosystems using an approach of ecological integrity, including the socio-ecological dimension. The ESFRI Landmark AnaEE (H&F) alternatively, provides experiments instead of observations, with stronger focus on agriculture and food security from a defined set of ecological and societal challenges and has a more anthropocentric approach. The ESFRI Landmark LifeWatch ERIC has a cross-domain approach and a focus on the Grand Challenges of preserving biological diversity and of protecting ecosystem health. LifeWatch ERIC is an e-Infrastructure that enables knowledge-based solutions to environmental managers by providing access to a multitude of sets of data, services and tools about the role of biodiversity in ecosystem functioning and conservation. The focus is made in the construction and operation of Virtual Research Environments (VRE), backed by strong computational capacity and metadata catalogues. On another side, the ESFRI Landmark ICOS ERIC also has a cross-domain approach to enable understanding the carbon cycle and to provide necessary information on the land-ecosystem exchange of CO<sub>2</sub>, CH<sub>4</sub> and N<sub>2</sub>O with the atmosphere. The digitization of biological collections and the connection to genomics is a game changer in biodiversity research aiming to close the taxonomic gap, which still is a major limitation to



biodiversity knowledge. The ESFRI Project DiSSCo is developing tools and resources to speed up digitization and virtual access to Natural History Collections (NHC).

In general, each RI has enough capacity to become a node in the Green Deal Data Space as contributors of observations and data of interest for the Green Deal Data Space stakeholders. Most of them are already able to share their data using standard protocols that can be adjusted for the GDDS requirements.

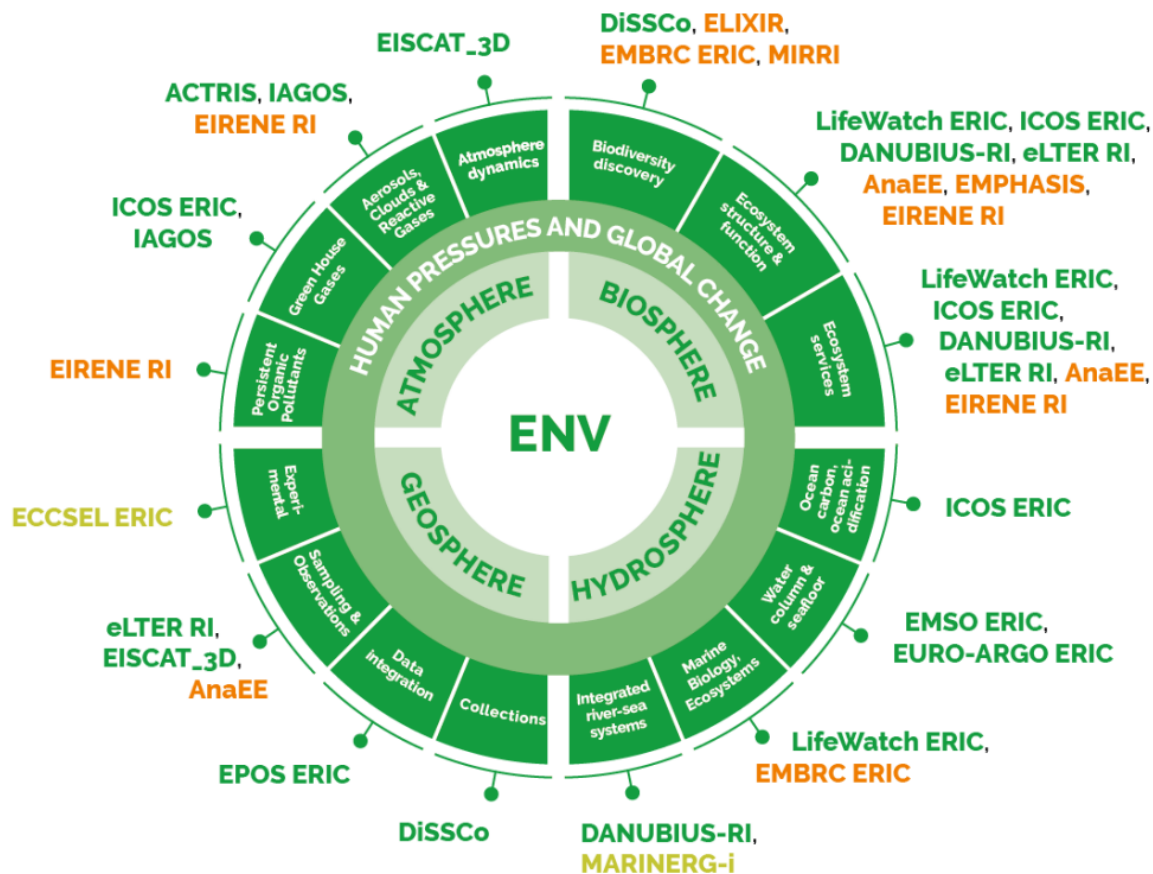


Figure 23: Environmental research infrastructures included in the ESFRI roadmap 2021.

Coordinated effort of RIs across various ESFRI domains and ENV sub-domains is required to support implementation of the European Green Deal and Horizon Europe Missions (e.g. Soil Health Mission) and protect the environment, soils and biodiversity. Carbon Capture and Storage (CCS), for instance, is an important element for the long-term storage of Carbon dioxide from the atmosphere (ERA-Net ACT and the ESFRI Landmark ECCSEL ERIC (ENE)). Green Infrastructures have been demonstrated to enhance nature protection and biodiversity beyond protected areas, to deliver ecosystem services such as climate change mitigation and re-creation, to prioritise measures for defragmentation and restoration in the agri-environment and regional development context, and to find land allocation trade-offs and possible scenarios involving all sectors.

### 3.2 THE ROLE OF CITIZEN SCIENCE

Citizen science is scientifically sound research conducted with the participation from members of the public (who are sometimes referred to as amateur/nonprofessional scientists). There are variations in the exact definition of citizen science, with different individuals and organisations having their own specific interpretations of what citizen science encompasses. However, many citizen science activities have a data gathering component, sometimes improving the scientific community's capacity (e.g. when a USA citizen

collects observations of blossoming of trees in his garden every year and contribute these observations to the USA national phenology network) or gathering evidence of an issue citizens are sensible to and want to demonstrate and force a change (e.g. when a neighbourhood gets organised because their streets are too noisy and they want to complain to the mayor with a compelling case supported by data). In other cases, citizen activities are recorded as data by their mobile devices and wearables or by their direct interaction in social media. This data can also be crowdsourced and aggregated for purposes initially not foreseen (if conveniently authorised by them). Whatever the motivations, the value of the in-situ observations is multiplied if citizen science individual observations (voluntarily or involuntarily) are recorded using standardised protocols and are contributed to a system capable of redistributing an aggregated dataset of in-situ data that complements official sources.

There are many citizen science portals that act as aggregation facilities or data portals such as the iNaturalist for biodiversity observations, GlobeAtNight for light pollution at night, Sensor.Community for air quality monitoring. There are also citizen science knowledge hubs and platforms such as SciStarter (global coverage) and EU-Citizen.Science that provide capabilities for discovering citizen science projects. At present, SciStarter and EU-Citizen.Science platforms contain databases of 1,549 and 260 citizen science projects respectively, however neither of these platforms provide formal descriptions of or direct access or links to the data collected by these projects. EU-Citizen.Science platform offers an API to automatically extract metadata about the registered CitSci projects, yet each project needs to be inspected manually to identify what data is being collected and offered for external use.

EU-Citizen.Science platform contains 211 active projects, 32 are completed, 11 are periodically active, 4 are not yet started, and 2 are on hold (Figure 24). As shown in Figure 25, most common participation tasks are observation (60 projects), data entry (55 projects), and identification (45 projects). Note that the projects can include multiple participation tasks. Regarding project topics (Figure 26), Ecology & Environment (102 projects, 13%), Biodiversity (91 projects, 11%), and Education (64 projects, 8%) are the top three of 29 topics covered. The topics used here are of varying scale and can be aggregated, e.g. Birds, Long-term species monitoring, Insects & pollinators can all be combined under the Biodiversity theme. As with the participation tasks, the projects can also fall under multiple topics.

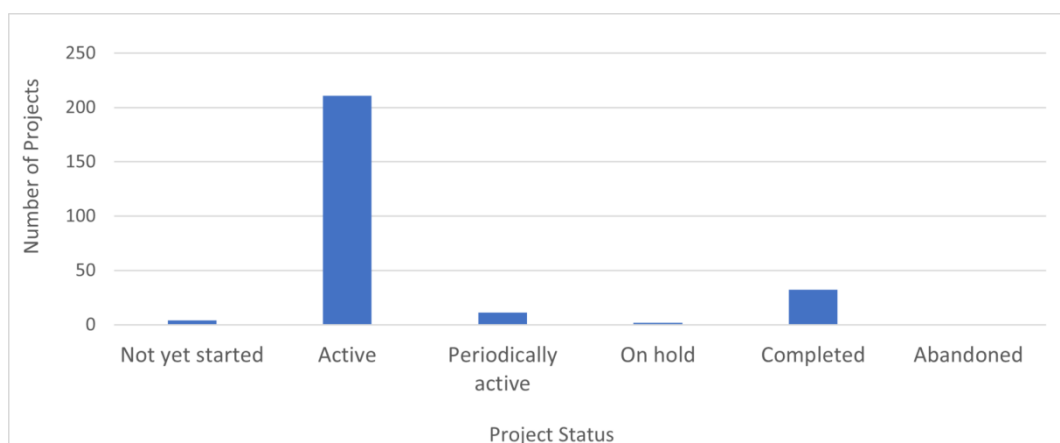


Figure 24: EU-Citizen.Science platform projects status.



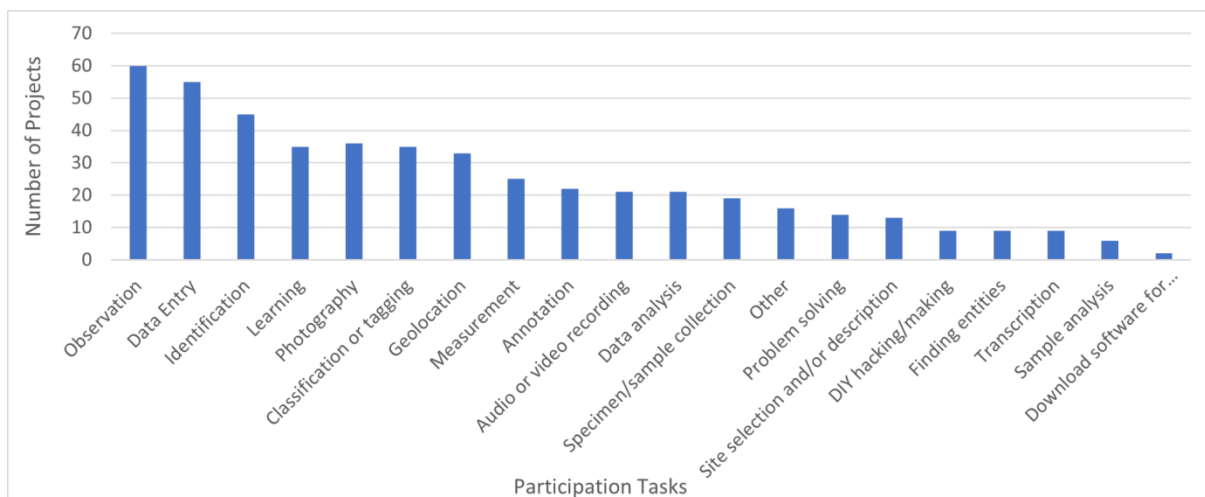


Figure 25: EU-Citizen.Science platform participation tasks.

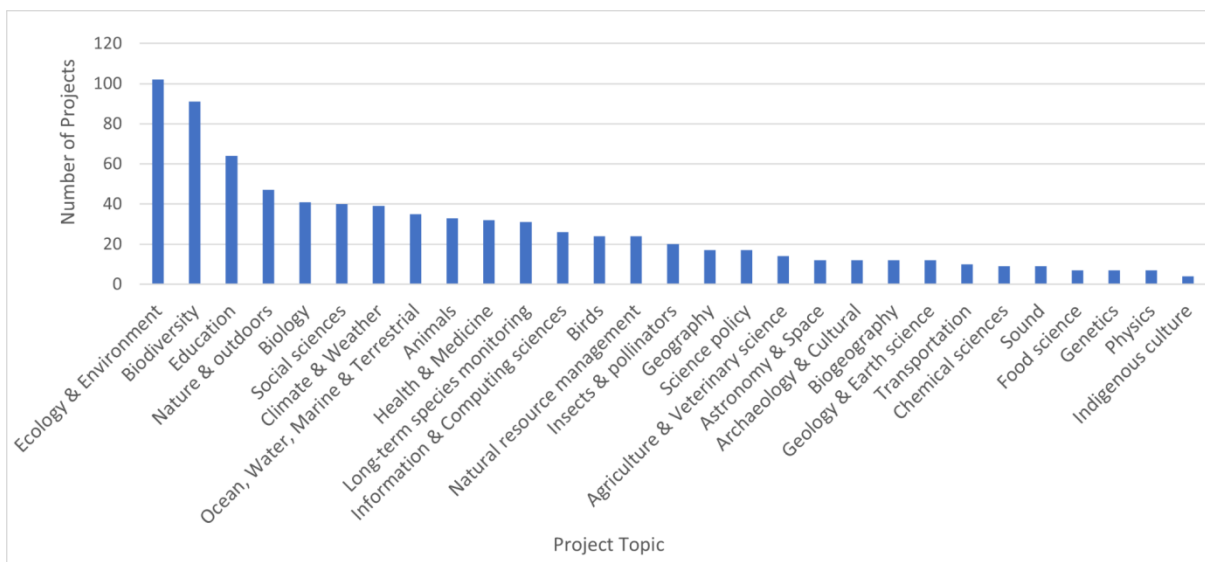


Figure 26: EU-Citizen.Science platform project topics.

The datasets from the citizen science aggregation portals can become interesting providers to the Green Deal Data Space. However, aggregation portals have some dynamics (some new appear every year while others are abandoned). In addition, it could be challenging for the aggregation portals to adopt the requirements for participating in the GDDS. Instead, we propose to create a reduced number of nodes that centralises the access to all citizen science datasets that can be relevant to the GDDS. We could start by setting up one node and experiment on how to implement the GDDS participation requirements.

### 3.3 THE ROLE OF INSPIRE AND ENVIRONMENTAL DATA

The Infrastructure for Spatial Information in the European Community (INSPIRE) was created under the Directive 2007/2/EC, which set the rules and technical guidelines for the establishment of such infrastructure based on the national spatial data infrastructures (SDI) operated by the Member States to enable data sharing across the EU to support environmental policies. INSPIRE defines in its Annexes I, II and III 34 spatial data themes (Figure 27). Annex 1 refer to basic geospatial reference data (e.g. geographical

names, standard grid system), Annex 2 refers to topographic data (e.g. elevation, geology), and Annex 3 refers to nature and socioeconomic data (land use, habitats, sea regions).

Annex I	Annex III
<ul style="list-style-type: none"> <li>• Coordinate reference systems</li> <li>• Geographical grid systems</li> <li>• Geographical names</li> <li>• Administrative units</li> <li>• Addresses</li> <li>• Cadastral parcels</li> <li>• Transport networks</li> <li>• Hydrography</li> <li>• Protected sites</li> </ul>	<ul style="list-style-type: none"> <li>• Statistical units</li> <li>• Buildings</li> <li>• Soil</li> <li>• Land use</li> <li>• Human health and safety</li> <li>• Utility and governmental services</li> <li>• Environmental monitoring facilities</li> <li>• Production and industrial facilities</li> <li>• Agricultural and aquaculture facilities</li> <li>• Population distribution demography</li> <li>• Area management/restriction /regulation zones &amp; reporting units</li> <li>• Natural risk zones</li> <li>• Atmospheric conditions</li> <li>• Meteorological geographical features</li> <li>• Oceanographic geographical features</li> <li>• Sea regions</li> <li>• Bio-geographical regions</li> <li>• Habitats and biotopes</li> <li>• Species distribution</li> <li>• Energy Resources</li> <li>• Mineral resources</li> </ul>
Annex II	
<ul style="list-style-type: none"> <li>• Elevation</li> <li>• Land cover</li> <li>• Ortho-imagery</li> <li>• Geology</li> </ul>	

Figure 27: INSPIRE data themes.

INSPIRE provides ancillary data that can help give context to the environmental studies (such as elevation, sea regions), help us to interpret the spatial distribution of some environmental variables (e.g. protected areas, land use, production and industrial facilities) and other datasets that were historically collected by public administrations that help us to characterise the environment (e.g. habitats and biotopes, soil, species distribution).

The INSPIRE Roadmap finished by December 2021. Since then, the INSPIRE effort is to advance towards the GDDS taking advantage of the experience gathered from the INSPIRE implementation. While INSPIRE focuses mainly on data from the public sector, the scope of GDDS is broad and considers the role of industry and the private sector to address emerging technological trends in the SDI. Two relevant documents were published by the Joint Research Centre (JRC) providing ideas and a vision for the evolution of INSPIRE in the context of the GDDS, cited as follows:

- *INSPIRE - A Public Sector Contribution to the European Green Deal Data Space*<sup>9</sup>. This report defines the future evolution of INSPIRE in the context of the emerging European data spaces envisioned by the EU Data Strategy, and, in particular, to the GDDS.
- *From Spatial Data Infrastructures to Data Spaces: A Technological Perspective on the Evolution of European SDIs*<sup>10</sup>. The authors of this paper provide a vision for a modernised architecture of a Spatial Data Infrastructure (SDI) in line with the new data spaces framework..

Thus, INSPIRE is currently in the phase of transformation to align to the new EC objectives with the data spaces architectures and rules and serving data for the Green Deal priorities. It links with a series of European policies and regulations, such as the Data Implementing Act and the High-Value Datasets.

<sup>9</sup> [https://publications.jrc.ec.europa.eu/repository/bitstream/JRC126319/inspire\\_web\\_single\\_pages.pdf](https://publications.jrc.ec.europa.eu/repository/bitstream/JRC126319/inspire_web_single_pages.pdf)

<sup>10</sup> <https://doi.org/10.3390/ijgi9030176>

The regulation (EU) 2023/138<sup>11</sup> defines in its Annex a list of specific High Value Datasets (HVD) and the arrangements for their publication and re-use in the context of the (EU) 2019/1024 Open Data Directive<sup>12</sup> (formerly Public Sector Information (PSI) Directive 2003/98/EC). Also known as the Implementing Act of HVD, it mentions the geospatial, environmental and climate datasets within the scope of the INSPIRE data themes defined in Annexes I-III to Directive 2007/2/EC.

In the HDV regulation, datasets are classified in six thematic data categories, as listed below:

1. Geospatial
2. Earth observation and environment
3. Meteorological
4. Statistics
5. Companies and company ownership
6. Mobility

Only Geospatial and Earth Observation and environment categories, refers to groups of particular INSPIRE data themes. The Earth Observation and environment category encompasses other legislations requesting monitoring of environmental aspects such Air Quality, Water, etc. that are collected by the administration as well as for the research infrastructures as in-situ datasets. Other HDV such as Meteorological datasets and Statistical data are relevant for the GDDS.

### 3.4 THE ROLE OF IOT

The Internet of things (IoT) describes physical objects (or groups of such objects) with sensors, processing ability, software and other technologies that connect and exchange data with other devices and systems over the Internet or other communications networks. Traditional fields of embedded systems, wireless sensor networks, control systems, automation (including home and building automation), independently and collectively enable the Internet of things. There is a part of the IoT for the consumer market with products pertaining to the concept of the "smart home" that is out of scope of the GDDS. Smart cities use Internet of Things (IoT) sensors in urban areas to collect data and automate systems such as traffic, energy use, and waste management. While this can be laterally connected to the GDDS, we should focus on applications of IoT in in-situ environmental monitoring. In these applications, sensors detect and measure every type of environmental change in different domains such as air and water pollution (by inexpensive sensors that allow for frequent sampling), extreme weather monitoring (allows early detection and early responses to prevent loss of life and property), water safety, endangered species protection, commercial farming, and more<sup>13</sup>:

As described in Section 2, the ENEON graph has been extended with new in situ networks including air quality and water quality networks that offer valuable IoT data for the GDDS. While the majority of these networks are well-established, there are some issues that need to be addressed before data can be integrated into the GDDS in a FAIR way.

**Findable:** citizen science IoT networks such as AirGradient, Sensor.Community and PurpleAir do not offer standardised metadata records that can be registered or indexed in a searchable resource. For water quality, Wasserportal offers descriptions of data as web pages and FRED supplies PDF metadata documents for their data, however both networks do not offer standard-compliant metadata documents that can be catalogued for data discovery.

**Accessible:** while IQAir, OpenAQ, and PurpleAir IoT networks offer APIs to access data, these APIs do not adhere to a common standard such as STA or STApplus which makes data acquisition and data integration

<sup>11</sup> [https://eur-lex.europa.eu/eli/reg\\_impl/2023/138](https://eur-lex.europa.eu/eli/reg_impl/2023/138)

<sup>12</sup> <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1561563110433&uri=CELEX:32019L1024>

<sup>13</sup> [https://www.tutorialspoint.com/internet\\_of\\_things/internet\\_of\\_things\\_environmental\\_monitoring.htm](https://www.tutorialspoint.com/internet_of_things/internet_of_things_environmental_monitoring.htm)

more challenging. AirGradient, Sensor.Community, FRED and Wasserportal do not offer any data APIs and data needs to be downloaded from historic archives in CSV or XML formats.

**Interoperable:** none of the air or water quality IoT networks discussed above use controlled vocabularies to describe data fields. Additionally, data field names are not consistent across data providers, for example, particulate matter 10 (PM10) is labelled as PM10 by OpenAQ, pm10.0\_atm by PurpleAir and P1 by Sensor.Community. This makes data integration more challenging since field names need to be reviewed and renamed manually. There is also a danger of misinterpreting the exact meaning of fields' names.

**Reusable:** as already noted, most IoT networks discussed here do not offer standardised metadata records, therefore the usage licence cannot be easily identified and needs to be located within the data providers' web pages or documentation. Additionally, the data is not uniquely identified which makes the recording of provenance information challenging.

There are a number of concerns about the risks in the growth of IoT technologies and products, especially in the areas of privacy and security. This is particularly important when trying to integrate this kind of data in the GDDS, as trust and privacy are fundamental elements in the data spaces.

While the IoT is a provider of in-situ, the analysis and problems of the IoT will be reported in the deliverables related to WP3.

### 3.5 HIGH PRIORITY DATASETS AND SERVICES WITH IN-SITU ON FOCUS AS IDENTIFIED BY THE GREAT PROJECT BLUEPRINT FOR THE GDDS

The Green Deal Data Space Foundation and its Community of Practice (GREAT) project, funded by the Digital Europe program, and ended last 29/02/2024, has provided a blueprint for implementing the GDDS in terms of 1/ community of practice, 2/ architecture, 3/ priority datasets, 4/ governance and 5/ the roadmap. Regarding the priority datasets, the GREAT project has performed an inventory of datasets and services relevant for the Green Deal that involve a variety of data providers. Specifically, a total of 406 high priority data-services and 94 high priority datasets have been selected, based on the information provided by the GREAT use cases and initiatives and on specific prioritisation criteria. The full inventory is available at: <https://www.greatproject.eu/dataset-data-services-inventory/>. The purpose of the following paragraphs is to analyse the GREAT data inventory concerning the quantity and classify the entries by the types of in-situ data defined in the previous sections.

- **High priority data-services**

Building on the inventory of 406 data-services, a list of top 30 high priority data-services representing the high overall prioritisation score for the Green Deal has been produced and published in the GREAT deliverable "D5.2: Prioritised Datasets and Gaps", available at: <https://www.greatproject.eu/wp-content/uploads/2024/04/D5.2-EGD-Prioritised-Datasets-and-Gaps-Initial-Inventory-plus-all-Reference-Use-Cases.pdf>. To facilitate a deeper understanding of the prioritized services, we analysed the top 30 data services and classified them according to the type of data provided (categorizing whether they offer in-situ data services or remote sensing data services), and the specific type of in-situ data source (identifying if the in-situ data originates from RI, CitSci, INSPIRE datasets, or IoT). The results of this classification are presented in Table 3, which enhances the original table from the GREAT D5.2 deliverable. Each entry in the table includes the data-service, its owner, the data access URL, the type of data source, and the type of in-situ data when applicable (being the two last fields those added by AD4GD).

Id	Data service	Service owner	Data access URL	Data source	In-situ data type
----	--------------	---------------	-----------------	-------------	-------------------

1	Copernicus Open Access Hub	EC - Copernicus	<a href="https://scihub.copernicus.eu/">https://scihub.copernicus.eu/</a>	Remote sensing	n/a
2	Copernicus Global Land Service	EC – Copernicus	<a href="https://land.copernicus.eu/global/products/">https://land.copernicus.eu/global/products/</a>	Remote sensing	n/a
3	EEA Data Hub	EC - EEA	<a href="https://www.eea.europa.eu/en/datahub">https://www.eea.europa.eu/en/datahub</a>	Remote sensing & In-situ	INSPIRE
4	INSPIRE	EC	<a href="https://inspire-geoportal.ec.europa.eu/pdv_home.html">https://inspire-geoportal.ec.europa.eu/pdv_home.html</a>	In-situ	INSPIRE
5	EEA Indicators	EC - EEA	<a href="https://www.eea.europa.eu/en/analysis/indicators">https://www.eea.europa.eu/en/analysis/indicators</a>	Remote sensing & In-situ	RI and IoT
6	ECMWF forecasts	ECMWF	<a href="https://www.ecmwf.int/en/forecasts/datasets/catalogue-ecmwf-real-time-products">https://www.ecmwf.int/en/forecasts/datasets/catalogue-ecmwf-real-time-products</a>	Remote sensing & In-situ	(Models)
7	Central Data Repository	EC - EEA	<a href="https://cdr.eionet.europa.eu/">https://cdr.eionet.europa.eu/</a>	Remote sensing & In-situ	INSPIRE
8	USGS Earth Explorer	US Government	<a href="https://earthexplorer.usgs.gov/">https://earthexplorer.usgs.gov/</a>	Remote sensing	n/a
9	Urban Atlas	EC – Copernicus	<a href="https://land.copernicus.eu/en/products/urban-atlas">https://land.copernicus.eu/en/products/urban-atlas</a>	Remote sensing & In-situ	INSPIRE
10	Data Catalogue Destination Earth	EC - JRC	<a href="https://data.jrc.ec.europa.eu/dataset">https://data.jrc.ec.europa.eu/dataset</a>	Remote sensing & In-situ	RI
11	EMODNET	EC - EMODnet	<a href="https://emodnet.ec.europa.eu/geonetwork/srv/eng/catalog.search#/search?resultType=details&amp;sortBy=sortDate&amp;from=1&amp;to=20">https://emodnet.ec.europa.eu/geonetwork/srv/eng/catalog.search#/search?resultType=details&amp;sortBy=sortDate&amp;from=1&amp;to=20</a>	In-situ	RI
12	GWIS Statistics Portal	EC - JRC & Copernicus	<a href="https://gwis.jrc.ec.europa.eu/apps/gwis.statistics/">https://gwis.jrc.ec.europa.eu/apps/gwis.statistics/</a>	Remote sensing	n/a
13	Corda	EC - EEA & Copernicus	<a href="https://corda.eea.europa.eu/">https://corda.eea.europa.eu/</a>	In-situ	INSPIRE
14	EO Browser	SentinelHub by Planet Labs	<a href="https://www.sentinel-hub.com/explore/eobrowser/">https://www.sentinel-hub.com/explore/eobrowser/</a>	Remote sensing	n/a
15	GWIS Situation viewer	EC - JRC & Copernicus	<a href="https://gwis.jrc.ec.europa.eu/apps/gwis_current_situation/index.html">https://gwis.jrc.ec.europa.eu/apps/gwis_current_situation/index.html</a>	Remote sensing	n/a
16	Joint Research Centre Data Catalogue	EC - JRC	<a href="https://data.jrc.ec.europa.eu/dataset">https://data.jrc.ec.europa.eu/dataset</a>	Remote sensing & In-situ	INSPIRE
17	Official portal for European data	EC	<a href="https://data.europa.eu/data/datasets">https://data.europa.eu/data/datasets</a>	In-situ	INSPIRE
18	ISRIC - World Soil Information	ISRIC - World Soil Information	<a href="https://data.isric.org/geonetwork/srv/eng/catalog.search#/home">https://data.isric.org/geonetwork/srv/eng/catalog.search#/home</a>	Remote sensing & In-situ	RI
19	Eurostat	EC	<a href="https://ec.europa.eu/eurostat/data/database">https://ec.europa.eu/eurostat/data/database</a>	In-situ	INSPIRE
20	ESA CCI Data portal	ESA	<a href="https://climate.esa.int/en/data/#/search">https://climate.esa.int/en/data/#/search</a>	Remote sensing	n/a

21	Copernicus Data Space Ecosystem	EC – Copernicus	<a href="https://browser.dataspace.copernicus.eu/">https://browser.dataspace.copernicus.eu/</a>	Remote sensing	n/a
22	NOAA National Centre for Environmental Information - Paleoclimatology Data	US Government	<a href="https://www.ncei.noaa.gov/products">https://www.ncei.noaa.gov/products</a>	Remote sensing	n/a
23	Copernicus Marine Service	EC – Copernicus	<a href="https://data.marine.copernicus.eu/products">https://data.marine.copernicus.eu/products</a>	Remote sensing	n/a
24	Water Data & Maps	EC	<a href="https://water.europa.eu/">https://water.europa.eu/</a>	Remote sensing & In-situ	INSPIRE, RI
25	EEA Analysis and Data	EC - EEA	<a href="https://www.eea.europa.eu/en/analysis">https://www.eea.europa.eu/en/analysis</a>	Remote sensing & In-situ	INSPIRE
26	Publieke Dienstverlening Op de Kaart (PDOK)	Dutch Ministry of the Interior and Kingdom Relations	<a href="https://www.pdok.nl/datasets">https://www.pdok.nl/datasets</a>	Remote sensing & In-situ	INSPIRE
27	European Forest Fire Information system	EC – Copernicus	<a href="https://effis.jrc.ec.europa.eu/">https://effis.jrc.ec.europa.eu/</a>	Remote sensing & In-situ	RI
28	Copernicus Atmosphere Monitoring Service	EC – Copernicus	<a href="https://ads.atmosphere.copernicus.eu/cdsapp">https://ads.atmosphere.copernicus.eu/cdsapp</a>	Remote sensing	n/a
29	EPOS - European plate observing system	EPOS ERIC	<a href="https://www.ics-c.epos-eu.org/">https://www.ics-c.epos-eu.org/</a>	In-situ	RI
30	Ramsar Sites Information Service	Ramsar Convention on Wetlands	<a href="https://rsis Ramsar.org/">https://rsis Ramsar.org/</a>	In-situ	RI

**Table 3: High priority data-services defined by the GREAT project classified according to the type of data provided (whether they offer in-situ data services or remote sensing data services), and the specific type of in-situ data source (identifying if the in-situ data originates from RI, CitSci, INSPIRE datasets, or IoT).**

From the analysis of the 30 high-priority data-services, we identified the following categories:

- In-Situ data-services: 7 services.
- Combined In-Situ and Remote Sensing data-services: 14 services.
- Remote Sensing data-services only: 9 services.

For services providing in-situ data (either alone or in combination with remote sensing data), INSPIRE and RI data are the main in-situ data classes identified, while IoT and CitSci data are a minority. This panorama reflects that despite the growing interest in IoT and CitSci data, current monitoring programs still rely on traditional data acquisition methods. These traditional methods, while reliable, have significant limitations, including low spatiotemporal resolution and high operational costs. These issues could potentially be overcome with the adoption of low-cost sensors and ICT devices provided by IoT and Citizen Science technologies. Although these novel approaches offer promising solutions to the limitations mentioned before (Section 3.4 and Section 3.5 of the deliverable specifically address the advantages and challenges

associated with using IoT and CitSci), they are still emerging technologies and require further development to become established as monitoring practices.

- **High priority datasets**

In addition to the high priority data-services, the GREAT project also catalogued high priority datasets in its inventory, which contains references to the existing data and products required for the achievement of the GREAT use cases objectives. These prioritized datasets constitute a sample of what data may be needed for the GDDS. One of the metadata attributes documented for each dataset, is the source of data/generation type, that was classified in the following categories: Airbone laster-scanning, Biodiversity, Cadastral Data, Hidro-geology maps, In-situ, IUCN GET classification, Land cover, Model, Reanalysis data, Remote Sensing, Renewable energy production, River network, Script, Soil, Statistics, Topographically derived drainage networks and ancillary layers. From the mentioned data sources categories, those that fit the in-situ definition provided by GEO (a.k.a; any EO data that is not acquired using a satellite) are:

- Biodiversity (field surveys)
- Cadastral Data
- Hydro-geology maps
- In-situ observations
- Land cover (when derived from field surveys)
- Renewable energy production (when using ground-based monitoring)
- River network
- Soil
- Statistics
- Topographically derived drainage networks and ancillary layers

A total of 25 datasets concerning the above in-situ categories have been inventoried by the GREAT project, each contributing valuable in-situ data for the GDDS. After exploring the concrete types of in-situ data inventoried, it has been seen again that, despite their potential, the IoT and CitSci are not represented in the list. As mentioned as regards the data-services, this fact reflects the challenges of integrating and consolidating these new data acquisition methods, which can be addressed by the use and development of common standards, formats, and observational protocols, leading to more reliable monitoring systems that can support the Green Deal and other environmental initiatives.

## 4 ESSENTIAL VARIABLES AS A COMMON FRAMEWORK TO SEMANTICALLY TAG IN-SITU DATA

Essential Variables are an abstract concept that is commonly associated with a set of measurements (e.g. for the Climatic set there are the Essential Climate Variables). Each set of EVs is classified in themes (e.g. Atmosphere, Land, Ocean) and has several individual variables in it.

The Global Climate Observing System (GCOS), which was set up to ensure that the observations needed to address climate-related issues are readily available to all interested parties, specifies a total of 54 Essential Climate Variables (ECVs) of which about 60% can be addressed by satellite data and the rest of the 40% by in-situ observations. The following Figure 28 provides an overview of the GCOS Essential Climate Variables<sup>14</sup>.

<sup>14</sup> <https://gcos.wmo.int/en/essential-climate-variables/table>





Figure 28: GCOS Essential Climate Variables overview.

GCOS' goal is to provide comprehensive data and climate information on the total climate system. Currently, the Climate Data Store (CDS) offers free and open access to products associated with 22 ECVs. Not all the defined products have been produced so far.

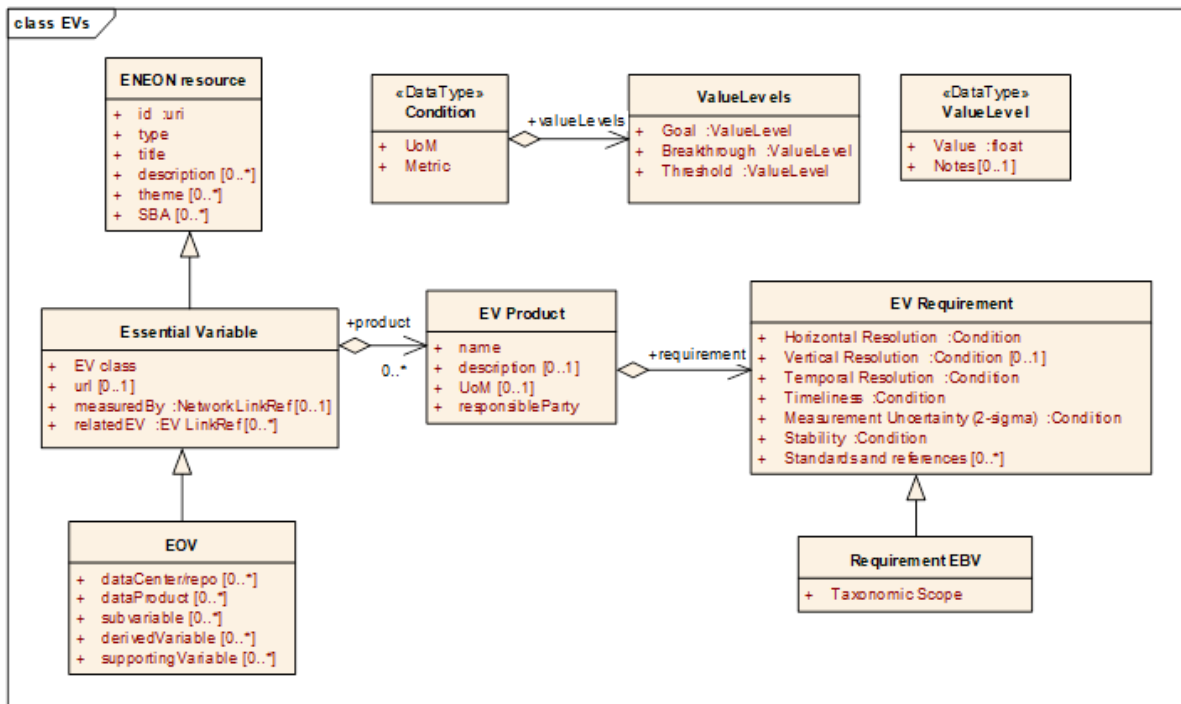


Figure 29: Data model used in ENEON for EVs and EV products.



The Essential Variable set used here is based on the GCOS definition but can be mapped to other EV sets and incorporates references to other EVs and to networks that measure the EV. The EV concept is extended into an EOVS that has a more detailed schema for defining EVs. EV sets can optionally be measured with one or more products as defined by GCOS, each one with a requirement class that has some generic parameters (resolutions, uncertainty, stability, etc). GEOBON (that is in charge of defining the EVs for biodiversity: EBV) also defined a very similar concept to the EV product, called “matrix” that has one extra parameter in the requirement class called “Taxonomic scope”. Figure 29 represents the data model used in ENEON for EVs and EV products. EuropaBON project (GA n° 101003553) has matured the definition of EBV products, specifying in detail all metrics and resolutions (spatial and temporal) related to a variable<sup>15</sup>. Other communities are also working on defining essential variables related to their domain mainly under the umbrella of GEO. This is the case of the Essential Agriculture Variables (EAV) defined by GEOGLAM<sup>16</sup>, the Essential Water Variables (EWV) defined by GEOGLOWS<sup>17</sup>, etc. More information on the current status of Essential Variables within GEO can be read at *Lehmann, Anthony et al. GEO community activity report: Mainstreaming EVs across GEO. 2023 doi: 10.13097/archive-ouverte/unige:166309*.

The EV framework and its EV products provides an excellent starting point to define a set of common concepts that can be used in remote sensing as well as in in-situ data (including Citizen Science). This gives us a way to semantically classify dataset around EV products independently of the technology and technique (in-situ, remote sensing, etc.) used to capture them. In practice, this is possible if the EVs and the EVs received a global and unique identifier. One possible approach is to use a different URI for each one. However, there is community recognized ontology of EV and its products that can be used. Meanwhile the ENEON graph provides a solution for the ontology that can be used in the GDDS.

Unfortunately the EV framework is not comprehensive as some variables that might be considered non-essential will not be present. A mechanism to detect these gaps and extend the EV or the EV products to areas that are not completely covered is needed. For example, air quality is not completely covered by ECV (in particular PM<sub>2.5</sub> is not defined as a ECV product). Other more specialized vocabularies include PM<sub>2.5</sub> such as <https://dd.eionet.europa.eu/vocabulary/aq/pollutant/> so we should extend the ECV if needed to cover the pilots in the project as well as the common variables in the air quality domain. Previous works on detecting gaps in EVs<sup>18</sup> should be reviewed and updated.

#### 4.1 USING EV VOCABULARIES TO IDENTIFY COMPATIBLE DATASETS

By its nature, in-situ data is fragmented. Data is acquired by different local actors with different sensors at different times many times by manual or semi-automatic processes. For example, in the marine domain, in-situ data rely on a very large number of national, regional, or global data providers. Ocean in-situ data observing networks (e.g. Argo vertical profilers, gliders, buoys or ships - research vessels and ships of opportunity) are owned or operated by a large number of institutions and agencies at national level, including, for example, national meteorological and oceanographic agencies or national research institutions<sup>19</sup>. A possible solution is to ensure a sustained operational flow of in-situ data into a single system by establishing interfaces with the global, regional, and coastal in-situ observing networks; a process that is called “aggregation”.

In the GDDS, data will cover different themes and variables and it should be carefully selected in order to be aggregated. If the data is not semantically tagged by the use of EV vocabularies, it will be almost impossible to be aggregated or selected as the right data as input for a numerical model in an automatic way.

<sup>15</sup> <https://github.com/EuropaBON/EBV-Descriptions/wiki>

<sup>16</sup> <https://agvariables.org/>

<sup>17</sup> <https://www.geogloWS.org/pages/working-group-3>

<sup>18</sup> [https://www.geoessential.eu/wp-content/uploads/2021/03/GEOessential-Deliverable-2.3\\_v3.pdf](https://www.geoessential.eu/wp-content/uploads/2021/03/GEOessential-Deliverable-2.3_v3.pdf)

<sup>19</sup> <https://insitu.copernicus.eu/state-of-play/copernicus-marine-environment-monitoring-service>

AD4GD is performing a harmonization and systematization of all Essential Variables based on the conceptual idea from GCOS of defining Variables and Products derived. In this process, an atomization of definitions and constraints for every variable is being done. By doing so, I-ADOPT<sup>20</sup> Framework Ontology defined by RDA revealed to be very appropriate and has been partially adopted. This framework is strongly focused on variables observed in environmental research to accurately encode what is measured, observed, derived, or computed in relation to the Earth systems, and to facilitate interoperability between existing variable description models.

#### 4.1.1 I-ADOPT FRAMEWORK ONTOLOGY APPLIED TO THE EBV

EuropaBON has developed a list of Essential Biodiversity Variables (EBVs) to advance the collection, sharing, and use of biodiversity information across Europe. EBVs provide a way to aggregate the many biodiversity observations collected through different methods such as in-situ monitoring or remote sensing. They can be visualized as biodiversity observations at one location over time, or in many locations, aggregated in a time series of maps. The EuropaBON EBV list has undergone a public review process to allow all interested stakeholders to provide input. The list has now been refined, and EBV workflow templates have been updated to offer guidance on how to collect and process data for each EBV. EBVs are scalable, meaning the underlying observations can be used to represent different spatial or temporal resolutions required for the analysis of biodiversity trends. When combined with social or economic information, EBVs can be used to identify indicators for biodiversity that reflect the interactions between human and natural systems. The initial EBV defined by GeoBON has been redefined as EBV classes. Each class has specialisations depending on the realm and other parameters that are defined in a way similar to the ECV products defined by GCOS. In this section we present a way to use I-ADOPT to parametrize EBV classes.

I-ADOPT is an ontology framework designed to facilitate interoperability between existing variable description models across research domains. It provides a common set of core components and relations to define machine-interpretable variable descriptions that re-use FAIR vocabulary terms.

I-ADOPT decompose a Variable in 3 classes Property, Entity, and Constraint, and specifies six object properties: hasProperty, hasObjectOfInterest, hasContextObject, hasMatrix, hasConstraint, constrains. The new v1.0, adds one optional new class (VariableSet) and four optional new object properties (hasApplicableProperty, hasApplicableObjectOfInterest, hasApplicableMatrix, hasApplicableContextObject). Thus, I-ADOPT enables the decomposition of complex observable properties into essential atomic parts represented through concepts in FAIR terminologies, serving as a common layer of abstraction to systematically align and extend concepts from different terminologies as needed. Annotating observational data with interoperable concepts to support findability and reusability of datasets across repositories.

---

<sup>20</sup> <https://i-adopt.github.io/>

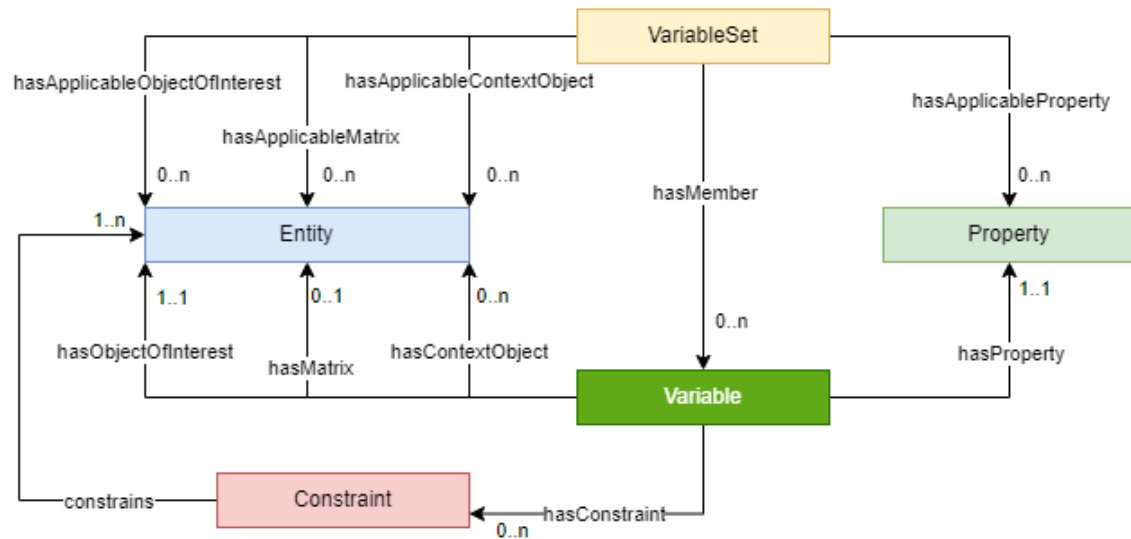


Figure 30: I-ADOPT Framework schema (extracted from <https://i-adopt.github.io/>)

The Variable is the top concept in I-ADOPT and represents the description of something observed or mathematically derived. In the case of EVs framework, this concept represents the EV definition, as it includes the observed target (I-ADOPT entity) and some properties (metrics) and constraints.

$EV = I-ADOPT:Variable = I-ADOPT:Entity + Object.Properties + Object.Constraints$

I-ADOPT:Entity is a similar concept to what OGC Sensor Things API defines as STA:ObservedProperty, and similar to what in OGC Observations, Measurements, and Samples is defined as OMS:ObservableProperty.

EV products are then at a hierarchically upper level, which includes the EVs and some new properties that define spatial and temporal resolution. These new properties don't exist in the current version of I-ADOPT, but AD4GD proposes to extend it to include a new concept defined as "Requirements". By doing so, EV products will be considered as:

$EV\ product = I-ADOPT:Variable + Object.Requirements$

If the I-ADOPT:Variable is assimilated to an EV definition, then the broader concepts of EV names and classes need to be considered as I-ADOPT:VariableSet. The VariableSet class can be connected to the Variable class using the property `ro:hasMember` from the OBO Relations Ontology <sup>21</sup>.

The UML representation of the AD4GD proposal for I-ADOPT extension to better fit the description of EVs is represented in the Figure 31.

<sup>21</sup> <https://obofoundry.org/ontology/ro.html>

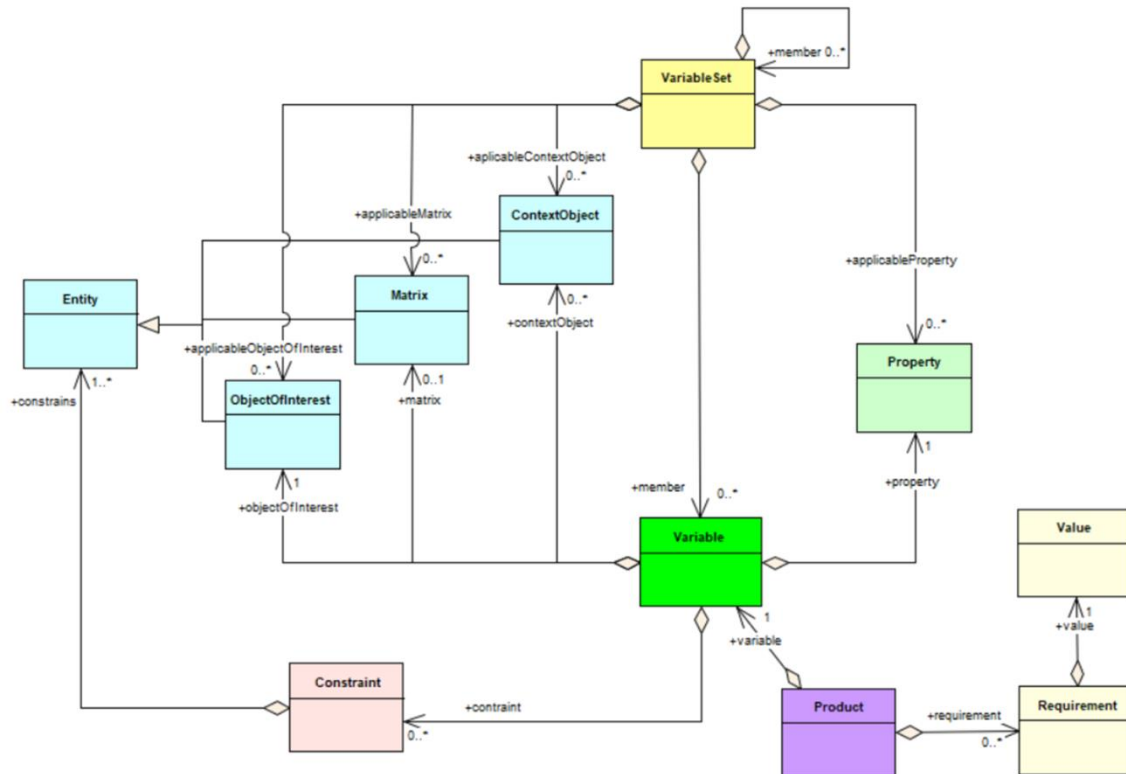


Figure 31: I-ADOPT UML schema expressing the AD4GD proposal of extension of I-ADOPT current version.

AD4GD is now a member of the RDA group on I-ADOPT definition and this proposal will be leveraged to the group for further discussion.

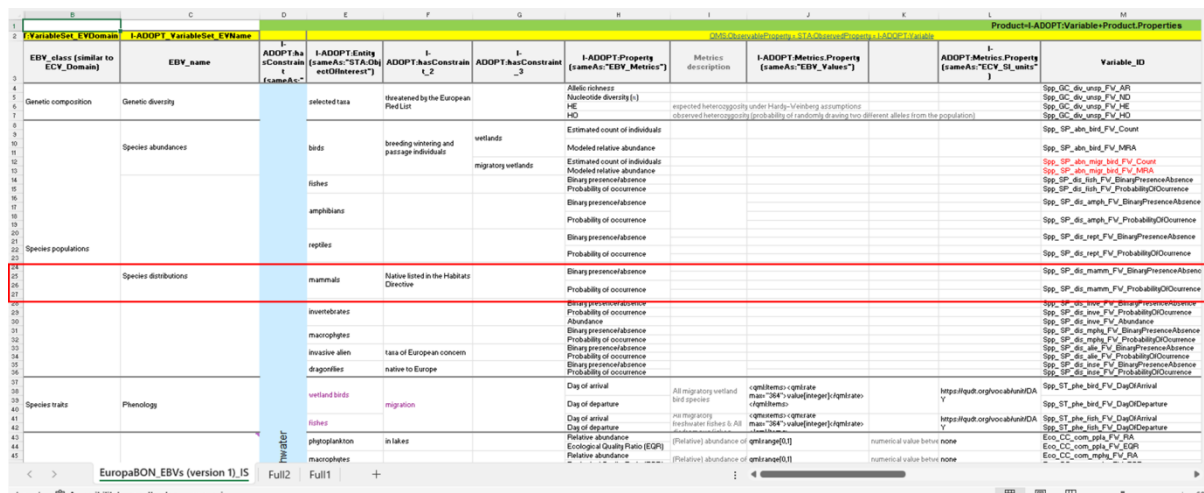
For the moment, only EBVs have been adapted to this new schema. If we take Species Distributions of Fresh water Mammals EVs as an example, the current definition of the variable from EuropaBON can be seen here:

## Species distributions of freshwater mammals

Key	Value
ID	Spp_SP_dis_mamm_FW
Realm	Freshwater
EBV class	Species populations
EBV name	Species distributions of freshwater mammals
Step in identification process	User & Policy Needs Assessment
Definition	The presence/absence or probability of occurrence of each European freshwater mammal species within contiguous spatial units (grid cells) over time.
Metric	<ul style="list-style-type: none"> <li>Binary presence/absence</li> <li>Probability of occurrence</li> </ul>
Spatial resolution unit	10x10km - 50x50km
Temporal resolution unit	3 to 6 years
Entity	Native freshwater mammal species listed in the Habitats Directive

Figure 32: Extracted from <https://github.com/EuropaBON/EBV-Descriptions/wiki/Freshwater-Species-distributions-of-freshwater-mammals>.

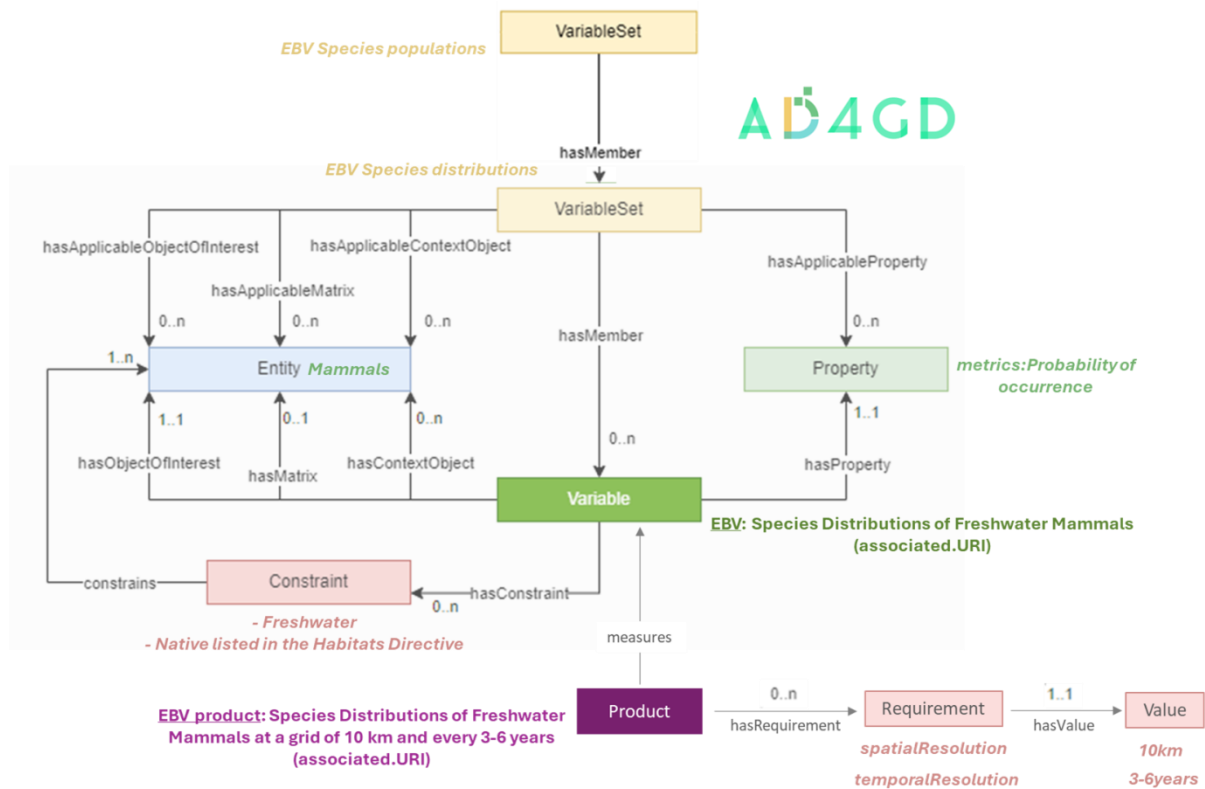
The same variable has been atomized following I-ADOPT and described in an EBV csv:



EBV class (similar to ECV_Domain)	EBV_name	I-ADOPT Entity (sameAs: "ST&Os")	I-ADOPT Property (sameAs: "EBV_Metrics")	Metrics description	I-ADOPT Metrics Property (sameAs: "EBV_Valuers")	Variable ID
Genetic composition	Genetic diversity	selected taxa	Allelic richness	Allelic richness	Spp_OC_dis_unsp_FW_AR	
Species abundances	birds	breeding wintering and passage individuals	Estimated species of individuals	Estimated species of individuals	Spp_SP_abn_birds_FW_MRA	
			Modeled relative abundance	Modeled relative abundance	Spp_SP_abn_mrg_birds_FW_MRA	
		wetlands	Estimated species of individuals	Estimated species of individuals	Spp_SP_abn_mrg_birds_FW_MRA	
			Modeled relative abundance	Modeled relative abundance	Spp_SP_abn_mrg_birds_FW_MRA	
		migratory wetlands	Binary presence/absence	Binary presence/absence	Spp_SP_dis_birds_FW_BinaryPresenceAbsence	
			Probability of occurrence	Probability of occurrence	Spp_SP_dis_birds_FW_ProbabilityOfOccurrence	
		amphibians	Binary presence/absence	Binary presence/absence	Spp_SP_dis_amph_FW_BinaryPresenceAbsence	
			Probability of occurrence	Probability of occurrence	Spp_SP_dis_amph_FW_ProbabilityOfOccurrence	
		Species populations	reptiles	Binary presence/absence	Binary presence/absence	Spp_SP_dis_rept_FW_BinaryPresenceAbsence
				Probability of occurrence	Probability of occurrence	Spp_SP_dis_rept_FW_ProbabilityOfOccurrence
Species distributions	mammals	Native listed in the Habitats Directive	Binary presence/absence	Binary presence/absence	Spp_SP_dis_mamm_FW_BinaryPresenceAbsence	
			Probability of occurrence	Probability of occurrence	Spp_SP_dis_mamm_FW_ProbabilityOfOccurrence	
		invertebrates	Binary presence/absence	Binary presence/absence	Spp_SP_dis_invertebrates_FW_BinaryPresenceAbsence	
			Probability of occurrence	Probability of occurrence	Spp_SP_dis_invertebrates_FW_ProbabilityOfOccurrence	
		macrophytes	Binary presence/absence	Binary presence/absence	Spp_SP_dis_macrophytes_FW_BinaryPresenceAbsence	
			Probability of occurrence	Probability of occurrence	Spp_SP_dis_macrophytes_FW_ProbabilityOfOccurrence	
		invasive alien	taxa of European concern	Binary presence/absence	Binary presence/absence	Spp_SP_dis_invasive_FW_BinaryPresenceAbsence
			Probability of occurrence	Probability of occurrence	Spp_SP_dis_invasive_FW_ProbabilityOfOccurrence	
		diapollites	native to Europe	Binary presence/absence	Binary presence/absence	Spp_SP_dis_diapollites_FW_BinaryPresenceAbsence
			Probability of occurrence	Probability of occurrence	Spp_SP_dis_diapollites_FW_ProbabilityOfOccurrence	
Species traits	Phenology	wetland birds	Day of arrival	All migratory wetland bird species	https://pub.org/vocab/abundantDA	Spp_ST_dis_birds_FW_DayOfArrival
			Day of departure	All migratory wetland bird species	https://pub.org/vocab/abundantDA	Spp_ST_dis_birds_FW_DayOfDeparture
		fishes	Day of arrival	All migratory freshwater fishes & All freshwater invertebrates	https://pub.org/vocab/abundantDA	Spp_ST_dis_fishes_FW_DayOfArrival
			Day of departure	All migratory freshwater fishes & All freshwater invertebrates	https://pub.org/vocab/abundantDA	Spp_ST_dis_fishes_FW_DayOfDeparture
phytoplankton	in lakes	Relative abundance	Relative abundance	numerical value betwe	Eco_CC_com_pHk_FW_RA	
	macrophytes	Ecological Quality Ratio (EQR)	Ecological Quality Ratio (EQR)	numerical value betwe	Eco_CC_com_pHk_FW_RA	

Figure 33: EBVs csv encoded following the I-ADOPT schema.

Finally, representing the same variable in the I-ADOPT schema, it can be expressed like this:



**Figure 34: I-ADOPT schema adapted by AD4GD expressing the Species populations | Species distributions EV in the case of Freshwater Mammals.**

EVs expressed under this new ontology schema are being translated to json and to RDF with the aim to upload them into the OGC RAINBOW. From there, the Essential Variable Product will have a code (for the moment only available from EuropaBON at the level of Entity) and a URI to which a dataset/in-situ data will uniquely refer to.

JSON files will also be used to describe EVs in the ENEON graph. With this, EVs will be linked to in-situ/citizen networks providing data in relation to these EVs.

**Community abundance** [Community abundance](#)

The abundance of organisms in ecological assemblages.

**Properties**

- *type*: EBV
- *supertype*: EV
- *theme*: [ Biodiversity | Biosphere | Ecosystems | Environment ]
- *GDThematicArea*: Biodiversity and ecosystems
- *SBA*: Biodiversity and Ecosystem Sustainability
- *EV\_class*: Community composition
- *product[0]*:
  - *Product Name*: Community abundance Freshwater phytoplankton
  - *Realm (Constraint 1)*: Freshwater
  - *Product Definition*: Composition of phytoplankton in the European catchments and rivers network system (ECRINS) as measured by the Ecological Quality Ratio (EQR) or based on total abundance (biovolume), taxonomic composition of indicator species, or bloom intensity (e.g. maximum biomass of cyanobacteria or percentage of cyanobacteria of the total biomass for all taxa)
  - *Observed Property Reference*: <https://github.com/EuropaBON/EBV-Descriptions/wiki/Freshwater-Community-composition-of-phytoplankton>
  - *Entity*: phytoplankton
  - *Constraint 2*: in lakes
  - *Constraint 3*:
  - *EBV Metric*: Relative abundance
  - *Spatial Resolution Unit*: <https://qudt.org/vocab/quantitykind/Area>
  - *Temporal Resolution Unit*: 1 year

[...]

**Links to**

(none)

**Backlinks to**

- *defines of*: [GEOBON](#) (GEOBON)
- *defines of*: [EuropaBON](#) (EuropaBON)

**Figure 35: Example of the Community Abundance EVs described following the I-ADOPT schema and encoded in json format for the ENEON graph**



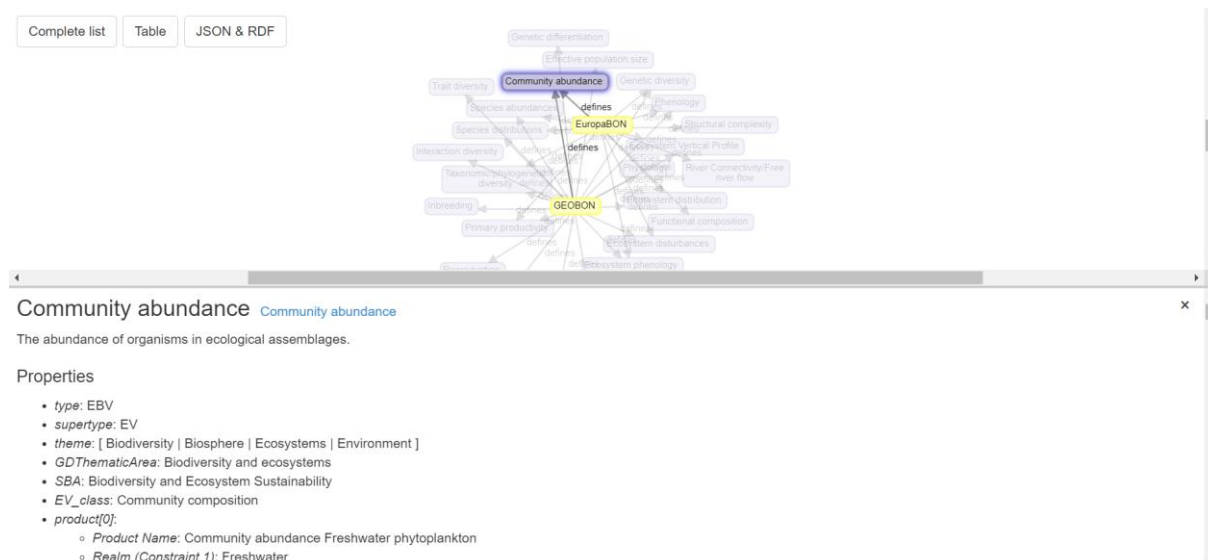


Figure 36: ENEON graph showing the EV of Community Abundance

After this work on EBVs, other EVs will follow, starting for the ones needed by other AD4GD pilots, such as the EWV and the ECV.

## 4.2 USING EV VOCABULARIES IN OGC STANDARDS

Only a few of the OGC standards and APIs provide a clean way to connect data with the described EV. One of them is the Sensor Thing API (STA). STA entity data model includes the concept of observedProperty (see Figure 37) that is associated to a set of observation values through a datastream. The observedProperty represents the variable or variable product definition and includes a pointer to a URI that defines the variable being measured.

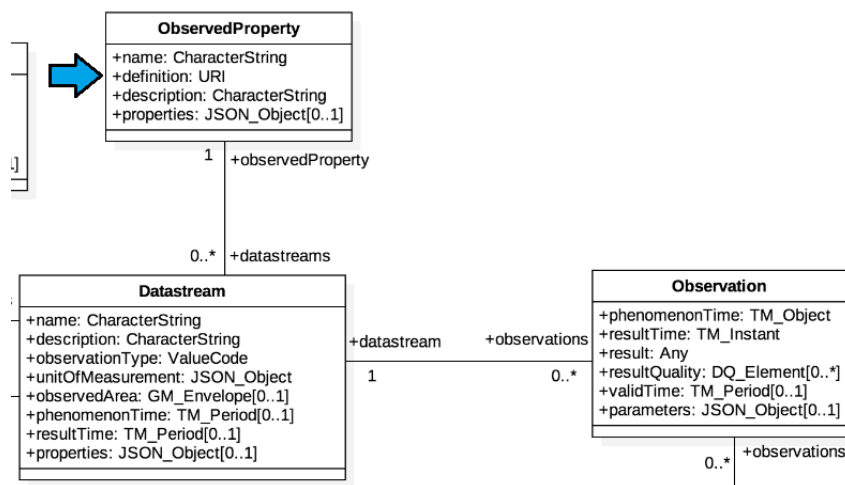


Figure 37: STA entity data model includes the concept of observedProperty.

In a STA, observations must have an observedProperty. We can enumerate all observed properties of a server by using the STA ODATA API<sup>22</sup>.

[https://api-samenmeten.rivm.nl/v1.0/observedProperties?\\$select=name.description](https://api-samenmeten.rivm.nl/v1.0/observedProperties?$select=name.description)

The response is as follows:

```
"value": [
```

<sup>22</sup> <https://developers.sensorup.com/docs/#queryparameters>

```

    {
      "name": "pres",
      "description": "atmosferische druk"
    },
    {
      "name": "rh",
      "description": "relatieve vochtigheid"
    },
    {
      "name": "temp",
      "description": "temperatuur"
    },
    {
      "name": "nh3",
      "description": "ammoniak"
    },
    {
      "name": "no2",
      "description": "stikstofdioxide"
    },
    {
      "name": "pm25_kal",
      "description": "fijnstof gekalibreerd < 2.5microm"
    },
    {
      "name": "pm10_kal",
      "description": "fijnstof gekalibreerd < 10microm"
    },
    {
      "name": "pm25",
      "description": "fijnstof < 2.5microm"
    },
    {
      "name": "pm10",
      "description": "fijnstof < 10microm"
    }
  ]
}

```

Now it is possible to formulate a query that extracts only one variable (e.g. 'temp') by using its name or its definition (full URI). For example, the following query requests only the observedProperty with name equal "temp" and the request expanding all observations (in the datastream) and selects only the unitOfMeasurement the result and the phenomenonTime.

[https://api-samenmeten.rivm.nl/v1.0/observedProperties?\\$filter=name%20eq%20%27temp%27&\\$select=name,description,Datastream&\\$expand=Datastream\(\\$expand=Datastream/Observations\(\\$select=result,phenomenonTime\);\\$select=unitOfMeasurement\)](https://api-samenmeten.rivm.nl/v1.0/observedProperties?$filter=name%20eq%20%27temp%27&$select=name,description,Datastream&$expand=Datastream($expand=Datastream/Observations($select=result,phenomenonTime);$select=unitOfMeasurement))

The response is as follows:

```

{
  "value": [
    {
      "name": "temp",
      "description": "temperatuur",

```

```

    "Datastreams": [
      {
        "unitOfMeasurement": {
          "definition":
"http://www.qudt.org/qudt/owl/1.0.0/unit/Instances.html",
          "symbol": "C"
        },
        "Observations": [
          {
            "phenomenonTime": "2023-04-14T14:00:00.000Z",
            "result": 21.76
          },
          {
            "phenomenonTime": "2023-04-14T13:00:00.000Z",
            "result": 16.4
          },
          {
            "phenomenonTime": "2023-04-14T12:00:00.000Z",
            "result": 14.41
          },
          {
            "phenomenonTime": "2023-04-14T11:00:00.000Z",
            "result": 13.8
          },
          {
            "phenomenonTime": "2023-04-14T10:00:00.000Z",
            "result": 11.7
          }
        ]
      }
    ]
  }...

```

If the Sensor Things API is used by several providers all using the same EV vocabularies it will be possible to request from all of them a single variable and create a composite dataset mixing all responses. Other OGC Web APIs lack this capacity of associating a property name with a URI defining it that is fundamental in an environmental data space. There is a need to extend the OGC Web APIs to support similar functionality.

## 5 RECOMMENDATIONS FOR IN-SITU DATA PROVIDERS

This section introduces what could be the steps that the in-situ data providers should take to be part of the GDDS, including the implementation of FAIR principles and their complementation with the other relevant principles (e.g., TRUST and GEO).

### 5.1 FAIR PRINCIPLES

In 2016, the 'FAIR Guiding Principles for scientific data management and stewardship' were published in Scientific Data (Wilkinson et al., 2016)<sup>23</sup>. The authors intended to provide guidelines to improve the Findability, Accessibility, Interoperability, and Reuse of digital assets. The principles emphasise machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with

<sup>23</sup> <https://www.nature.com/articles/sdata201618>

none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.<sup>24</sup>

The FAIR principles are well-known by the scientific community by the 4 words (Findable, Accessible, Interoperable, and Reusable) that give rise to its acronym, and which are defined as follows:

- **Findable** - The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services.
- **Accessible** - Once the user finds the required data, she/he/they need to know how they can be accessed, possibly including authentication and authorisation.
- **Interoperable** - The data usually needs to be integrated with other data. In addition, the data needs to interoperate with applications or workflows for analysis, storage, and processing.
- **Reusable** - The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

However, the whole approach of FAIR consists of 10 principles, some of them with sub-principles, with a total of 15 criteria. These are the FAIR principles as defined by Go-FAIR. The detailed version of FAIR is presented in Table 4 below.

Principles Name	Principles ID	Principles Description
Findable	F1	(Meta) data are assigned globally unique and persistent identifiers
	F2	Data are described with rich metadata
	F3	Metadata clearly and explicitly include the identifier of the data they describe
	F4	(Meta)data are registered or indexed in a searchable resource
Accessible	A1	(Meta)data are retrievable by their identifier using a standardised communication protocol
	A1.1	The protocol is open, free and universally implementable
	A1.2	The protocol allows for an authentication and authorisation procedure where necessary
	A2	Metadata should be accessible even when the data is no longer available
Interoperable	I1	(Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
	I2	(Meta)data use vocabularies that follow the FAIR principles
	I3	(Meta)data include qualified references to other (meta)data
Reusable	R1	(Meta)data are richly described with a plurality of accurate and relevant attributes
	R1.1	(Meta)data are released with a clear and accessible data usage licence
	R1.2	(Meta)data are associated with detailed provenance
	R1.3	(Meta)data meet domain-relevant community standards

**Table 4: Full version of the FAIR principles: Findable (F1, F2, F3, F4), Accessible (A1, A1.1, A1.2, A2), Interoperable (I1, I2, I3), and Reusable (R1, R1.1, R1.2, R1.3).**

<sup>24</sup> <https://www.go-fair.org/fair-principles/>

## 5.2 RELATION OF FAIR WITH THE GEO DMP

In 2015, the 'GEO Strategic Plan 2016-2025: Implementing GEOSS' was released containing the Data Management Principles (DMPs), that, if followed, should maximize the value and benefits of the data shared in GEOSS<sup>25</sup>. Consisting of 10 key principles, the DMPs cover all phases of the live cycle of the geospatial data management and are classified in five headings: discoverability, accessibility, usability, preservation, and curation. In summary (Table 5), the GEO DMP principles say that Earth observations should be catalogued or otherwise advertised on the internet so that they can be discovered (DMP-1), and should be accessible online using open-standard encodings and services (DMP-2). Data and services should be comprehensively documented using international or community-approved standards, and to the extent possible, peer-reviewed publications, so that users can understand and make use of the data (DMP-3). Metadata should include access and use conditions (DMP-4), the results of quality control procedures (DMP-6), and provenance statements indicating the origin and processing history of the dataset or product (DMP-5). Data and associated metadata should be protected from loss (DMP-7), and periodically verified to ensure integrity, authenticity and readability (DMP-8). Corrections and updates to data and metadata records should be performed as required (DMP-9). Finally, persistent identifiers should be assigned to data so that they can be tracked and cited and data providers can be acknowledged (DMP-10)<sup>26</sup>.

GEO DMP headings	GEO DMP number	GEO DMP name	GEO DMP description
Discoverability	DMP-1	Discovery	Data and all associated metadata will be discoverable, through catalogues and search engines, and data access and use conditions, including licenses, will be clearly indicated.
Accessibility	DMP-2	Access	Data will be accessible via online services, including, at a minimum, direct download but preferably user-customizable services for access, visualization and analysis.
Usability	DMP-3	Encoding	Data should be structured using encodings that are widely accepted in the target user community and aligned with organizational needs and observing methods, with preference given to non-proprietary international standards.
	DMP-4	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 possible extent, data will be described in peer-reviewed publications referenced in metadata records.
	DMP-5	Provenance	Data will include provenance metadata indicating the origin and processing history of raw observations and derived products, to ensure full traceability of the product chain.
	DMP-6	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.

<sup>25</sup> [https://www.d-geo.de/docs/GEO\\_Strategic\\_Plan\\_2016\\_2025\\_Implementing\\_GEOSS\\_Reference\\_Document.pdf](https://www.d-geo.de/docs/GEO_Strategic_Plan_2016_2025_Implementing_GEOSS_Reference_Document.pdf)

<sup>26</sup> [https://www.earthobservations.org/documents/dswg/201504\\_data\\_management\\_principles\\_condensed\\_final.pdf](https://www.earthobservations.org/documents/dswg/201504_data_management_principles_condensed_final.pdf)

Preservation	DMP-7	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.
	DMP-8	Verification	Data and associated metadata held in data management systems will be periodically verified to ensure integrity, authenticity and readability.
Curation	DMP-9	Review and Processing	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.
	DMP-10	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.

**Table 5: Summary table of the current text of the GEO DMP.**

In parallel to the GEO DMP, in 2016, the ‘FAIR Guiding Principles for scientific data management and stewardship’ were published (Section 5.1). Instead of focusing on data management, the FAIR principles emphasize data actionability in a machine to machine environment. There is an overlap with the FAIR principles that will be addressed by the GEO DMP subgroup in the near future. The FAIR principles have become popular in the scientific community and have been embraced as part of the open science movement, while the GEO DMPs became limited to the GEO community. This section proposes some changes in the DMPs definitions to allow for a better alignment with the FAIR principles. The main aim is to later define a profile of the FAIR principles for GEO that reuses the modified version of the GEO DMPs.

The main problem with the FAIR principles is that the 4 words that the acronym includes can be interpreted in an intuitive way that often leads to an interpretation that is different from its original intend. To avoid this, is it necessary to go deep into the sub-principles as defined by the "three-point FAIRification Framework"<sup>27</sup>.

### 5.2.1 FINDABLE

This principle focuses on how to make data findable by machines and humans. Machine-readable metadata are essential for automatic discovery of datasets and services.

*F1. (Meta)data are assigned a globally unique and persistent identifier*

This FAIR principle is almost identical to the GEO DMP-10: "Data will be assigned appropriate unique, persistent, resolvable identifiers to enable documents to cite the data on which they are based and to enable data providers to receive acknowledgement of use of their data."

Small details differ in the definitions: In FAIR, metadata also receives an identifier and the GEO DMP-10 the identifier should be resolvable (what seems necessary for machine-to-machine findability).

<sup>27</sup> <https://www.go-fair.org/fair-principles/>

**Recommendation 1:** Include in DMP-10 that "Metadata will be assigned appropriate unique, persistent, resolvable identifiers".

*F2. Data are described with rich metadata (defined by R1 below)*

This is equivalent to DMP-4 "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 ."

The GEO-DMPs include the possibility to describe the data in peer-reviewed publications. It is normal that the FAIR principles do not mention this possibility as this is not useful for machine-to-machine interactions.

*F3. Metadata clearly and explicitly include the identifier of the data they describe*

There is no mention about metadata including the persistent identifier (PID). Since this is good practice in the geospatial world, we have to assume this is implicit in the DMP-4 and DMP-10 but it is not explicitly stated.

**Recommendation 2:** Modify DMP-10 to explicitly request that the PID is included in the metadata.

*F4. (Meta)data are registered or indexed in a searchable resource*

This is exactly what DMP-1 says: "Data and all associated metadata will be discoverable through catalogues and search engines, ..."

## 5.2.2 ACCESSIBLE

This principle is requesting a mechanism to know how data can be accessed.

*A1. (Meta)data are retrievable by their identifier using a standardised communications protocol*

While DMP-2 deals with accessibility, there is one detail that is not requested in DMPs: the capacity to retrieve the dataset via de PID; instead DMP-2 proposes the use of web services. It is worth noting that the FAIR sub-principle is considering the retrieval of the whole dataset while the DMP-2 contemplates the possibility to "preferably user-customizable services for visualization and computation". It is unclear what "customizable services" means, but one possible interpretation is the capacity to extract a subset of the data via a query. Geospatial resources have the tendency of being too big for a full download. The two final words are also unclear to me: "visualization and computation" seems to suggest remote processing (e.g. in the cloud) but it is not clearly stated.

**Recommendation 3:** In DMP-2, include the need of "resolvable" PID (i.e. the capacity to access the data using the PID). This is not current common practice in the geospatial world.

**Recommendation 4:** Add to DMP-2 the use of services or Web APIs (to consider the migration of the OGC web services to APIs) (While this is not suggested by the FAIR-DMP comparison it is a necessary update)

**Recommendation 5:** Add to DMP-2 the word "remotely" in front of "analysis" to emphasize that for big data moving code close to the data might be more efficient than a simple download.

*A1.1 The protocol is open, free, and universally implementable*



There is no reference to "open standards" in DMP-2. Since the Data Sharing (DS) and the DMPs were designed with the open data and open science in mind, it could be good to include the reference to it. In DMP-3 there is a reference to "non-proprietary international standards" but it talks about data formats and not about protocols.

**Recommendation 6:** In DMP-2 include a reference to "open and universally implementable protocol" after mention "web service and APIs".

**Recommendation 7:** In DMP-3 replace "non-proprietary international standards." By "open and universally implementable standards".

*A1.2 The protocol allows for an authentication and authorisation procedure, where necessary*

There is no reference to this in the DMPs and this might be an omission done in a time period when GEO was pushing for open data. On the other hand we have to recognize that the EC is now pushing for systems that contemplate digital economy requirements such as the data spaces. GEO is also trying to attract the private sector and, in addition, the GEO portal allows for authentication. In the spirit of open data, it is unclear if GEO needs to adhere with this sub-principle.

**Recommendation 8:** Add a reference to possible authentication and authorisation procedure to DMP2.

*A2. Metadata are accessible, even when the data are no longer available*

This recognizes that we cannot preserve all existing data forever and compromises need to be made. In the DMP-7 there is a reference that suggests retiring data and moving it to data archives. It says "planning will be for the long term ... disposal or transfer procedures. "

**Recommendation 9:** Add " metadata will remain accessible, even when data have been disposed and transferred to an archive" to the end of DMP-7.

### 5.2.3 INTEROPERABLE

In the FAIR principles, interoperability is used in a restricted meaning, referring to the capacity to integrate data with other data.

*II. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.*

This principle talks about "a knowledge representation model". It is talking about a data model that can be annotated with semantic concepts such as RDF annotated with Dublin core, OWL or JSON-LD. Nothing in the DMPs refers to knowledge representation models. Instead the DMPs discuss data formats (almost nothing in the FAIR principles about data formats; that is only suggested in R1.3) without any reference to data models.

**Recommendation 10:** Rephrase the DMP-3 by adding references to data models. "Data will be structured using encodings and **knowledge representation models (i.e. data models or data schemas)** that are widely accepted in the target user community and aligned with organizational needs and observing methods, with preference given to non-proprietary international standards.

*I2. (Meta)data use vocabularies that follow FAIR principles*

Data and metadata should be annotated using common vocabularies. This idea is not very visible in the DMPs. DMP-4 seems to suggest that, when it talks about elements for "understanding" the data. Vocabularies could be considered complementary to metadata so it could be added to DMP-4.

**Recommendation 11:** Include a reference to vocabularies in DMP-4 in the following way: "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. (Meta)Data should be annotated with references to external and accepted vocabularies preferable in machine readable formats. To the extent possible, data will also be described in peer reviewed publications referenced in the metadata record.

### *I3. (Meta)data include qualified references to other (meta)data*

This requires that a dataset links to another dataset and specifies the reason why this link is done. DMP-5 indirectly allows for one particular reason for that, that is mentioned in the description of FAIR I3 too: A derived dataset can be originated by processing or analysing other dataset. In ISO metadata this is recorded in the lineage (a.k.a. provenance).

**Recommendation 12:** Extend DMP-5 to other kinds of links between datasets by adding: "Datasets will also include other relations to other datasets when appropriate, such as being a part, being derived from, a new version of or a different resolution of another dataset."

## 5.2.4 REUSABLE

Reuse of data means replicability but also the capacity to be combined in different settings.

### *R1. (Meta)data are richly described with a plurality of accurate and relevant attributes*

This is the exact intention of DMP-4 that asks for comprehensive metadata for use, understanding and processing the data.

#### *R1.1. (Meta)data are released with a clear and accessible data usage license*

DMP-1 asks for metadata that includes "use conditions, including licenses". The inclusion of licenses in the discoverability DMP-1 is confusing, as licenses do not influence discoverability but the later use and reuse. These needs to be fixed by moving the license reference to another DMP.

**Recommendation 13:** Move the reference to "use conditions, including licenses" from DMP-1 to DMP-4.

#### *R1.2. (Meta)data are associated with detailed provenance*

The DMP-5 recommends precisely that.

#### *R1.3. (Meta)data meet domain-relevant community standards*

DMP-3 mentions "data encodings" the "target user community" with the same intention that this FAIR sub-principle does.

## 5.2.5 ABOUT THE OTHER DMPS

Note that, in go-fair, the detailed explanation of the R1.3 sub-principle specifically states: "Note that **quality** issues are not addressed by the FAIR principles. The data's reliability lies in the eye of the beholder and

depends on the intended application."<sup>28</sup>. This means that **DMP-6** cannot be mapped with FAIR. Other DMPs that cannot be mapped to any of the FAIR sub-principles are: **DMP-8** and **DMP-9** referring to data verification and data updates respectively.

If the 13 recommendations in this document are applied to the DMPs text, the similarities between FAIR and some DMPs increase and it is possible to create a better FAIR profile for GEO recognizing that FAIR is not only about data management. This will make it acceptable that 3 DMPs cannot be mapped to FAIR and eventually suggests that the profile name needs to reflect this. A proposal is to call them **FAIR and Management Principles** (FAIRM Principles: Findable, Accessible, Interoperable, Reusable and Managed). We should also consider **resorting** to the DMP to better align with the FAIR order. DMP-10 should be the first and we should consider DMP-6 being after DMP-7 to group together the 3 DMPs that cannot be included in FAIR.

**Recommendation 14:** Move DMP-10 to the first position to align it with the FAIR principles Also consider DMP10 as Discoverability.

**Recommendation 15:** Switch DMP-6 and DMP-7 to group the 3 principles that go beyond the scope of FAIR.

The correspondence between the FAIR principles, the GEO DMP, and the recommendations described above, is illustrated in Figure 38 and summarised in Table 6.

---

<sup>28</sup> <https://www.go-fair.org/fair-principles/r1-3-metadata-meet-domain-relevant-community-standards/>

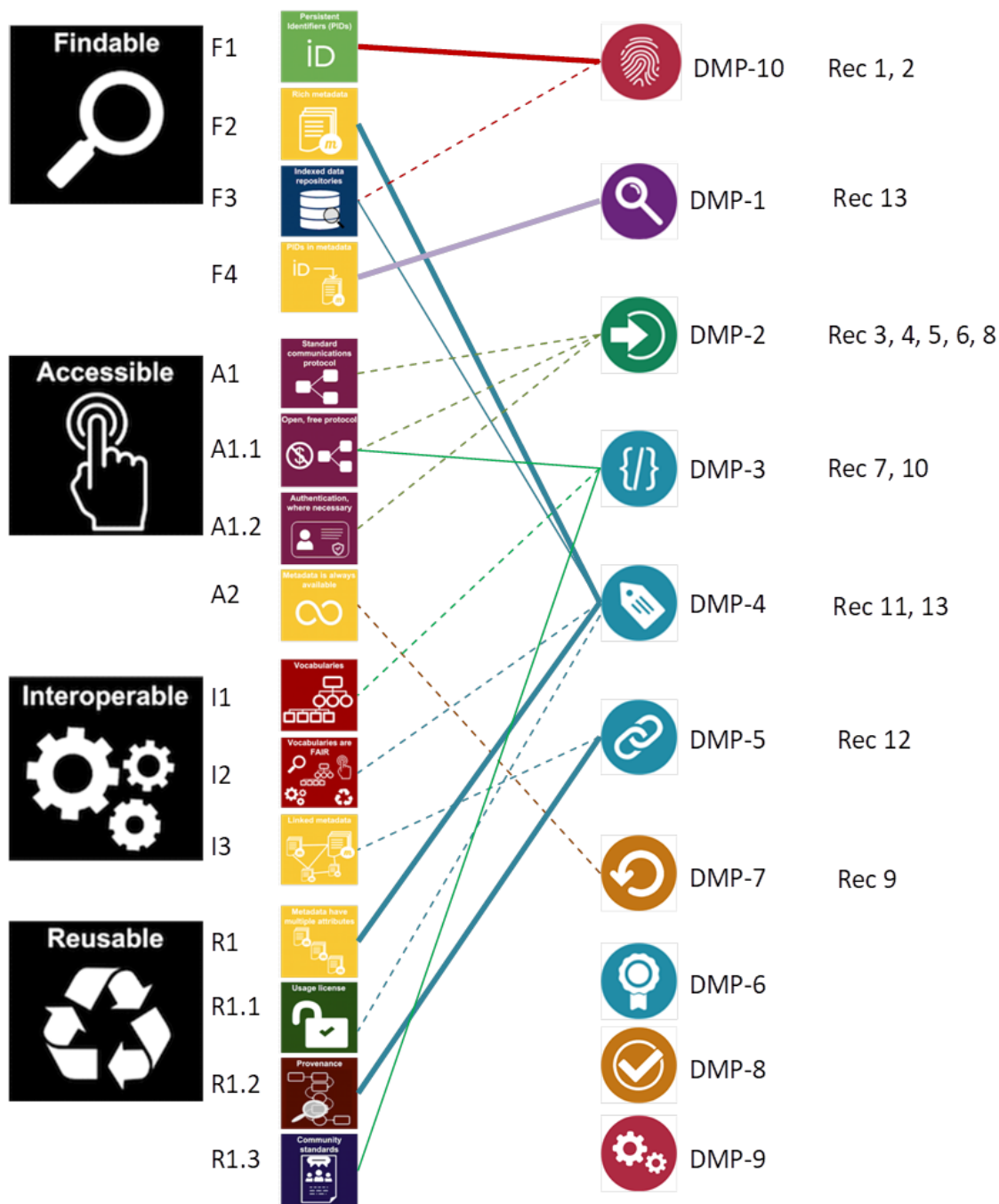


Figure 38: Correspondence between the FAIR principles and the GEO DMP.

FAIR	GEO DMP Number	GEO DMP Name	Description
Findable	DMP-10	Identifiers	Data and metadata will be assigned appropriate persistent, unique and resolvable identifiers to enable documents to cite the data and the metadata on which they are based and to enable data providers to receive acknowledgement for use of their data. Metadata records will contain the PID of the data they describe.

	DMP-1	Discovery	Data and all associated metadata will be discoverable, through catalogues and search engines.
Accessible	DMP-2	Access	Data will be accessible via online services or web APIs using open and universally implementable protocol that will support authentication and authorisation procedure where necessary, including, at a minimum, direct download but preferably user-customizable services for access, visualization or remote analysis.. Data will also be accessible with its resolvable permanent identifier.
	DMP-3	Encoding	Data should be structured using encodings and knowledge representation models (i.e. data models or data schemas) that are widely accepted in the target user community and aligned with organizational needs and observing methods, with preference given to open and universally implementable standards.
	DMP-4	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. (Meta)Data should be annotated with references to external and accepted vocabularies preferable in machine readable formats. To the possible extent, data will be described in peer-reviewed publications referenced in metadata records. Data access and use conditions, including licenses, will be clearly indicated.
Interoperable	DMP-5	Provenance	Data will include provenance metadata indicating the origin and processing history of raw observations and derived products, to ensure full traceability of the product chain. Datasets will also include other relations to other datasets when appropriate, such as being a part, being derived from, a new version of or a different resolution of another dataset.
	DMP-7	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. Metadata will remain accessible, even when data have been disposed and transferred to an archive.
Reuse and other	DMP-6	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.
	DMP-8	Verification	Data and associated metadata held in data management systems will be periodically verified to ensure integrity, authenticity and readability.
	DMP-9	Review and Processing	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.

**Table 6: Summary table of the proposed revision of the GEO DMPs.**

### 5.3 IMPLEMENTING FAIR PRINCIPLES

In order to facilitate and promote the implementation of the FAIR principles, the European Commission expert group on FAIR data published the report “Turning FAIR into reality”<sup>29</sup>. This report analyses what is needed to implement FAIR and provides an Action Plan with concrete recommendations and actions for stakeholders for advancing in Open Science.

A total of twenty-seven recommendations are drawn, grouped into Priority Recommendations and Supporting Recommendations. Fifteen Priority Recommendations are made. These relate to the key concepts of FAIR Digital Objects, which are then implemented through interoperability frameworks. The whole set of recommendations is presented as follows.

#### ■ PRIORITY RECOMMENDATIONS

Step 1: Define – concepts for FAIR Digital Objects and the ecosystem

- Rec. 1: Define FAIR for implementation
- Rec. 2: Implement a model for FAIR Digital Objects
- Rec. 3: Develop components of a FAIR ecosystem

Step 2: Implement – culture, technology and skills for FAIR practice

- Rec. 4: Develop interoperability frameworks for FAIR sharing within disciplines and for interdisciplinary research
- Rec. 5: Ensure Data Management via Data Management Plans
- Rec. 6: Recognise and reward FAIR data and data stewardship
- Rec. 7: Support semantic technologies
- Rec. 8: Facilitate automated processing
- Rec. 9: Develop assessment frameworks to certify FAIR services
- Rec. 10: Professionalise data science and data stewardship roles and train researchers
- Rec. 11: Implement curriculum frameworks and training

Step 3: Embed and sustain – incentives, metrics and investment

- Rec. 12: Develop metrics for FAIR Digital Objects
- Rec. 13: Develop metrics to certify FAIR services
- Rec. 14: Provide strategic and coordinated funding
- Rec. 15: Provide sustainable funding

#### ■ SUPPORTING RECOMMENDATIONS

- Rec. 16: Apply FAIR broadly
- Rec. 17: Align and harmonise FAIR and Open data policy
- Rec. 18: Cost data management
- Rec. 19: Select and prioritise FAIR Digital Objects
- Rec. 20: Deposit in Trusted Digital Repositories
- Rec. 21: Encourage and incentivise reuse of FAIR outputs
- Rec. 22: Use information held in Data Management Plans
- Rec. 23: Develop FAIR components to meet research needs
- Rec. 24: Incentivise research infrastructures and other services to support FAIR data
- Rec. 25: Implement FAIR metrics to monitor uptake
- Rec. 26: Support data citation and next generation metrics
- Rec. 27: Open EOSC to all providers but ensure services are FAIR

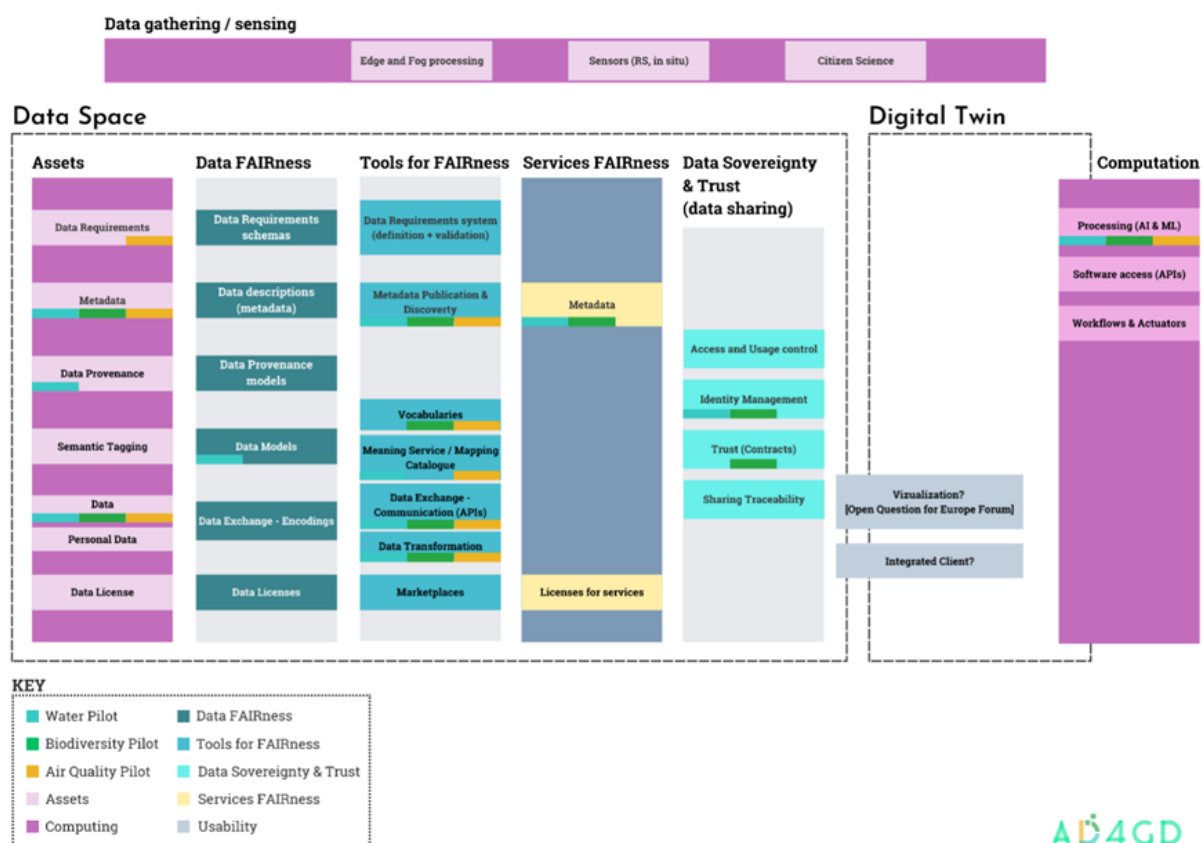
<sup>29</sup> <https://op.europa.eu/en/publication-detail/-/publication/7769a148-f1f6-11e8-9982-01aa75ed71a1/language-en/format-PDF/source-80611283>

## 6 STANDARDS AND COMPONENTS FOR FAIR DATA IN GDDS

AD4GD has adopted the Data Spaces Support Centre (DSSC)<sup>30</sup> building blocks to establish the elements required to support the implementation of the GDDS. This adoption resulted in an evolution of the original DSSC building blocks based on what was first proposed by the USAGE sister project and published within the USAGE D3.2 at: <https://drive.google.com/file/d/1FyvNkWkAWkuKKHh-3l529VUu5LIKy5E/view>.

The extension of the DSSC building blocks evolved by mapping them to the current OGC standards and web APIs tested in AD4GD according to the needs of the pilot case studies, the FAIR principles, the GEO DMP, and the European Interoperability Framework principles and recommendations.

As a result of this analytic process, the building blocks to support the implementation of the AD4GD proposal of the GDDS were agreed by consensus in the Torino AD4GD GA in February 2024, where the consortium participated, and externals, i.e. Andreas Matheus, in the AD4GD Advisory Board and Lukas Mocek from Sensor.Community, who was invited to the meeting. The last version of the building blocks is presented in Figure 39 - directly extracted from AD4GD Deliverable D6.1. Building blocks are structured on Assets, Data, Tools and Services for data FAIR, and Data Sovereignty and Trust aspects for secure data sharing. Other data environments such as the digital twins for data processing are reflected, as it is expected they will coexist and will be provisioned from data in the data space.



**Figure 39: Building blocks to support the implementation of the AD4GD proposal of the GDDS, including a mapping of the planned pilots' components. Credits: AD4GD D6.1 (L: AU)**

In alignment with these building blocks, the AD4GD results so far have been structured in a total of **12 practical components**. These components focus on the integration and reuse of existing and tailored tools

<sup>30</sup> <https://dssc.eu/>



and workflows to support the development of three pilot case studies (on Water, Biodiversity and Air Quality topics), emphasising the specific challenges in terms of data and metadata consumption, use and production that each pilot demonstrates. In other words, the components set the context for the pilots and position them within the AD4GD proposal of the GDDS.

The 12 components are listed and briefly described below (the numbering comes from AD4GD D6.1):

- **Component 1 - Automated Ingestion of Data from Diverse Low-cost Sensors.** Integration of several data sources in a single one to be applied in the Air Quality Pilot.
- **Component 2 – Evaluation of Connector Solutions.** Investigation of IDSA connectors APIs and OGC APIs in place.
- **Component 3 – Semantic uplift: SPARQL/JSON-LD For Data Exchange.** A Green Deal Information Model is developed as a common vocabulary to provide the basis of a common green deal data space, enabling interoperability and integration of different systems, potentially from different vendors.
- **Component 4 – Mobilising Data to STApplus.** TAPIS, "Tables from OGC APIs for Sensors" as a client application developed in HTML and JavaScript to support data mobilisation in STApplus.
- **Component 5 – Mobilising sensor data with STA.** Routines to import observations to STA (FROST implementation) and transformation of observations to a standard service.
- **Component 6 – Data Cubes for Consuming, Publishing and Processing Multidimensional Data.** A habitat connectivity open data cube as a service to access the data through Jupyter notebooks.
- **Component 7 – Open workflows for habitat connectivity computation.** Open graph-oriented workflows for habitat connectivity computation applied to the Biodiversity Pilot.
- **Component 8 – Water modelling and prediction.** Water modelling and prediction that models the status and evolution of water quality and quantity on Berlin lakes applied to the Water Pilot.
- **Component 9 – Data catalogue and metadata.** A GeoNetwork data and metadata catalogue for all data consumed and produced in the context of three pilots.
- **Component 10 – Data catalogue and metadata with Semantic Uplift.** A Data catalogue with Semantic Uplift using GeoDCAT (and linked to GeoNetwork).
- **Component 11 – Data Trustworthiness Framework.** A Data Trustworthiness Framework to validate data and express its quality with QualityMl and SensorML is developed and applied to all 3 pilots.
- **Component 12 – GDDS Semantic Terms - RAINBOW Server.** The OGC RAINBOW as a Web accessible source of information about things ("Concepts") the OGC defines or that communities ask the OGC to host on their behalf. It applies FAIR principles.

It is out of the scope of this document to explain each component in detail. Instead, the aim is to situate each component according to the FAIR principles it contributes to, with the focus on in-situ observations. It is recommended that the reader refers to AD4GD Deliverable D6.1 (<https://zenodo.org/records/10839023>) for a detailed technical description of each component.

As mentioned, the building blocks research involved mapping them with common standards and the FAIR principles. Figure 40 shows to which FAIR principle each building block contributes to, and the practical components (C+number) associated.

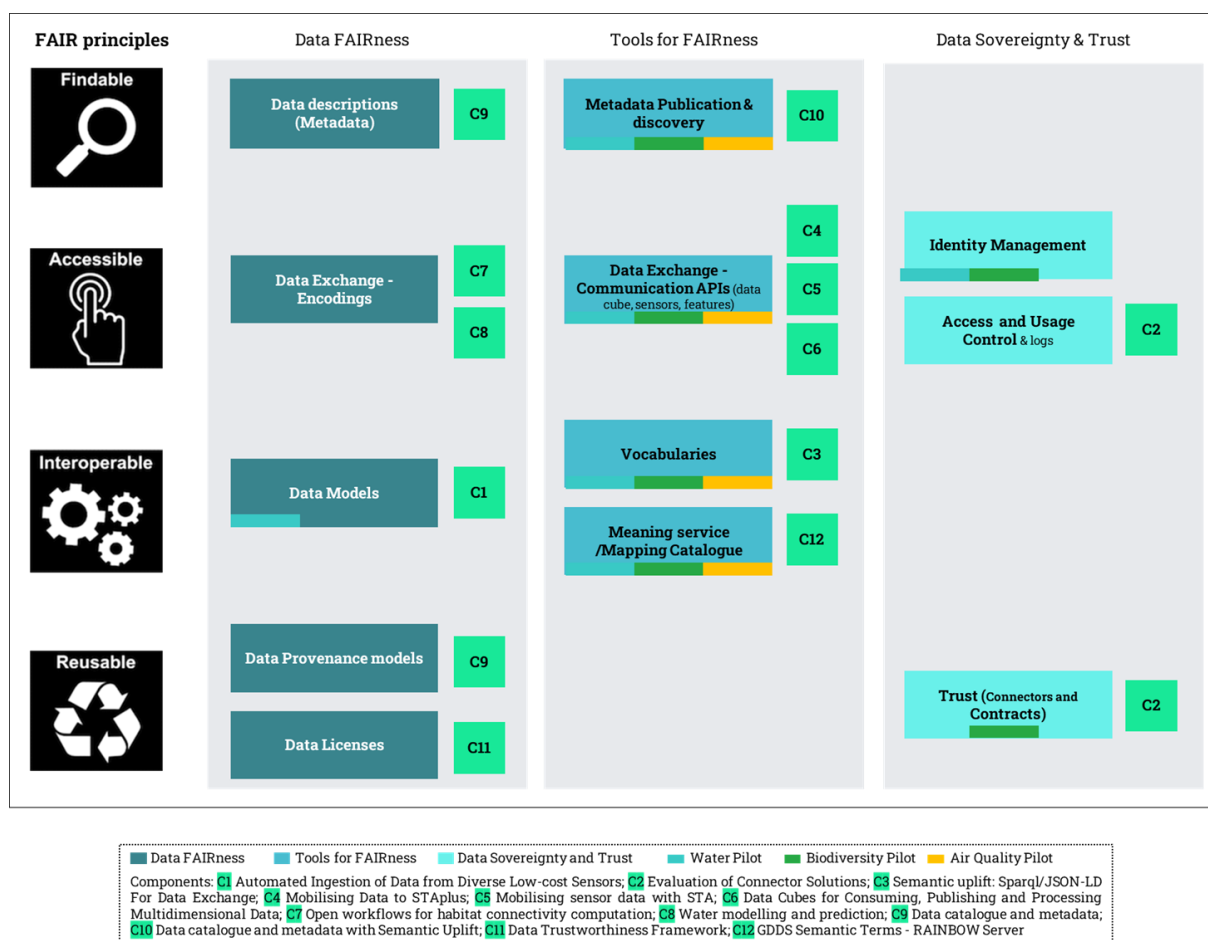


Figure 40: AD4GD Building blocks and Components in contribution to FAIR.

The standards identified for FAIR in-situ data linked to the building blocks are presented in the following sections. These sections aim to provide guidance for implementing the building blocks and components that facilitate the application of these standards. While the primary focus of this document is on in-situ observations, many of the standards and components discussed can also be applied to gridded or raster data, as you will see below.

## 6.1 STANDARDS AND COMPONENTS FOR FINDING IN-SITU DATA IN THE GDDS

While most people rely on internet browsers such as Google to find textual information, finding data is more complex. Since most datasets are just sequences of numbers, it is hard to express their meaning through the natural text and keywords currently used by common browsers and they cannot be indexed without metadata. Metadata is composed of textual descriptions of the data. Formal standards for geospatial environmental data metadata are Dublin Core and ISO 19115 and DCAT. Individual metadata records are stored in metadata catalogues that are indexed databases that allow for querying the metadata and retrieving entries that match with the requested characteristics such as topic, spatial extent, temporal extent, resolution and data quality indicators. Common metadata catalogues are STAC, OGC CSW and CKAN (Comprehensive Knowledge Archive Network), GeoNode or GeoNetwork. Another way to improve findability is to associate each and every dataset with a resolvable and permanent identifier (PID). A reference of the dataset PID in a report or in scientific literature will make the dataset immediately findable. One of the main difficulties in geospatial information is that it is structured in long temporal and spatial dataset series sometimes resulting in flooding the results with a lot of similar results. Another problem in geospatial information is the almost absence of criteria to prioritise or rank the results. The use of Geospatial

User Feedback (GUF) can help in ranking the results and providing additional information to users based on previous experiences using the same dataset.

It is expected that the GDDS will have one or more nodes that will be dedicated to data discovery and finding information.

#### **AD4GD Components contributing to data Findability**

- **Component 9 – Data catalogue and metadata**

Data consumed and produced in ADG4D pilots are being catalogued in GeoNetwork according to the ISO 19115 standard using predefined metadata templates. GeoNetwork is also designed to be compliant with the INSPIRE Directive specifications and data encoding rules for open data. GeoNetwork generates globally unique identifiers (IDs) and allows for referencing and linking any specific record, within the catalogue and with external systems. The aim is to ensure that, as a first step, data and metadata are **findable**, while linking data catalogued with standard vocabularies as part of the semantic enrichment effort. The AD4GD GeoNetwork catalogue can be accessed at: <https://catalogue.grumets.cat/geonetwork/ad4gd/>

- **Component 10 – Data catalogue and metadata with Semantic Uplift**

Data and metadata catalogued in GeoNetwork are being semantically enriched using Data Catalogue Vocabulary (DCAT)/GeoDCAT extension for ontologies of spatially referenced data. Experiments have been performed to convert ISO 19115 datastreams into DCAT/GeoDCAT. The goal is a functional solution that allows to retrieve metadata from GeoNetwork (and potentially other ISO 191\*\* compatible catalogues), convert it to DCAT/GeoDCAT, and publish it so that it can be queried by clients. The semantic enrichment of the data increases the success of geoportals and search engines in **finding** the required data.

## **6.2 STANDARD AND COMPONENTS FOR IN-SITU DATA ACCESS IN THE GDDS**

One of the main problems of geospatial information is the variety of formats and data models of the geospatial information. However, there are 3 main data types that result in a number of standards:

- Data can be a continuous coverage represented by a raster file, a data cube, a point cloud or a TIN data structure. In this case, the OGC Web Coverage Service and the OGC API Coverages are two standards for data access of a subset of the coverage data.
- Data can also be a collection of sparse points, lines or polygons representing geospatial features on the ground. In this case, the OGC Web Feature Service and the OGC API Features are two standards for data access of a subset of the feature data collections.
- Data can also be sets of observations done by sensors over the terrain. The OGC Sensor Observation Service and the OGC Sensor Things API are two standards for data access of a subset of the observations and measurements

There are also standards that focus on geospatial queries in a way that are agnostic of the data type such as the OGC Environmental Data Retrieval.

The separation of the metadata in catalogues and the data in access services has one main issue. Sometimes the data provider discontinues the product or moves it to another URL without notifying the catalogue about the change. A user search finds hits that will never give access to the disconnected data. In an environment such as the GDDS, this will not only generate frustration but also lack of trust on the data space.

Most of these services are maintained by the data providers themselves so each provider will have a node in the GDDS. The citizen science data could be an exception to this, as discussed in previous sections.

### **AD4GD Components contributing to data Accessibility**

- **Component 2 – Evaluation of Connector Solutions**

Connectors are software components that automatically extract data from one or more data sources and land that data in a third-party database making the use of contracts and pre-established rules for **accessing** the data. In AD4GD we evaluated two data space connector frameworks: the Eclipse Dataspace Components (EDC) and the International Data Spaces Association (IDSA) implementation. Given the limited software development resources available in AD4GD, EDC and IDSA connectors should be considered still in an early stage for its robust implementation. The most practical approach to implement the functionalities that connectors should perform is through an extension of an OGC service that supports authentication (identity verification of users or services) and authorization (access rights) based on a number of rules which serve as a preliminary step to the signature of the contracts. A possible candidate is the STAplus extension of the OGC STA that provides specific solutions for the privacy management of citizen science observations.

- **Component 4 – Mobilising Data to STAplus**

This component has involved the development of the TAPIS (Tables from APIs for Sensors) tool, available at: <https://www.tapis.grumets.cat/>. The open-source code can be accessed from GitHub at: <https://github.com/joanma747/TAPIS>. TAPIS is a client application developed in HTML and Javascript to support data mobilisation to STAplus in a GUI that allows semantic annotation of data tables and facilitates the mapping from tables to the OGC STA data model. Using TAPIS, CSV tables can be semantically enriched by generating CSVW files. TAPIS maintains the provenance and lineage information of the data. Beyond the scope of mobilising data, the goal is to discover sensor-data and make it **accessible**. STAplus data model is part of the OGC family of standards, extending the aforementioned OGC STA, both of them specifically developed for sensor-based in-situ observations.

- **Component 5 – Mobilising sensor data with STA**

The IoT lab partner infrastructure has been used to perform initial testing to transform water quality and water level data in CSV format to STAplus format. The component includes the FROST server which is compliant to the STAplus data model using a dedicated plugin. Harmonised sample data can be **accessed** at: <https://frost.iotlab.com/sensorthings/v1.1/Things>.

- **Component 6 – Data Cubes for Consuming, Publishing and Processing Multidimensional Data**

Open Data Cube technology is being developed on various platforms as a service to dynamically access multi-dimensional data. One of the most common ways to use data cubes is through interactively writing Python code within a Jupyter Notebook, as experimented in AD4GD the project, notably by the Biodiversity Pilot. Data cubes were in origin conceptualised for handling data generated by satellites and made it **accessible** from different data repositories. In AD4GD, together with FAIR-i-CUBE and B3 sister projects, the goal is to integrate species occurrence (in-situ data) in a data cube in order they can be combined with satellite data and derived products (gridded data) in the same multi-dimensional environment.

- **Component 7 – Open workflows for habitat connectivity computation**

Open graph-oriented semi-automatic workflows for habitat connectivity computation

- **Component 8 – Water modelling and prediction**

As part of Water Pilot, past archived water quality and water availability measurements coming from IoT or citizen science are being used to build and evaluate Machine Learning pipelines for predicting water trends on Berlin lakes. The aim is to improve the results of the predictions to potentially set a forecast system based on past water measurements with an emphasis on making this data easily **accessible** by reducing the complexity of the new ML models. The AD4GD project prioritises solutions for sharing information through interoperable public platforms or databases useful for the pilots' stakeholders.

### 6.3 STANDARD FOR IN-SITU DATA (SEMANTIC) INTEROPERABILITY IN THE GDDS

Traditionally geospatial information data models and formats cared about the description of the geographic properties and assumed that the thematic component of the information was somehow solved. One of the main examples of this is the GML. The long document describing GML 2.0 defines how to encode geographical features in XML objects. To do so, it defines XML schemas (XSD) for geographic objects (e.g. points, lines, polygons). To define your geospatial features you are supposed to define a class in an XSD with thematic properties and geographical properties. While in XSD you can define a property type you cannot add units of measure or attach a variable description (e.g. an EV product name or URI concept definition). The syntax of XSD is too restrictive. A similar situation happens with a Shapefile where the thematic properties are described in DBF tables only by up to 10 character column name. Metadata describing the data model could fix this but it does not. For example, ISO19110 Feature Catalogue (see Figure 41) incorporates valueMeasurementUnit (to add a unit of measure) and valueType (if it is Integer or String, etc).

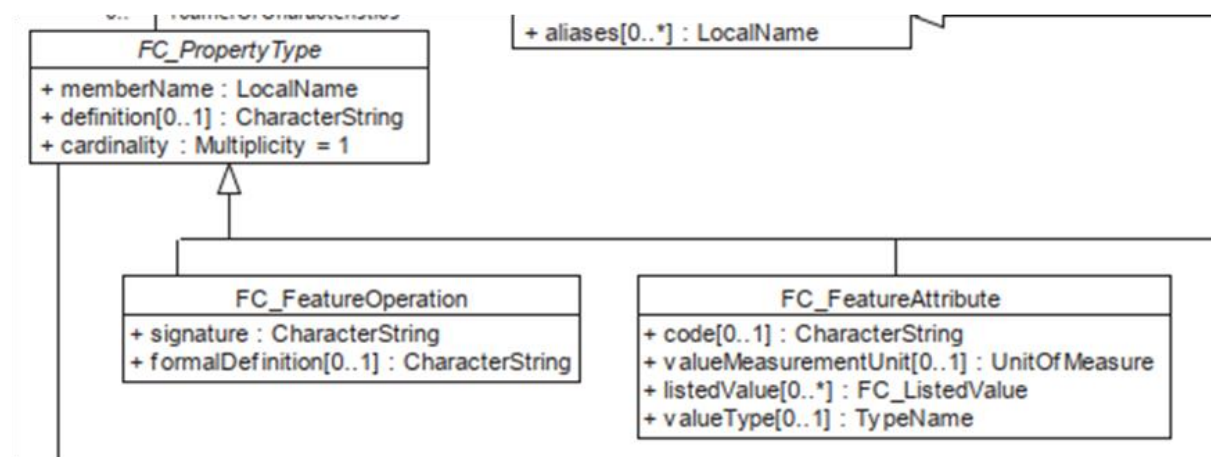


Figure 41: Detailed view of the ISO19110 Feature Catalogue model.

Using the FC\_PropertyType and extending "definition" to include a URI as well as a name (as explained by ISO19139) could be done (see Figure 42). Unfortunately, the ISO19110 is rarely used and this level of detail is difficult to see in Feature services.

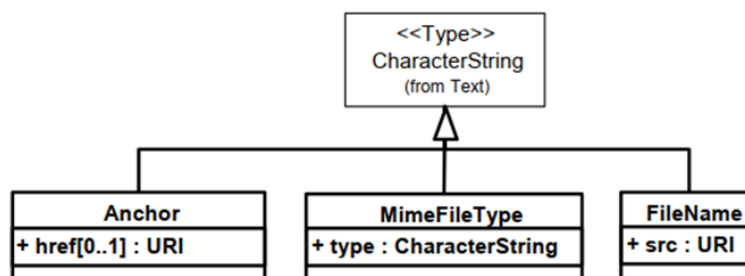


Figure 42: How to extend CharacterString to include a URI.

The ISO 19115-1 fixes this only for coverages incorporating a "name" in MD\_RangeDimension (that is a MD\_identifier that e.g. allows pointing to a vocabulary of variables) and units in the MD\_SampleDimension

(Figure 43). Unfortunately, coverage services do not usually provide coverage data accompanied by ISO 19115-1 records and if they do, it rarely includes this level of detail.

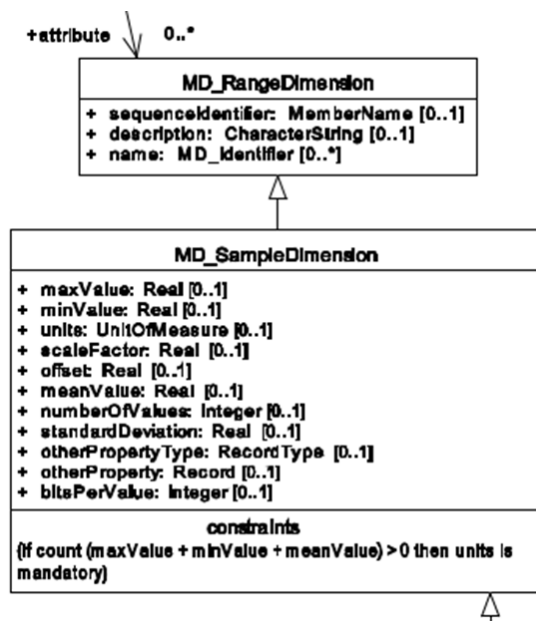


Figure 43: Description of the Range types included as values in a coverage following the ISO 19115-1.

As we have said before, the Sensor Things API is almost the only data access standard that has a mandatory observedProperty defined by a URI.

But having data access services that are capable of linking to external vocabularies is not going to fix the problem if there is no common vocabulary to link with. That is why the OGC RAINBOW (formerly known as OGC definition service) is proposed as a vocabulary repository for observedProperties. In our opinion it should contain a list of EV and its products and other extended vocabularies of variables that we could reuse and point from the data that is being included in the data space.

Another component of interoperability and trust is the need for describing the provenance and the data quality of the data. These aspects are reasonably covered by the ISO 19115-1 and the ISO19157 standards respectively.

**AD4GD Components contributing to data Interoperability**

- **Component 1 - Automated Ingestion of Data from Diverse Low-cost Sensors**

This component relies on well-documented findable data to succeed in the integration of data from several sensor platforms. The prototype proposed by AD4GD automatically ingests data from different Non-commercial air quality sensor platforms (Plume Flow (<https://plumelabs.com/en/flow/>), EarthSense Zephyr (<https://www.earthsense.co.uk/zephyr>), Sensor.Community open source AirRohr kit (<https://sensor.community/en/sensors/airrohr/>) which comes in different data formats, ranging from web scraping, csv downloads, or via web APIs. Beyond the solution for accessing and retrieving sensor data that this component allows, it has demonstrated to be very useful in facilitating **interoperability** experiments with semantic mapping and republishing too.

- **Component 3 – Semantic uplift: Sparql/JSON-LD For Data Exchange**

Based on previous experiences, the Green Deal Information Model (GDIM) have been defined to establish the basis of a common node of the GDDS that enables semantic interoperability, available at <https://github.com/AD4GD/GDIM>. SPARQL is the standard query language and protocol for Linked Open Data



on the web or for RDF triplestore. Original data in CSV format is transformed to RDF. RDF models can be implemented in different formats, including JSON-LD (JavaScript Object Notation for Linked Data). The data remains in the original source (e.g., relational database, JSON/CSV from REST API) and we generate semantic data dynamically as an output. The SPARQL mappings from CSV format to the GDIM can be found on GitHub at: <https://github.com/AD4GD/GDIM/tree/main/AD4GD-pilots-mappings/sensor-community>. The GDIM approach involves both reusing existing components and ontologies, developing new code and introducing new domain-specific vocabularies to enable semantic **interoperability**, such as the Essential Variables framework, which apply to all types of EO data including in-situ observations.

- **Component 12 – GDDS Semantic Terms - RAINBOW Server**

To host AD4GD's GDIM (Component 3), the project requires an easily accessible and extensible semantic resource capable of supporting semantic interoperability with other repositories. To make this possible, the project has chosen the OGC RAINBOW to define ontologies from the Essential Variables framework for organizing EO-environmental data by themes and domains. The aim is transforming controlled vocabularies into RDF format and aligning terms in different vocabularies to avoid duplication of information. The AD4GD RAINBOW environment can be accessed at: <http://defs-dev.opengis.net/vocprez-hosted/>. Recently in the project, new experiments are being done with I-ADOPT Framework Ontology in the same effort of increase interoperability between existing variable description models (see Section 4.1). Standards and components for in-situ data Reusability in the GDDS

The main goal is to optimise the reuse of data in a secure environment with the use of licences. To achieve this, as stated by the FAIR principles, metadata must be richly described with a plurality of accurate and relevant attributes (ontologies), released with a clear and accessible data usage licence (linked with contracts), associated with detailed provenance (data quality) and meet domain-relevant community standards to interoperate. All the standards mentioned in the previous subsections contribute to the reuse of data.

#### **AD4GD Components contributing to data Reusability**

- **Component 11 – Data Trustworthiness Framework**

The Data Trustworthiness Framework to validate data and express its quality with QualityML and SensorML is developed and applied to all 3 pilots to enable data **reuse**. The quality assessments and usability guidance of this prototype will be made available in an open and standardised manner via a semantically enriched API, thus ensuring that producers and users of IoT and CitSci can replicate the quality assurance procedure followed. This approach has been initially tested in the context of the data consumed and produced in the context of Air Quality Pilot and it will be extended to the Water and Biodiversity pilots as well.

**Component 9 – Data catalogue and metadata** for data provenance (quality) and **Component 2 – Evaluation of Connector Solutions** for data trust are also considered as main contributions to the data reusability in the AD4GD proposal of the GDDS.

## **6.4 AUTHENTICATION AND PRIVACY ASPECTS**

While most of the data in the Green Deal data space should remain open, it could be good to have controlled access to sensitive data and private data when needed. Standards for SSO such as OpenID Connect can be used. The Authenix service developed and tested in the Cos4Cloud project is a possible solution (<https://marketplace.eosc-portal.eu/services/authenix>). Another issue in the GDDS is the need for trust in the data and the services. Certificates and HTTPS can be used to ensure that services are genuine and have not been tampered with. We should adopt solutions coming from IT mainstream



## 7 CONCLUSIONS

The ENEON graph produced in the ConnectinGEO project is still a valid tool for presenting and analysing the status of the in-situ networks in Europe. To make it useful for the GDDS, EU Green Deal thematic areas needed to be applied. The evolution of the Essential Variables (and in particular the EBV done by EuropaBON), the emergency of other sets of global targets beyond the SDGs (such as the Kunming-Montreal Global Biodiversity Framework - GBF - 2030 targets), the inclusion of other Essential Variables (such as the EWV and the air quality variables), the emergence of new observation networks and the consideration citizen science initiatives as in-situ data contributors imposes a continuous requirements for updating the graph to keep it relevant for the analysis of the current status for the in-situ variables.

The EVs are too general to be used as a vocabulary to characterise the available dataset provided by the in-situ networks. The essential variable products concept introduced by GCOS for the ECV is the right way to go. In the EBV, the I-ADOPT ontology has been used to define the EBV products defined by EuropaBON. This specialisation will help the ENEON graph to better integrate the EBV. The definition created will be stored in the OGC Rainbow.

The FAIR principles should inspire the architecture of the GDDS as well as the DMP in GEO. In this deliverable we demonstrate that with minimum changes the GEO DMP can be mapped to the FAIR principles. In that respect, 15 concrete recommendations are formulated.

In addition the AD4GD three pilots have been designed in WP6. This deliverable demonstrates that the pilots use components of the AD4GD architecture that contribute to make the pilot data FAIR. Indeed, the pilot needs are covered by a set of 12 components that once combined together form the AD4GD architecture.

## REFERENCES

- [1] Giuliani, G., Egger, E., Italiano, J., Poussin, C., Richard, J. P., & Chatenoux, B. (2020). Essential variables for environmental monitoring: what are the possible contributions of earth observation data cubes?. *Data*, 5(4), 100.
- [2] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1), 1-9.