



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: D4.1
TECHNICAL DEVELOPMENTS REQUIRED FOR THE IMPLEMENTATION OF GREEN DEAL DATA SPACE

Due date of deliverable: 31/05/2024

Actual submission date: 20/07/2024

Version: 2.0

Main Authors: Thomas Hodson (ECMWF), Franco Di Pietro, Carlos Cob Parro, Ignacio EliceGUI (ATOS), Cédric Crettaz (IoT Lab), Adrian Quesada Rodriguez, Sébastien Ziegler (MI), Ivette Serral, Joan Maso (CREAF)



DOCUMENT METADATA

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

Deliverable title	Technical Developments required for the implementation of Green Deal Data Space
Deliverable number	D4.1
Deliverable version	1
Contractual date of delivery	31/05/2024
Actual date of delivery	20/07/2024
Document status	Ready for submission
Document version	2.0
Online access	
Dissemination	Public
Work package	WP4: Satellite and Green Deal Data Space Integration
Partner responsible	ATOS-IT
Author(s)	Thomas Hodson (ECMWF), Franco Di Pietro, Carlos Cob Parro, Ignacio Elicegui (ATOS), Cédric Crettaz (IoT Lab), Adrian Quesada Rodriguez, Sébastien Ziegler (MI), Ivette Serral, Joan Maso (CREAF).
Editor(s)	Franco Di Pietro, Ignacio Elicegui (ATOS)
Reviewer(s)	Francesca Noardo (OGC)
EC Project Officer	Lara Congiu



Abstract	This deliverable reports the work achieved during the first phase of the AD4GD project concerning the WP4 on the technical developments required to bring new data sources into the Green Data Space. It presents a description of the existing earth observation data, then the incorporated internet of things, low-cost sensors and multidimensional data are described as well as the work performed in this project to do so. A description on how this data is used in the context of each pilot is provided.
Keywords	Data sources, IoT, multidimensional data, low-cost sensors, Data Spaces, Green Deal, Data Cube, Data Integration, Homogenisation, Data curation.
Disclaimer	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

Version history			
Version	Date	Modification reason	Modified by
0.1	26/03/2024	Initial version of the document	Franco Di Pietro (Atos)
0.2	05/04/2024	Inputs from IoT Lab	Cédric Crettaz (IoT Lab)
0.3	26/04/2024	Inputs from Berlin pilots	Malte Zamzow (KWB), Cédric Crettaz (IoT Lab)
0.4	10/06/2024	Inputs from CREAM	Ivette Serral (CREAF), Joan Maso (CREAF)
0.5	27/06/2024	Socio-economic data contribution	Cédric Crettaz (IoT Lab)
0.6	28/06/2024	Contribution for the socio-economic data	Mandat International
1.0	03/07/2024	Compilation of missing contributions	Ignacio Elicegui (Atos)
1.1	11/07/2024	Introduction and Conclusions sections	Ignacio Elicegui/Franco Di Pietro (Atos)
2.0	18/07/2024	Internal review	Francesca Noardo (OGC)



ABBREVIATIONS

Abbreviation	Definition
AD4GD	All Data for Green Deal
ACS	Atmospheric Data Store
ADSAPI	Ads Application Program Interface
AI	Artificial Intelligence
API	Application Programming Interface
C3S	Copernicus Climate Change Service
CAMS	Copernicus Atmosphere Monitoring Service
CDS	Climate Data Store
CDSAPI	Climate Data Store (CDS) Application Program Interface (API)
CESSDA	Consortium of European Social Science Data Archives
CIS	Coverage Implementation Schema
CLMS	Copernicus Land Monitoring Service
CMEMS	Copernicus Marine Environment Monitoring Service
CSV	Comma-Separated Values
DC	Data Catalogue
EC	European Commission
ECB	European Central Bank
ECMWF	European Centre for Medium-Range Weather Forecasts
EFTA	European Free Trade Association
EO	Earth Observation
ERA5	ECMWF fifth reanalysis



ESA	European Space Agency
EU	European Union
EUMETSAT	European Organisation for the Exploitation of Meteorological Satellites
FADN	Farm Accountancy Data Network
FAPAR	Fraction of Absorbed Photosynthetically Active Radiation
FDB	Fields Database
FROST	FRaunhofer Opensource SensorThings
GBIF	Global Biodiversity Information Facility
GD	Green Deal
GDDS	Green Deal Data Space
GeoJSON,	Geospatial JavaScript Object Notation
GeoTIFF,	Geographic Tagged Image File Format
GUI	Graphical User Interface
HTTP	HyperText Transfer Protocol
IEA	International Energy Agency
IMF	International Monetary Fund
INSPIRE	Infrastructure for Spatial Information in Europe
IoT	Internet of Things
IPCC DDC	Intergovernmental Panel on Climate Change Data Distribution Centre
IPR	Intellectual Property Rights
JSON	JavaScript Object Notation
LULC	Land-Use/Land-Cover
ML	Machine Learning
MQTT	Message Queuing Telemetry Transport



NetCDF	Network Common Data Form
NO2	Nitrogen dioxide
NUTS	Nomenclature of Territorial Units for Statistics
ODATA	Open Data Protocol
ODC	Open Data Cube
OECD	Organization for Economic Co-operation and Development
OGC	Open Geospatial Consortium
pH	potential of Hydrogen
PM	Particulate Matter
RESTful	REpresentational State Transfer
SITS	Strategic Information Technology Services
STA	SensorThings API
STAC	SpatioTemporal Asset Catalogs
STApplus	SensorThings API plus
UNECE	United Nations Economic Commission for Europe
URL	Uniform Resource Locator
WCS	Web Coverage Service
WDPA	World Database on Protected Areas
WFS	Web Feature Service
WMS	Web Map Service
WPx	Work Package number x
WRI	World Resources Institute



TABLE OF CONTENTS

1	Introduction	11
2	Data for the Green Deal Data Space	13
2.1	Remote Sensing	13
2.2	INSPIRE Data Annexes	13
2.3	Socio-economic Data relevance for GD, applicability, etc.	14
2.3.1	Identifying and Selecting Relevant Data Sources	15
2.3.2	Addressing Heterogeneous Data Sources	16
2.3.3	Next Steps for Socio-economic Data Integration	17
3	IoT & Low-cost Sensors	21
3.1	Data Sources in pilots	21
3.1.1	Pilot 1: Water Availability/Quality	21
3.1.2	Pilot 2: Biodiversity	22
3.1.3	Pilot 3: Air quality	22
3.2	Technical developments: solutions proposed	23
3.2.1	STA and STApplus (Components 4&5)	23
3.2.2	IonBeam (Component 1)	24
3.3	Future work	26
4	Multidimensional Data Cubes (component 6)	27
4.1	Data sources in pilots	27
4.1.1	Pilot 1: Water Availability/Quality	28
4.1.2	Pilot 2: Biodiversity	28
4.1.3	Pilot 3: Air quality	29
4.2	Technical developments: solutions proposed	30
4.2.1	Kriging Interpolation Algorithm	30
4.2.2	Open Data Cube Extensions	31
4.2.3	Rasdaman Data Cube	33
4.3	Future work	36
5	Conclusions and recommendations	38
6	References	39



TABLE OF FIGURES

Figure 1:	AD4GD.....	11
Figure 2:	Population retrieved from the Eurostat service.....	18
Figure 3:	Nitrogen value retrieved from Eurostat	18
Figure 4:	IoT sensors installed in the lakes in Berlin	19
Figure 5:	Average of water temperature in Berlin.....	19
Figure 6:	Air temperature in Berlin during one year.....	20
Figure 7:	A high level block diagram showing how IonBeam fits into ECMWFs existing infrastructure...	24
Figure 8:	The current architectural implementation of IonBeam in Kubernetes, using RabbitMQ as a message queue.....	25
Figure 9:	Different implementations of data cubes (extracted from https://www.ogc.org/initiatives/gdc).	27
Figure 10:	Description of CAMS European air quality dataset on the ADS web portal including scientific documentation and metadata.....	29
Figure 11:	Monthly mean PM2.5 concentration for December 2022 across the Netherlands with background determined from CAMS Europe regional model analysis and circles indicating the PM2.5 concentration at a given IoT station.....	30
Figure 12:	Distribution of available IoT stations (left) and gridded PM2.5 concentration derived from those stations (right) in Sofia for December 2022. Color shading indicates monthly mean PM2.5 concentration. The regridding was performed on a 250x250m2 resolution.....	31
Figure 13:	Resulting arrays from a query on a specific BBOX and time range over the whole S2 pool of COG images.	33
Figure 14:	Visualisation of the NIR band for a certain date in the array, and for the whole array of date values.	33
Figure 15:	Example of WCPS query in FAIRiCUBE to visualise dynamics in Index of Terrestrial Connectivity (1987-2022) based on the bounding box, covering central Catalonia.....	34
Figure 16:	Gain and loss in Index of Terrestrial Connectivity (1987-2022).	35
Figure 17:	The fragments of FAIRiCUBE platform to access Index of Terrestrial Connectivity timeseries...	36



EXECUTIVE SUMMARY

This document (Deliverable 4.1 for the AD4GD project) presents the initial results achieved in the context of Tasks 4.1 “Inclusive Green Deal Data Space(s) design, including socio-economic data” and 4.2 “Green Deal Data Space Implementation” as part of Work Package 4 “Satellite and Green Deal Data Space Integration”. Current text is aligned with the previous work carried out by WP6 and the technical requirements identified to perform the AD4GD pilots’ main targets and so, follows the components’ architecture defined by D6.1.

This way, results are presented in two parts directly derived from the mentioned tasks: i) first part shows an analysis of a set of current available data sources (external and internal to AD4GD) with corresponding integration approaches; and ii) the technical components required to execute the integration for a subset of these data sources.

Specifically, this first round of achievements includes a list of available socio-economic dataset considered relevant as part of the Green Deal Data Space (GDDS). Continuing with the integration stage, the document includes an overview of the technical developments needed to incorporate new data sources into the GDDS, particularly for the cases of low cost/IoT sensor data and multidimensional data based in data cubes technology. In this line, a per-pilot description is provided with the aim of clearly presenting the relevance of these data sources in the context of the project concrete applications, as well as a clear description of the identified problems in each case together with the approach to target them in the future.

The continuation of this work with further results will be presented on the next deliverable (D4.2).

1 INTRODUCTION

WP4 is devoted to support the development of the three AD4GD pilots from the point of view of integrating, managing and serving the required datasets for their performance. This way, WP4 contributes to the general Green Deal Data Space(s) by supporting (and offering) a harmonised AD4GD common data space, in close collaboration with WP1, WP2 and WP3.

For this purpose, WP4 first executes an analysis of existing data sources to identify and provide an initial catalogue of infrastructures and/or assets considered potentially relevant to the objectives of the pilots (and of the project). This catalogue is composed of raw sources provided by the AD4GD pilots and additional open datasets offered by external but verified sources.

In parallel, WP6 has executed an analysis of the different requirements needed by the AD4GD designed pilots to be performed, in terms of expected outcomes, user interfaces, data pipelines, integration and semantics. This analysis resulted in a set of twelve identified building blocks, mapped as technical components (Figure 1:), required to build and deploy the three sets of project pilots.

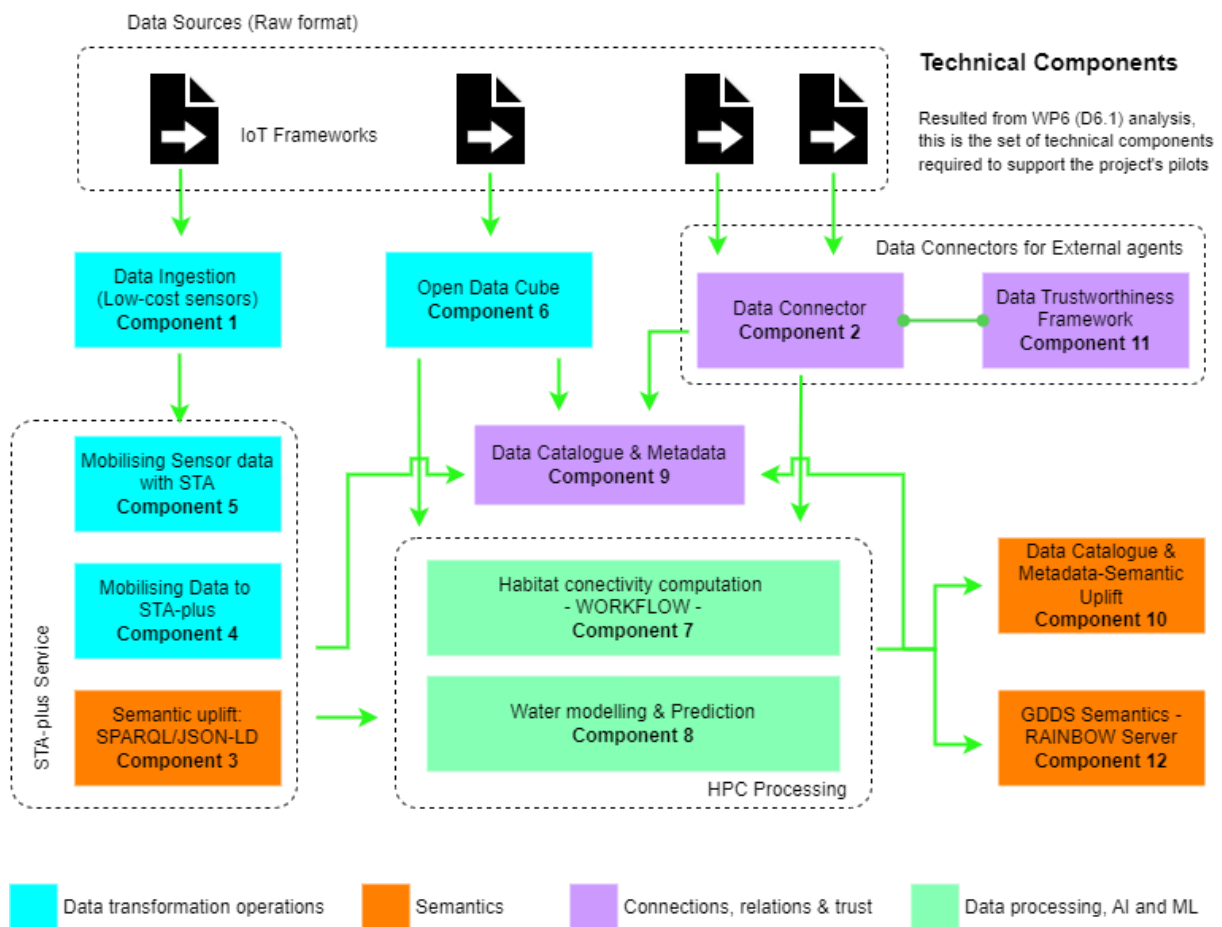


Figure 1: AD4GD Technical components (derived from D6.1 [1]).

According to their main operations, these technical components were classified in four main functional blocks, which are addressed by the different WPs:

- **Data transformation** components connect with the original (and existing) raw data sources and perform the corresponding operations (data curation, data cleansing, etc.) to enable homogeneous data sharing and exploitation within the scope of the pilot and in line with AD4GD objectives. The included components are supported by **WP4** and **WP3** in a close collaboration action.



- **Semantic** components reinforce the data homogeneity by adapting the meaning, format and structure of the captured data, supporting the Green Deal Data Space data model. **WP1** provides this set of functionalities.
- **Data processing, AI and ML** developments represent specific AD4GD work undertaken to extend and enrich reference datasets by combining data sources and applying AI and ML technologies. These are addressed in **WP5**.
- **Connection, relations and trust** blocks are intended to expose and share the datasets resulting from the components of the previous functional blocks. This contributes in a trustworthy manner to the Green Deal Data Spaces and it is addressed by **WP4**.

The orchestration of all these functional building blocks, through their corresponding technical components, collects from the selected data sources, transforms, combines and exploits these datasets to support the pilot's performance and, finally, produces valuable new information to expand the GDSS.

Presented deliverable focuses on the first round of achievements within WP4, related to the schema of Figure 1: . This way, the document is divided into two main parts, as follows:

- Part 1, composed by **Section 2** (*Data for the Green Deal Data Space*), presents the results achieved by Task 4.1 and the analysis carried out with the identified existing data sources, mainly IoT infrastructures, and socio-economic available information. This also covers the integration of these data sources within the project's data space, which directly connects with the next parts of the document.
- Part 2, directly involving the Data transformation technical components shown in Figure 1: , and reflecting part of the work carried out by Task 4.2 linked with WP3. This is shown in **Section 3** (*IoT & Low-cost sensors*) and **Section 4** (*Multidimensional Data Cubes*) which describe the STA and STApplus framework (components 1, 4 and 5 of Figure 1:) and the Open Data Cube (component 6) respectively.

Both the data sources analysis and the technical components development have been done from the point of view of the AD4GD proposed pilots, and it is thus reflected in the corresponding sections, showing the covered data sources by each technical component and describing how these are transformed to be later exploited and shared.

Finally, the deliverable presents a set of conclusions derived from the results obtained in Task 4.1 and 4.2. These are linked with the current work in progress within the rest of WPs and their future exploitation lines.



2 DATA FOR THE GREEN DEAL DATA SPACE

Before describing the work done on dataspace components as part of this deliverable, we will first briefly review the landscape of data that can be contributed to the Green Deal Data Space, with special emphasis on the socioeconomic data that was the focus of the Task 4.1 in the project.

This first review of data sources is the first step in the implementation of AD4GD pilots, as it provides the raw material necessary to achieve the pilot's objectives: each pilot can be developed considering the available (and usable) datasets.

This section analyses external relevant sources, such as the Remote Sensing information provided by the **Copernicus** system and the Spatial data offered by the **INSPIRE** [1] infrastructure.

2.1 REMOTE SENSING

The Copernicus Data Space Ecosystem is the official gateway to satellite imagery from the Copernicus Programme, providing free instant access to the largest collection of Earth observation data and services in the world. Key features of the Copernicus Data Space Ecosystem include the largest EO data offering, with the full archive and newly collected images from Sentinel-1, Sentinel-2, Sentinel-3, and Sentinel-5P satellites, as well as complementary datasets like the Copernicus digital elevation model. In addition, it incorporates a powerful data analytics environment with high-quality data processing tools to extract valuable information for public, private or commercial activities. The Copernicus Data Space Ecosystem was launched in early 2023 by a consortium of leading European cloud and Earth observation service providers, with guidance from ESA.

Besides the low-level products available in the Copernicus Data Space Ecosystem¹, the Copernicus services cover six main thematic areas: Atmosphere (CAMS), Marine (CMEMS), Land (CLMS), Climate Change (C3S), Security and Emergencies. They are managed by the European Commission (EC) in partnership with organisations including the ESA, EUMETSAT, ECMWF and others.

Both, the Copernicus Data Space Ecosystem and the Copernicus Services should be incorporated in the GDDS. In AD4GD we work on integrating the Copernicus Atmosphere Monitoring Service (CAMS) and the Copernicus Climate Change Service (C3S) into the GDDS. The CAMS provides information related to air pollution, solar energy, greenhouse gases, and climate forcing (defined as the imbalance in the Earth's energy budget caused by changes in the climate system). The C3S provides information about the past, present, and future climate to support adaptation and mitigation policies. ECMWF implements CAMS and the C3S on behalf of the EC. ECMWF maintains the Atmospheric Data Store (ACS) and the Climate Data Store (CDS). Each of these services provides a web catalogue and data access methods for multiple disparate data sources including essential climate variables, forecast model output and EO. They implement many of the functions that are required by the GDDS such as discovery, authentication, contracts negotiation, translation and data access. These functions together are prototypical examples of federated dataspace components. AD4GD gets inspiration from the CAMS and C3S and is proposing an architecture that is compatible with them.

2.2 INSPIRE DATA ANNEXES

The INSPIRE (Infrastructure for Spatial Information in Europe) directive is a European Union (EU) framework for spatial data and services that supports the creation of European policy relating to the environment. It aims to create a harmonised spatial data infrastructure across EU member states to facilitate the sharing of open access environmental spatial information among public sector organisations.

The INSPIRE infrastructure integrates existing spatial data infrastructures data and services provided by the member states and demands the harmonisation of existing data rather than the collection of new data.

¹ <https://dataspace.copernicus.eu/ecosystem/services>



The directive establishes that data should be collected only once and maintained where it can be most effective. It includes the concept of interoperability defined as the ability to combine seamless spatial information from different sources across Europe and share it with many users and applications. It implements mechanisms to make data discoverable and usable by means of cataloguing metadata about the data. It defines services such as discovery, viewing, downloading, and transformation. The INSPIRE directive would bring several benefits, that facilitates the work of policymaking across boundaries by providing access to spatial information in a transparent way. However, there are also challenges and limitations: Complex data models for harmonisation add an extra burden, requiring technical experts and the right tools. In addition, budget constraints limited the ability of Member States to implement and maintain the infrastructure so far.

The INSPIRE directive requires public authorities in the EU to publish certain datasets related to the environment according to specific standards. The INSPIRE directive defines 34 spatial data themes that cover a wide range of geospatial data related to the environment, infrastructure, and human activities. The EU are required to publish data for this themes that are:

- Addresses
- Administrative units
- Cadastral parcels
- Coordinate reference systems
- Geographical grid systems
- Agricultural and aquaculture facilities
- Area management/restriction/regulation zones & reporting units
- Atmospheric conditions
- Bio-geographical regions
- Buildings
- Energy Resources
- Environmental monitoring facilities
- Habitats and biotopes
- Human health and safety
- Land use
- Meteorological geographical features
- Mineral Resources
- Natural risk zones
- Oceanographic geographical features
- Population distribution and demography
- Production and industrial facilities
- Sea regions
- Soil
- Species distribution
- Statistical units
- Utility and governmental services

The INSPIRE Geoportal and Registry provide access to these datasets published by public authorities across Europe. Common ways to publish INSPIRE data include using web services for visualisation (WMS) and downloading (WFS/Atom) the data. The INSPIRE Geoportal allows sharing of geographically referenced thematic information between organisations. The INSPIRE Registry lists priority datasets related to environmental reporting that should be made available through the European Spatial Data Infrastructure.

2.3 SOCIO-ECONOMIC DATA RELEVANCE FOR GD, APPLICABILITY, ETC.

A comprehensive mapping of relevant socio-economic datasets and resources available for integration in the GDDS has been done, which includes technical considerations associated with APIs, the availability of

data, the accessibility limitations, the privacy and the intellectual property rights (IPR). This section presents the results of the analysis related to the socio-economic data in the context of the AD4GD project and the upcoming GDDS. Data from GBIF is used in pilot 2.

2.3.1 Identifying and Selecting Relevant Data Sources

Task 4.1 performed a review of available online data sources containing environmental and socio-economic data of relevance for AD4GD. After analysing a large variety of existing data sources, a selection has been made, by considering:

1. Relevance of the datasets for AD4GD;
2. Availability of the Open Access datasets without financial restrictions or binding licences;
3. Reliability of the data sources.

After review, the following data sources were considered as priority data sources for the GDDS:

- Eurostat
- EU Open Data Portal
- OECD Data
- OECD Data Explorer
- World Bank Open Data "THE WORLD BANK Data Catalog"
- EnergyData.info
- "ECB Eurosystem (ECB Data Portal)"
- United Nations Economic Commission for Europe (UNECE) Data Portal
- "European Environment Agency (Charts, Maps, Indicators, Datahub)"
- International Monetary Fund (IMF) Data
- International Energy Agency IEA
- "Consortium of European Social Science Data Archives (CESSDA) DC Data Catalogue"
- Global Biodiversity Information Facility (GBIF)
- "Intergovernmental Panel on Climate Change Data Distribution Centre (IPCC DDC)"
- Climate Policy Database
- UN Department of Economic and Social Affairs Statistics - SDG Indicators Database"
- World Resources Institute (WRI)

The GBIF dataset is used by pilot 2. Pilot 2 is also using data from human build infrastructures (e.g. roads) that can condition habitat connectivity and could consider Farm Accountancy Data Network (FADN) as another source of connectivity barriers created by the presence of livestock. Pilot 3 is considering using some of this socioeconomic data to correlate the evolution of emission with population or industrial development parameters. Due to the local scope of the pilot 1, the kinds of data needed were not included in this general study that ignores datasets coming from municipalities and similar sized entities.

The **tables in the Annex I** present the identified sources of socio-economic data which could be relevant to the GDDS. The methodology to collect the data goes as follows: first of all, different providers of possible socio-economic data were identified among the consortium partners. For each of them, the types of available socio-economic data were distributed into different categories corresponding to the second column of the table. These categories consist of the domains where the socio-economic data are relevant for. The third column lists all the databases associated with a given data provider or owner; it encompasses some statistics about the datasets if they are publicly available. The URLs and domain names to get the datasets are written in the fourth column. Generally, this is a main portal from where the users can get the datasets. The fifth column lists the APIs to obtain the datasets. Often, there is more than one way to get the datasets, in particular if the data are available under different file formats. The technical documentation related to a given API is also mentioned if the API is publicly available. This allows the automation of the retrieval and the collection of socio-economic data by third parties. The sixth column describes the accessibility to read



the data and the limitations related to the datasets. This column also provides the URL to the terms and conditions published by a data owner. This column is somehow linked to the last column which takes care of the IPR aspects. The types of licences associated to the datasets are cited and most of them are open licences such as Creative Commons licences, by applying open access licences, data owners are encouraging the reuse of the socio-economic datasets. When some restrictions related to the use of a particular dataset are in place, related information is available in the metadata or in the documentation of the dataset. Finally, the data privacy is also mentioned in the table and for each data owner, the URL to the privacy policy is displayed.

2.3.2 Addressing Heterogeneous Data Sources

Heterogeneity of socio-economic data formats

The various heterogeneous sources of socio-economic data provided by the different organisations may use different formats. Typically, different ways are employed to retrieve socio-economic data and related metadata, which encompass shapefiles, GeoJSON, CSV files, relational databases, RESTful APIs. The Task T4.1 has collaborated with the WPI “Semantic Interoperability Space” to identify and find some common solutions to address the heterogeneous data sources. In the general context of the AD4GD GDDS, the use of the OGC SensorThings API and its associated data models was proposed to ensure the interoperability and the integration of the data, but socio-economic data does not fit well with the observations model used in SensorThings API. Of course, the tools described and developed in the context of semantic interoperability in the WPI are useful to solve the issues associated with the heterogeneous data models and formats. Indeed, different formats, units, scales and resolutions are available through the different sources of socio-economic data. Another problem is the geographical and resolution scales concerning socio-economic data, IoT data and satellite data; they are different and this can condition the results of an analytical model that combines them. In many cases when these different kinds of data are put together, the heterogeneity of the data appears in terms of visual representations. Generally, the locations of IoT sensors are represented as a point which corresponds to the exact location where the IoT sensors are installed. Data derived from satellite data, like those provided through the Copernicus services, are presented in formats such as GeoTIFF, Shapefiles, NetCDF and Zarr representing regular grids of multidimensional cubes. The socio-economic data are linked to the countries, the administrative boundaries. Furthermore, any change on the administrative borders should be also taken into account when retrieving socio-economic data; indeed, the updates on the administrative borders are not forcibly reflected in the socio-economic datasets, in particular if they are relatively old. The statistical data, which can be part of the socio-economic data, can present delays in terms of publication; indeed, typical examples are the economic data which are compiled after the end of the current year and published the following year. Concretely, we cannot have near real time data in the same way as we can get it from IoT sensors. This latency can affect the choice of particular socio-economic data if other means to get similar ones are possible, notably through open data servers hosted and managed by third parties.

Concerning the semantic interoperability of the socio-economic data, some challenges were identified by AD4GD project:

- Finding the relevant datasets;
- Understand the structure of the datasets;
- Finding the meaning of the codelists and vocabularies used by the socio-economic datasets.

These challenges were tackled during the development and the usage of the semantic interoperability tools deployed in the context of the WPI.

Heterogeneity of socio-economic APIs

The heterogeneity of the APIs to reach out the socio-economic data is a challenge which is addressed by implementing these APIs at the border of the GDDS for receiving socio-economic datasets relevant to the

pilots. Each source of socio-economic data is using its own API, which requires additional work to integrate them in the GDDS. The OGC APIs are proposing a way to distribute socio-economic data in the GDDS to collect and store the socio-economic data and to expose it to the data space users. Semantic interoperability can play an important role to reduce the heterogeneity of the data in the GDDS: for instance, semantic interoperability can provide a common vocabulary across different socio-economic datasets published by different organisations through different APIs and data models. This allows a common understanding of the terms used in these datasets.

2.3.3 Next Steps for Socio-economic Data Integration

A basic demonstration has been produced by AD4GD that was focused on pilot 1, in the small lakes located in Berlin. Heterogeneous data were displayed on a map built by different layers of information. Different sorts of data were used such as: IoT sensors located on the small lakes, the satellite data and the socio-economic data. First, the background layer of the map can be selected by the users and encompasses several options consisting of satellite data or administrative regions. The satellite data option is provided by the Sentinel-2 satellite, which is part of the Copernicus programme. The second option displays the borders of the European countries and their administrative regions. This representation is particularly useful to join it with the socio-economic data provided by the Eurostat service. Indeed, the Eurostat socio-economic data are retrieved using a tool developed in the context of the semantic interoperability work done in the WPI and a parameter based on NUTS (Nomenclature of Territorial Units for Statistics) was used to limit the requests to the desired administrative regions. NUTS is a European standard defining the regions of the European Union, the EFTA countries and possible candidate countries to EU². Three versions of NUTS exist:

- NUTS 1 encompasses the major socio-economic regions. The population of such a region is between 3 million and 7 million. A country can typically be a NUTS 1 region.
- NUTS 2 is composed of basic regions populated between 800,000 and 3 million of citizens. These regions are applying regional policies which can be reflected on the collected socio-economic data.
- NUTS 3 consists of small regions with a limited number of citizens, from 150,000 to 800,000.

As the first demonstration was focused on the city of Berlin, we used the NUTS related to Germany. Concretely, NUTS 1 corresponds to the German states (Bundesland), NUTS 2 to a mix of German states and government regions (Regierungsbezirk) and NUTS 3 to the districts (Kreis).

Different sources of data are shown on a map and encompass IoT data from the small lakes, the weather data in Berlin, and the socio-economic data available from Eurostat and satellite data. Some screenshots are shown in the following pages (Figures 2 to 6).

Figure 2: presents the map with the population in the region of Brandenburg as available through the Eurostat online service. The numbers related to the population in the different regions is available at this URL: https://ec.europa.eu/eurostat/databrowser/product/page/demo_r_pjanaggr3

² NUTS Overview: <https://ec.europa.eu/eurostat/web/nuts>

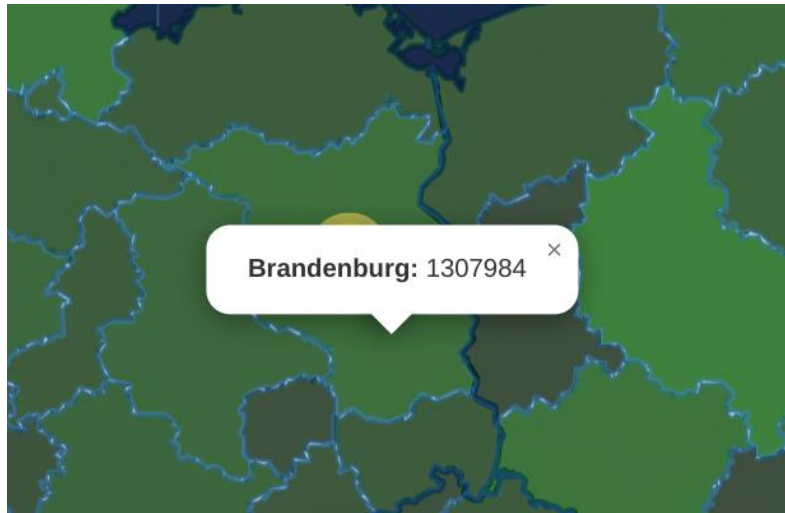


Figure 2: Population retrieved from the Eurostat service.

Figure 3: displays the nitrogen in the region of Oberbayern (Upper Bavaria) and the data are provided by Eurostat, too. The link to retrieve the dataset related to the nitrogen in the different European regions is https://ec.europa.eu/eurostat/databrowser/product/page/ENV_AIR_NO2



Figure 3: Nitrogen value retrieved from Eurostat.

Figure 4: represents the IoT sensors installed in the lakes of Berlin. They are mainly measuring the water level and temperature of the lakes. Currently, 18 lakes of Berlin are equipped with the IoT sensors which generate data collected and stored in the STAplus server (see the components 4 and 5 later in this document).



Figure 4: IoT sensors installed in the lakes in Berlin.

The aggregation of the water temperature in the monitored lakes in Berlin was realised and the result is shown on the map in the Figure 5: . The result is obtained by calculating the value of each water temperature sensor installed in the small lakes. The individual values are retrieved from the STApplus server.



Figure 5: Average of water temperature in Berlin.

Finally, the weather data in Berlin was also retrieved and is available through a chart on the map. All the data can be filtered to display only the interesting measurements for the users at a given time. For instance, Figure 6: displays the air temperature in Berlin during 24 hours. The data related to the weather in Berlin are coming from this public API: <https://brightsky.dev/docs/#/operations/getWeather>. This API retrieves the data published by the official Deutscher Wetterdienst (DWD)³ for the city of Berlin.

³ DWD Open Data Server: <https://www.dwd.de/EN/ourservices/opendata/opendata.html>

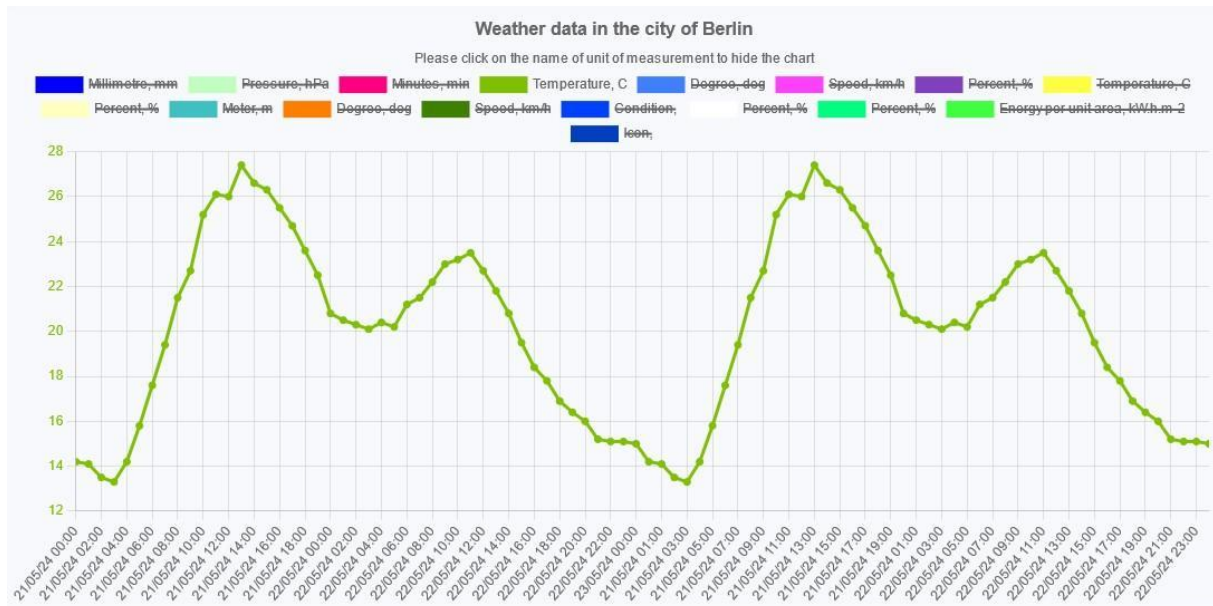


Figure 6: Air temperature in Berlin during one year

The demonstration through the map has showcased that it is possible to represent different kinds of data (socio-economic, satellite, IoT) through different layers but at the same online service. Of course, more data integrations and improvements will be done in the second part of the project.



3 IOT & LOW-COST SENSORS

Over the last 15 years, IoT infrastructures have seen considerable growth from simple development and test environments to substantial deployments that now offer relevant and reliable services in a wide set of environments. This is mainly due to advances in two technology lines: communication protocols for low-power, long-range and wide-area networks, such as LoRa, SigFox or NarrowBand IoT; and smaller, more versatile and power-efficient sensor devices. More specifically, the evolution of sensing technologies for a multitude of physical parameters, supported by the requirements of these low-power networks, facilitates the rapid and low-cost deployment of multi-purpose IoT infrastructures. In many of today's monitoring application scenarios (smart cities, environmental evaluation, nature observation, etc.), these low-cost sensor deployments are replacing expensive individual sensors as aggregated measurements of several low-cost sensors is, at least, as accurate as a single dedicated device, while being cheaper and more resilient.

In turn, these IoT infrastructures, built by dozens or hundreds of low-cost sensors, are a very valuable source of data, as they can constantly report large amounts of data without loss.

3.1 DATA SOURCES IN PILOTS

This section provides a quick overview of the datasets coming from IoT infrastructures involved in each pilot, according to the pilots' targets.

3.1.1 Pilot 1: Water Availability/Quality

Water bodies are typically analysed using water samples or sensors. Standard sensors for describing water quality measure:

- pH,
- electrical conductivity,
- oxygen concentration,
- chlorophyll content or
- water temperature.

The hydrological conditions of water bodies are often assessed using water level and flow measurements.

Water sensors usually have a data logger, so they only need to be removed from the water and read out sporadically. More and more frequently, data is also transmitted directly from the sensor in real time.

In Berlin, water quality data measured in the laboratory as well as sensor data are made available in real time via an interface called *Wasserportal*⁴. It contains current and historical measured values for rivers, lakes, soil water and groundwater. Small lakes (< 50 ha) are currently rarely equipped with sensors. An exception are lakes with bathing spots, where online measurements of water level and water temperature are available. The data is publicly available (*Data licence Germany - attribution - version 2.0*) and can be downloaded via API. However, one obstacle to data utilisation is currently the language barrier, as the metadata is only available in German. In addition, the measured water parameters are not linked to existing vocabularies.

In AD4GD, the data relevant for pilot study 1 is integrated via STA in the FROST server. CSV files containing the data related to the water level and temperature are retrieved from the *Wasserportal*. A process described in detail in the deliverable D3.1 "Heterogeneous IoT Data Integration Model" [2] allows to convert the data from the CSV files to a data model compliant with the one specified by the OGC SensorThings API. Then, requests containing the data from Berlin pilot are sent to the FROST/STA server to store the data for further data processing actions.

⁴ <https://wasserportal.berlin.de/>

In order to improve the data availability of small lakes, three sensors for determining the oxygen concentration were also procured as part of AD4GD. The sensors are placed in a lake for an entire season (spring to autumn). Three different sensors were acquired in order to take into account different common data transmission protocols. The sensors are as follows:

- In-Situ VuLink CI with RDO Blue sensor, data transfer via Mobile network to In-Situ HydroVu or any other FTP server
- Narrowband IoT logger with RDO Pro-X sensor
- Aquatos web LTX with GSM/GPRS data transfer to TerraTransfer server and automated export to other FTP server

In AD4GD, the oxygen concentration, oxygen saturation and water temperature are integrated via STA in the FROST server. These different types of data are converted into data models used by the OGC SensorThings API through a data conversion process described in the deliverable D3.1. This conversion ensures that all the properties related to the metadata and data are using the right format compliant with the OGC SensorThings API. Afterwards, the data is sent to the FROST/STA server through the different endpoints provided by the OGC SensorThings API on the FROST/STA server. Of course, it is possible to retrieve the data stored on the FROST/STA server through the ODATA queries requests available in the OGC SensorThings API.

All additional data provided by KWB will be publicly available under the creative commons licence.

3.1.2 Pilot 2: Biodiversity

Most biodiversity observations about species occurrences and absences are done by human observers that are equipped with cameras. Observations and pictures are catalogued and shared in web portals such as *iNaturalist*⁵ and do not require IoT sensors. This is particularly true for plants, birds and butterflies. However, some mammals are particularly tricky to observe and citizen scientists and experts are turning to camera traps as an alternative. These cameras are capable of detecting motion and generating pictures. They are also capable of illuminating the scene with infrared light for night vision.

Camera traps have been recognized as a way to demonstrate the efficiency of biodiversity corridors between habitat patches with low connectivity for the biodiversity case in AD4GD. Unfortunately, most of these cameras are not collected by the network (due to energy consumption and the lack of mobile telephone networks in remote areas) and require users to visit them and collect the data. An application should be built to get the data from the micro-SD card in the camera trap and extract the relevant images and upload them into a common portal such as *Mammal Web*⁶. A more automatic way of doing this will be preferable for an IoT system as well as a way to integrate these observations in STApplus and a FROST server. AD4GD is working in this line and some experiments in north-east Catalonia are being deployed.

3.1.3 Pilot 3: Air quality

Air pollution poses a major health hazard globally, yet effectively monitoring it remains difficult. On one hand, ground-based measurement networks are often sparse; on the other, high-resolution urban/regional air quality modelling is resource-intensive. Utilising Internet of Things (IoT) and low-cost sensors (e.g., from citizen science projects), offers a promising solution to these problems by providing detailed insights into air pollution on a local scale. However, these new observations come with several limitations.

⁵ <https://www.inaturalist.org>

⁶ <https://www.mammalweb.org/>

To investigate the potential of how IoT data can enhance air quality monitoring and thus also assessment of health impacts, the focus so far has been on 2 datasets from CitizenScience initiatives: *Sensor.Community*⁷ and *PurpleAir*⁸.

Both networks use low-cost nephelometers to determine PM_{2.5}⁹ and PM₁₀¹⁰ concentrations. Within Europe, around 14,000 stations are operated in the *Sensor.Community* network, the majority of which are located in Central and (Southern) Eastern Europe. *PurpleAir*, on the other hand, is mainly found in the UK and operates around 700 stations in Europe.

The data provision is different for the two networks: *Sensor.Community* pursues an open data policy and makes the data available online in `csv` format in real time and in a long-term archive simply via `http`. The *PurpleAir* data, on the other hand, is subject to payment, but until March 2024 it was also possible to obtain these measurements via the *OpenAQ*¹¹ platform via Amazon Web Services.

Despite the general availability of data, there are some challenges and difficult boundary conditions for carrying out further processing with the data. These include:

- Insufficient technical documentation of the low-cost sensors,
- Potential malfunctions of the devices (e.g., due to vulnerable electronics),
- Maintenance by non-experts.

Moreover, the data is not standardised and has undergone quality control. The pilot study will therefore initially focus on establishing such a framework/pipeline. Components of this could be, for example, a scheme for filtering outliers or an implementation for correcting systematic biases.

3.2 TECHNICAL DEVELOPMENTS: SOLUTIONS PROPOSED

All technical developments presented in this section reference the components that were presented and shortly described in section 6 of deliverable "Pilot technical implementation planning, implementation and assessment".

3.2.1 STA and STAplus (Components 4&5)

The Work Package 3 has elaborated a model for the integration of heterogeneous IoT data to a STAplus server instance. It allows the mobilisation of data, in particular, IoT sensor data, to a STAplus server. Open Geospatial Consortium (OGC) has specified the SensorThings API (STA) and an implementation of a server using the SensorThings API was installed in the IoT Lab research infrastructure. Furthermore, a plugin named STAplus has been added to this server to make it aligned with the FAIR principles. Then, the STA server became a STAplus server, providing an extended data model which is able to manage the parties, the licences, the campaigns, the observation groups and the relations.

Each entity representing a *Thing*, a *Sensor*, a *Datastream* or an *Observation* is stored into the STAplus server and can be reached by a unique URL. There are two communication protocols which can be used to send the data to the STAplus server through the **OGC SensorThings API**: using *HTTP* and *MQTT* protocols.

A first approach to integrate IoT data from the pilot organised in Berlin was elaborated and tested by using the `.csv` format files which contain the data generated by the IoT sensors installed in the small lakes. Based on these datasets, different types of measurements were identified and pushed to the STAplus server. This requires a process which consists of creating *JSON* files encompassing the metadata to be stored within the IoT data in the STAplus server. Each *JSON* file was pushed to the STAplus server through the OGC SensorThings API. Then, each individual `.csv` file was read by column and all useful data was extracted and

⁷ <https://sensor.community/>

⁸ <https://www2.purpleair.com/>

⁹ PM_{2.5} is particulate matter 2.5 µm or less in diameter.

¹⁰ PM₁₀ is particulate matter 10 µm or less in diameter.

¹¹ <https://openaq.org/>

eventually converted in the correct format, notably the timestamp. Finally, the extracted content of each .csv file was sent to the STApplus server through as usual the OGC SensorThings API.

This approach was tested and validated; indeed, it was possible to retrieve the data generated in Berlin from the STApplus server instance using the different methods offered by the OGC SensorThings API.

A second approach was elaborated and successfully tested using TAPIS (Tables from OGC APIs for Sensors), a tool developed and maintained by CREAM. A new process consists of combining the .csv files with semantic tagging. A Graphical User Interface (GUI) is employed to realise the semantic tagging.

Another approach will be tested using images coming from camera traps deployed in pilot 2 to detect endangered terrestrial fauna. This data will be encoded in STApplus considering privacy issues (i.e., the exact location of the camera and the specie observed) and then send to FROST server.

More details can be found directly in the AD4GD deliverable D3.1 "Heterogeneous IoT Data Integration Model".

3.2.2 IonBeam (Component 1)

As part of T4.2, ECMWF is developing open-source infrastructure named *IonBeam* whose role is to help ingest high volume IoT datasets into ECMWF's existing infrastructure, see Figure 7: , including the Climate and Atmosphere data stores. IonBeam is a tool for building scalable data processing pipelines out of modular and reusable software components.

Our approach aims to complement the existing work on STA and other technology by focussing on high volume use cases such forecasting and bulk data archival. This focus generates a very different set of priorities and trade-offs from those made by STA and other data models.

3.2.2.1 IonBeam

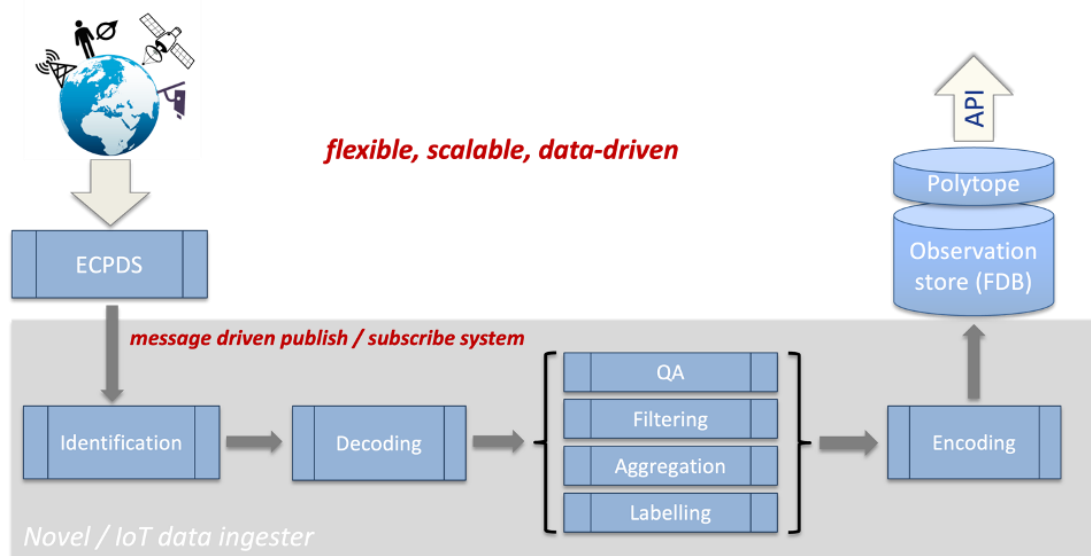


Figure 7: A high-level block diagram showing how IonBeam fits into ECMWF's existing infrastructure.

IonBeam is based on a Message/Action architecture designed to support near real time processing. In this model, actions both consume and produce messages, including decoding binary chunks of data, reformatting tabular data and attaching appropriate metadata, re-aggregating data into chunks according to appropriate time windows, encoding data into standard formats and outputting the data into an instance of ECMWF's Fields Database (FDB) for later use. Actions specify which messages they match, according to their metadata. And as such, the metadata is used to route the messages through the system from arrival to final output.

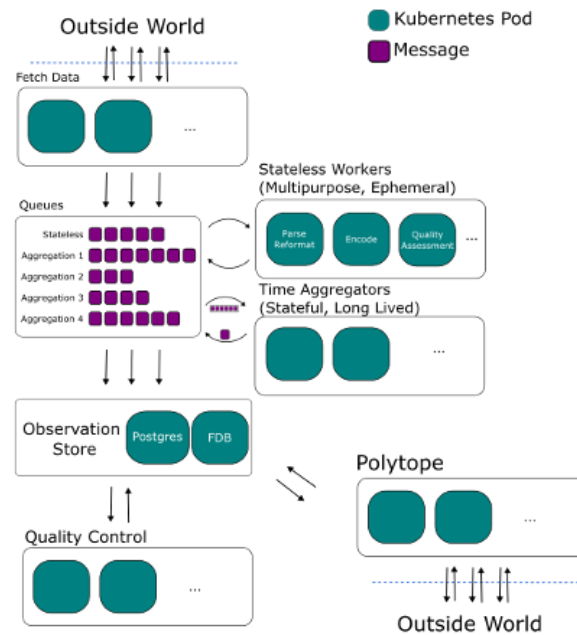


Figure 8: The current architectural implementation of IonBeam in Kubernetes, using RabbitMQ as a message queue.

Stateless actions are handled by generic workers and thus the system throughput is scalable to multiple worker nodes running in parallel. All actions except time aggregation are stateless: they take one message and produce one or more output messages whose contents are entirely determined by that single input. Time aggregation requires combining information from multiple messages and so requires a longer lived, stateful process to manage it.

Scalability. IonBeam is designed with high-frequency, high-volume IoT data in mind, anticipating a gradually increasing throughput and short-term peaks in demand.

Configuration. IonBeam data pipelines are created using .yaml format based config files which will be accompanied by extensive documentation on their structure and use to define flexible data processing pipelines.

Fault Tolerance. IoT data does not come with guarantees that it will be well formed or correct, it will sometimes be invalid or violate our expectations. IonBeam is designed defensively to handle this. This includes mechanisms to quarantine problematic data until a human operator can intervene and to replay data or remove data once the issues have been resolved. Systems themselves may also fail, and the infrastructure gracefully handles failures of its own, or external organisations, infrastructure, and to catch up after failures, avoiding data loss.

Reproducibility. As the ingested IoT data will ultimately be used scientifically at ECMWF and elsewhere, it is important that all data processing be reproducible. The system archives raw input data and stores provenance information such that the original data and exact series of transformations made to the data can be reproduced later. For example, if observations are added to the system a long time after they are taken, those observations may affect the outcome of analyses that have already been run. These ensure that bit-identical results can be obtained even after later data has arrived.

Curated metadata. All of ECMWF's data infrastructure and workflows are driven by semantically and scientifically meaningful metadata. All data is uniquely described by a metadata language, according to a predefined schema. IoT data is highly heterogeneous and consequently requires a careful approach to bringing this data into such a metadata-driven ecosystem.

3.2.2.2 Connection to the Air Quality Pilot



The first application of the IonBeam infrastructure will be the automation of the IoT collection, cleaning and ingestion workflow that is currently being manually performed as part of the Air Quality Pilot.

3.3 FUTURE WORK

One of the identified problems is how to automatically attach metadata to data generated by heterogeneous IoT sensors deployed in the pilots. Indeed, a large amount of data provided by the IoT sensors installed in the small lakes in Berlin is available, but the metadata is missing or not associated directly to the data. Possible future work is to create automated processes to add the metadata to each IoT sensor data which is collected in the STAplus server.

Similar issues exist for IonBeam, currently the raw data can be tagged with additional metadata specified in .yaml config files. However, it remains an open question whether this format can be sufficiently flexible to support heterogeneous use cases without becoming overly complex. Authentication and authorisation for the data space are other areas of future work. This work requires specific plugins and settings to ensure the enforcement of user access policies on the STAplus server and other components sharing or using the data notably generated by the IoT sensors deployed in the three pilots.

4 MULTIDIMENSIONAL DATA CUBES (COMPONENT 6)

Current challenges imply the use of data from diverse sources, formats and thematic approaches, and from different temporal resolutions. All these require multidimensional analysis and aggregation methodologies. Data Cubes aim to provide a technological solution to these needs. Following the definition from the OGC¹², “a data cube is a multi-dimensional (“n-D”) array of values; a data structure that can be 1- dimensional, 2- dimensional, 3- dimensional, or higher-dimensional. The dimensions may be coordinates or enumerations, e.g., categories” (Figure 9:).

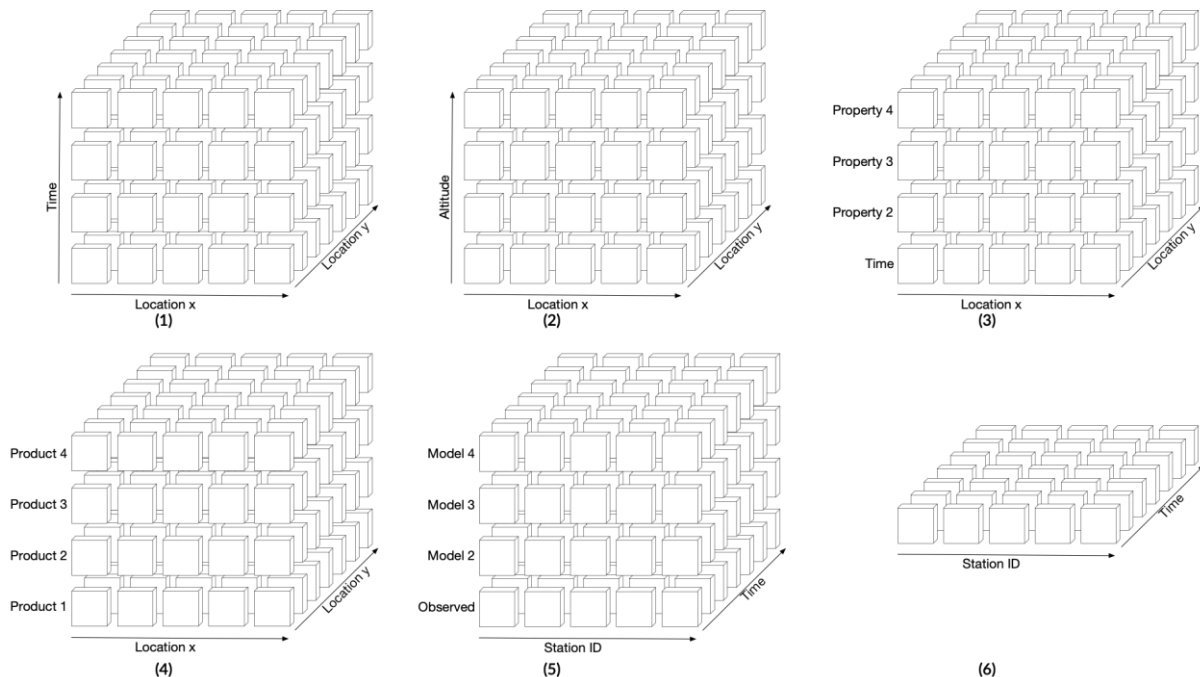


Figure 9: Different implementations of data cubes (extracted from <https://www.ogc.org/initiatives/gdc>).

Several technologies exist nowadays, among them: Open Data Cube (ODC), Rasdaman, Data Cubes with STAC, etc. Some successful examples using ODC are the Brazilian Data Cube, the Swiss Data Cube Platform as a Service, and the Ukrainian Open Data Cube (used for computationally expensive processes such as crop type identification and land cover classification and change detection over time based on machine learning).

Also, several sets of client-server implementations are being developed upon these technologies allowing to interact with the data behind, such as OGC API – Maps, - Tiles, or -Coverages, WCS, openEO [4] (an API and processing environment with data cubes generated on the fly based on users' needs), etc. Interoperability among all these technologies is still an important issue to solve. AD4GD, together with the sister projects B-cubed and Fairicube, is working in the technological part (by implementing some experiments in ODC and Rasdaman), as well as in addressing interoperability challenges by analysing OGC API standards along with openEO.

Rasdaman combines well-matured and standardised models such as the OGC Coverage Implementation Schema (CIS) with the Web Coverage Service (WCS).

In AD4GD, Data Cubes are specially being tested in pilot 2. Some approaches will also be considered for pilot 1.

4.1 DATA SOURCES IN PILOTS

¹² <https://www.ogc.org/initiatives/gdc/>



4.1.1 Pilot 1: Water Availability/Quality

In the case of pilot 1, no technological approach has been yet decided as a back-end to store and deal with multidimensional data. Lessons learnt from pilot 2 in terms of Data Cube technologies and standards to interrogate and process data, will be adopted here, considering the particularities of the pilot as well regarding data sources, temporal and spatial resolutions of the small lakes in Berlin.

4.1.2 Pilot 2: Biodiversity

A few sources of multidimensional data have been used to obtain connectivity outputs, including time series of land-use/land-cover (LULC) raster data, expert-derived impedance and affinities data, vector data subsets retrieved from Open Street Map portal, refining LULC data, protected areas through from WDPA API (World Database on Protected Areas)¹³. The potential of GBIF species occurrence data and camera traps data from Citizen Science projects has been explored to verify connectivity outputs and provide additional visualisation for front-end GUI application.

As soon as connectivity data are provided as time series (1987-2022), may be extended through the same semi-automatic workflow and used to forecast connectivity based on different scenarios of LULC change, the combination of input LULC data and output connectivity with relevant, consistent and overarching metadata is considered multidimensional data.

Regarding the structure of this multidimensional data in pilot 2, two Data Cube technologies are being tested: Open Data Cube and FAIRiCUBE Rasdaman.

All Sentinel-2 images for Catalonia from 2015 to the present with a cloud filter of 80%, have been downloaded from the Copernicus Data Space and prepared, similarly to other ODC experiences already working (i.e. the Swiss Data Cube). All images have been first transformed to Cloud Optimized GeoTiff standard¹⁴ and then ingested into the PostGree Database which ODC uses for organising the arrays of data.

Apart from these basic reference datasets, other thematic data is uploaded which have different dimensional properties. In this sense, the already computed Terrestrial Connectivity Maps for 1987, 1992, 1997, 2002, 2007, 2012, 2017 and 2022 have been successfully uploaded into the system. However, these resulting products from LULC maps could be tried to derive directly from S2 images using the Data Cube and ML workflows like the Brazilian SITS API does and the Ukrainian Data Cubes also provide. It's still in an experimental phase in AD4GD between ATOS and CREAM. This will also allow for future scenario connectivity computation.

Other datasets that could be relevant to understand Terrestrial Connectivity in a multidimensional scope and that will be ingested into the ODC are: time-series climatic data from Catalonia (temperature, pluviometry), biodiversity indicators (FAPAR, GPP, landscape indicators, etc.), socioeconomic variables (population density, agrarian activities indices, infrastructures, etc.), among others.

Besides these rather more "classical layers", GBIF species occurrence information will be also analysed and incorporated. In this particular case, data is distributed in point vector format which needs to be aggregated in grids (typically in 1 km or 10 km grids). Several approaches can be used here, from individually downloading these vectorial point data and being aggregated *a posteriori*, to directly using the GBIF API for downloading data already aggregated in grids. These particular tests are being shared between AD4GD, B-cubed and FairiCube projects. Dedicated meetings are being held in this sense, and, in particular, AD4GD has participated in the B-cubed Hackathon on Biodiversity Data Cubes, to specifically work in this direction.

Finally, data collected from Camera traps will also be incorporated into the Data Cube. In this case, apart from the multidimensional aspect and the aggregation procedures, this type of data brings other particularities to be treated, such as the privacy of sharing sensible information regarding camera location, species location and description. Sensible information is a crucial point when developing a GDSS.

¹³ <https://api.protectedplanet.net/documentation>

¹⁴ <https://docs.ogc.org/is/21-026/21-026.html>

4.1.3 Pilot 3: Air quality

Various multi-dimensional datasets are used to further improve the processing. Typical model data in meteorology are provided on gridded multidimensional data, which allows for comprehensive spatial and temporal analysis. This standardised format is crucial for effectively integrating and comparing different datasets.

In our case, we focus on data from CAMS regional model analysis and the ECMWF fifth reanalysis (ERA5). Both the CAMS and ERA5 datasets are provided via the Copernicus Climate Data Store (CDS, <https://cds.climate.copernicus.eu/#!/home>) and the Copernicus Atmosphere Data Store (ADS, <https://ads.atmosphere.copernicus.eu/#!/home>) respectively in the form of a data cube in netCDF4 format, facilitating further data processing for the pilots. These data cubes already combine various dimensions of data (i.e., time, longitude, latitude, and model level), eliminating the need to create a new data cube from scratch, as was done in Pilot 2 with satellite images.

Additionally, both datasets are available via dedicated APIs (CDSAPI and ADSAPI, respectively), simplifying data access and integration. These APIs allow users to programmatically retrieve data, enabling automated workflows and efficient data handling.

4.1.3.1 European Air Quality Forecast

To combine the IoT data, gridded data from the analysis fields of the CAMS European Air Quality Forecast are used. These fields are produced by CAMS for the European domain at a spatial resolution of 0.1° in longitudes and latitudes (approx. 10km) every hour. The production is based on an ensemble of eleven air quality forecasting systems across Europe. This product provides a median of the ensemble calculated from individual outputs, which is beneficial since ensemble products yield, on average, better performance than the individual model products. Our focus is on the data fields of particulate matter (PM_{2.5}, PM₁₀) and NO₂, depending on the availability of corresponding observations from IoT/low-cost sensors.



The dataset provides daily air quality analysis and forecasts for Europe.

CDS provides specific daily air quality analysis and forecasts for the European domain at significantly higher spatial resolution (0.1 degrees, approx. 10km) than is available from the global analysis and forecasts. The production is based on an ensemble of eleven air quality forecasting systems across Europe. A median ensemble is calculated from individual outputs, since ensemble products yield on average better performance than the individual model products. The spread between the eleven models are used to provide an estimate of the forecast uncertainty. The analysis combines model data with observations provided by the European Environment Agency (EEA) into a complete and consistent dataset using various data assimilation techniques depending upon the air quality forecasting system used. In parallel, air quality forecasts are produced once a day for the next four days, with the analysis and the forecast are available at hourly time steps at seven height levels.

Note that only nitrogen monoxide, nitrogen dioxide, sulphur dioxide, ozone, PM_{2.5}, PM₁₀ and dust are regularly monitored against in situ observations, and therefore forecasts of all other variables are unvalidated and should be considered experimental.

More details about the product are given in the documentation section.

DATA DESCRIPTION	
Data type	Gridcell
Horizontal coverage	Europe (west boundary=25.0° W, east=45.0° E, south=30.0° N, north=70.0° N)
Horizontal resolution	0.1°x0.1° (10km x 10km)
Vertical coverage	Surface, 50m, 100m, 250m, 500m, 750m, 1000m, 3000m, 5000m
Temporal coverage	Pre-season rolling archive
Temporal resolution	1-hourly
File format	GRIB, NetCDF
Update frequency	Daily

Figure 10: Description of CAMS European air quality dataset on the ADS web portal including scientific documentation and metadata.

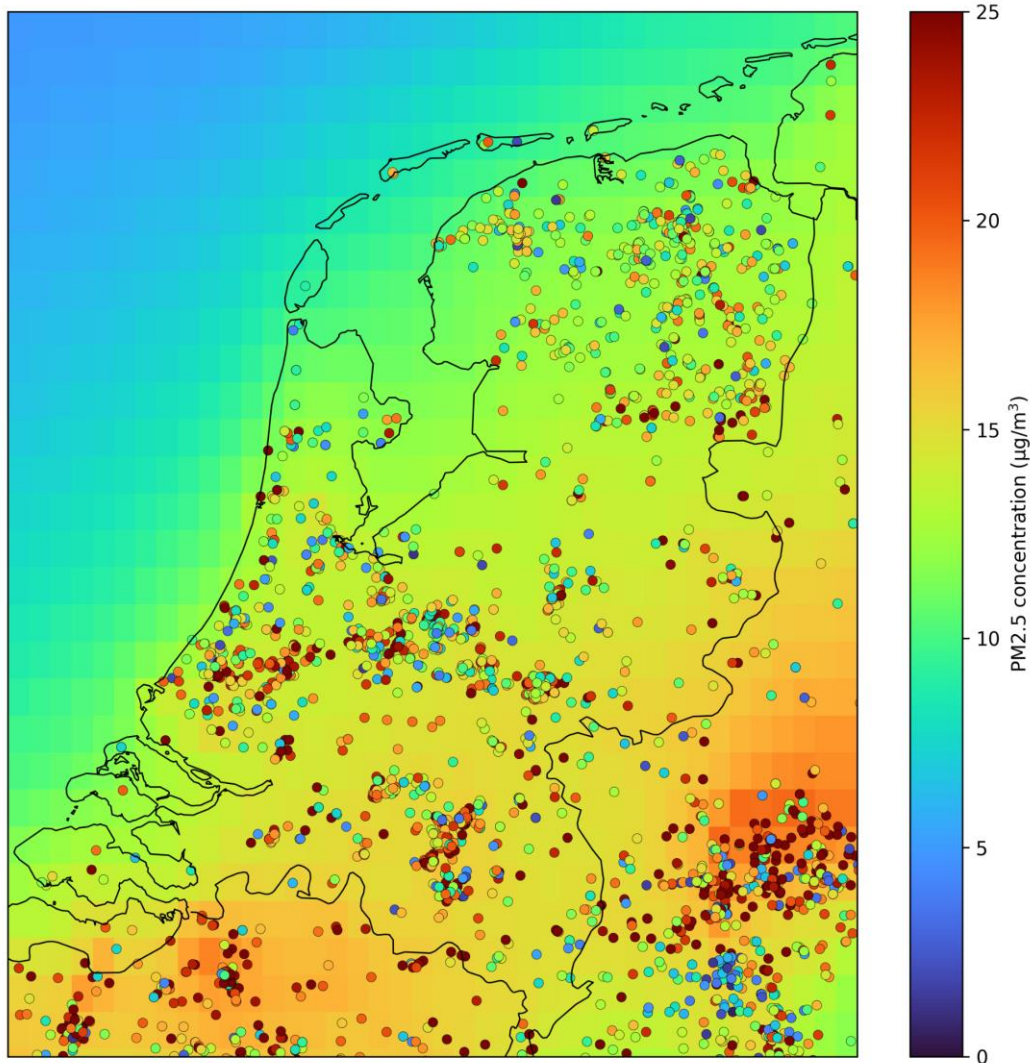


Figure 11: Monthly mean PM2.5 concentration for December 2022 across the Netherlands with background determined from CAMS Europe regional model analysis and circles indicating the PM2.5 concentration at a given IoT station.

4.1.3.2 Reanalysis ERA5 data

Furthermore, we also include meteorological data from the fifth-generation ECMWF reanalysis ERA5 (Hersbach et al., 2020). This reanalysis dataset combines model data with in-situ and remote sensing observations from various different measurement platforms. This results in a multidimensional dataset with an hourly temporal resolution, 0.25° horizontal spatial resolution on several different vertical height levels for all types of meteorological variables, providing a unique overview of the atmospheric state from the past to the present.

4.2 TECHNICAL DEVELOPMENTS: SOLUTIONS PROPOSED

All technical developments presented in this section reference the components that were presented and shortly described in section 6 of deliverable “Pilot technical implementation planning, implementation and assessment”.

4.2.1 Kriging Interpolation Algorithm

Within the air quality pilot, one approach that we are using is Kriging interpolation. This is a geostatistical method that allows one to optimally interpolate observations from arbitrary spatio-temporal coordinates to a structure grid, essentially a datacube. The method is optimal subject to assumptions about the statistical behaviour of the observed parameters. More precisely, the interpolation weights are determined from the relationships between the individual point measurements (i.e., their spatial autocorrelation or covariances).

For this purpose, a fundamental tool in Kriging, the variogram models the spatial autocorrelation between sample points, thereby representing how the similarity between values changes with distance, as commonly, points closer to each other exhibit similar values. This variogram is then used to derive the interpolation weights, ensuring that the estimates are unbiased and have minimum variance.

A decisive advantage over methods such as optimal interpolation is the non-necessity of a priori information or, in comparison to inverse-distance-weighted interpolation, the prevention of so-called bull's-eye effects. Moreover, Kriging also provides a variance or error of the interpolation, enabling an estimation of the quality of the result.

In the air quality pilot, we are using the Ordinary Kriging method to interpolate the Sensor. Community PM_{xx} readings (e.g., to the same grid as the CAMS Europe air quality forecast). This alignment allows for a more direct comparison for instance to satellite or model data sets, and using a logarithmic scale prevents unphysical interpolation results, such as negative PM_{xx} values. Overall, Kriging offers a robust method for spatial data analysis, creating detailed and accurate spatial representations from diverse datasets.

Figure 12: shows an exemplary application of the Kriging technique for the city of Sofia [5] in Bulgaria. The PM_{2.5} observations from irregularly distributed IoT stations are transformed to a regular latitude-longitude grid with a resolution of about 250x250m². This allows for an identification of pollution hotspots within the city, but also a simplified further combination/processing with other geophysical/geoinformation datasets.

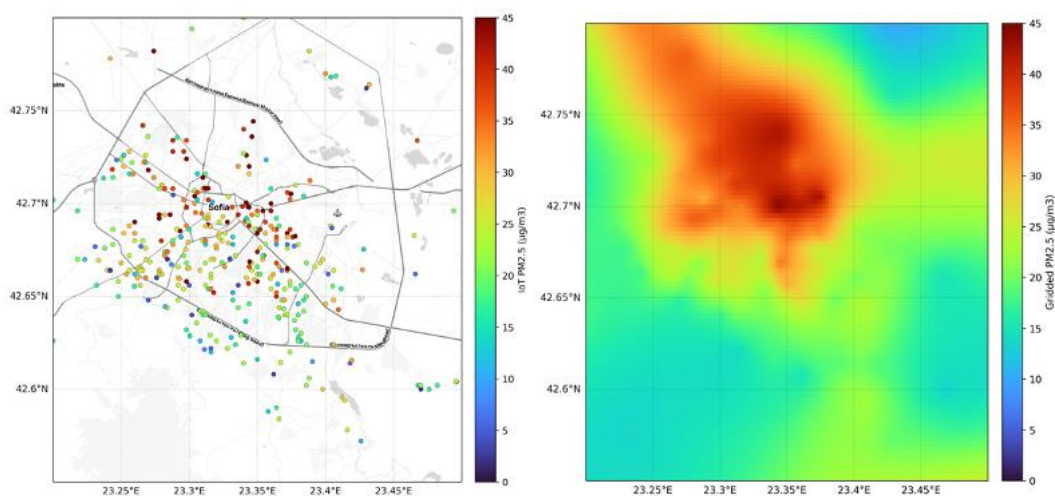


Figure 12: Distribution of available IoT stations (left) and gridded PM_{2.5} concentration derived from those stations (right) in Sofia for December 2022. Color shading indicates monthly mean PM_{2.5} concentration. The regridding was performed on a 250x250m² resolution.

4.2.2 Open Data Cube Extensions

In terms of accessing the ODC technology, several standards are being analysed:

- OGC Web Coverage Service (WCS)¹⁵, which has been used and tested for a long time and in several server/clients. WCS offers multi-dimensional coverage data for access over the Internet, but does not allow for computing among data. Another standard is needed for this, the Web Coverage

¹⁵ <https://www.ogc.org/standard/wcs/>



Processing Service (WCPS)¹⁶ which defines a protocol-independent language for the extraction, processing, and analysis of multi-dimensional coverages representing sensor, image, or statistics data. The current status of a WCS implementation in AD4GD is that a basic development has been done to allow GetCoverage requests to the ODC server, but other basic requests such as GetCapabilities (for discoverability) and GetMap (for visualising) need to be deployed.

- An example of a GetCoverage instance allowing to retrieve a particular data from a particular date from the Data Cube: [https://callus.ddns.net/cgi-bin/mmdc.py?SERVICE=WCS&VERSION=2.0.1&REQUEST=GetCoverage&SUBSET=E\(260145,527685\)&SUBSET=N\(4488645,4747995\)&COVERAGEID=TerrestrialConnectivityIndex&FORMAT=image/tiff&RANGESUBSET=Forest&SUBSET=ansi\(%222017-01-01%22\)](https://callus.ddns.net/cgi-bin/mmdc.py?SERVICE=WCS&VERSION=2.0.1&REQUEST=GetCoverage&SUBSET=E(260145,527685)&SUBSET=N(4488645,4747995)&COVERAGEID=TerrestrialConnectivityIndex&FORMAT=image/tiff&RANGESUBSET=Forest&SUBSET=ansi(%222017-01-01%22))
- OGC API - Coverages¹⁷, which is the new building block API proposed by OGC to interact with gridded coverages in multi-dimensional coordinate space. This API allows for computing with data directly from the server.
- OGC API - Maps¹⁸ which allows query operations from dynamically rendered maps retrieved from the underlying data store based upon simple selection criteria defined by the client. This standard does not allow for computation among data either.
- openEO API¹⁹ which is now proposed as an OGC standard (but is not a OGC API) standardises how local clients (R, Python, and JavaScript) can access and operate with EO data in cloud service providers. OpenEO has been chosen by Copernicus Data Space Ecosystem for exploring and processing all Copernicus data. Some approaches are being developed to interact between openEO and ODC technology²⁰.
- OGC API Coverages - Part 2. A proposal for an OGC API Coverage - Part 2: "Filtering, deriving fields, aggregation and convolution" is being moved in the OGC. The proposal adds functionalities that can be combined with Part 1 requests. Part 2 Leverages on OGC Common Query Language (CQL2) for expressing filtering, derived fields, aggregation and convolution in a very natural intuitive syntax. The solution will be further developed in the Testbed 20. The solution is considering a syntax that will be capable of expressing a band index as a filter operation. For example, and Enhanced Vegetation Index can be expressed as:
 - /ogcapi/collections/sentinel2-l2a/coverage?
 - properties=2.5 * (B08 / 10000 - B04 / 10000) / (1 + B08 / 10000 + 6 * B04 / 10000 + -7.5 * B02 / 10000)&
 - filter=(SCL >=4 and SCL <= 7) or SCL=11&
 - subset=Lat(38.9:39.1),Lon(-4.8:-4.6),time("2017-09-04T11:18:26Z")&
 - crs=[EPSG:4326]&scale-size=2000,2000

A deep and thorough analysis is currently being done within AD4GD, B-cubed and FAIRiCUBE projects, regarding the best option to retrieve multidimensional data in data cubes in terms of feasibility, usability and standardisation.

For the moment, AD4GD ODC Data Cube, can be queried through one Jupyter notebook approach. Below, some screenshots on an example query (Figure 13: and Figure 14:):

¹⁶ <https://www.ogc.org/standard/wcps/>

¹⁷ <https://ogcapi.ogc.org/coverages/>

¹⁸ <https://ogcapi.ogc.org/maps/>

¹⁹ Schramm, Matthias et al. (2021). The openEO API—Harmonising the Use of Earth Observation Cloud Services Using Virtual Data Cube Functionalities. Remote Sensing. 13. 1125. 10.3390/rs13061125.

²⁰ <https://openeo.org/documentation/1.0/developers/backends/opendatacube.html>

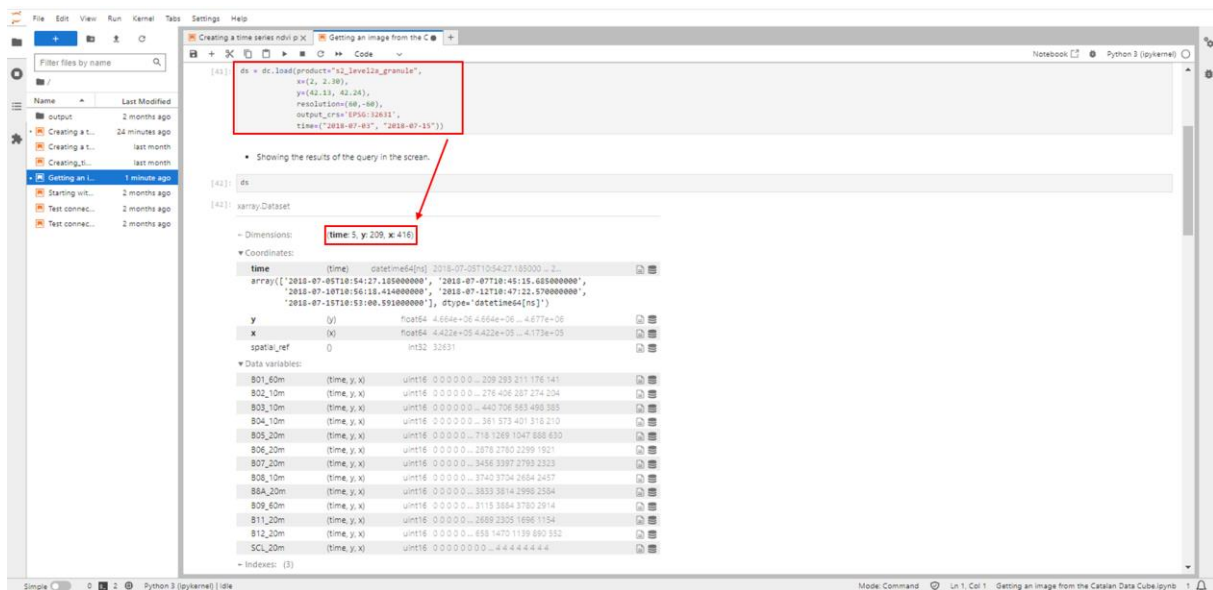


Figure 13: Resulting arrays from a query on a specific BBOX and time range over the whole S2 pool of COG images.

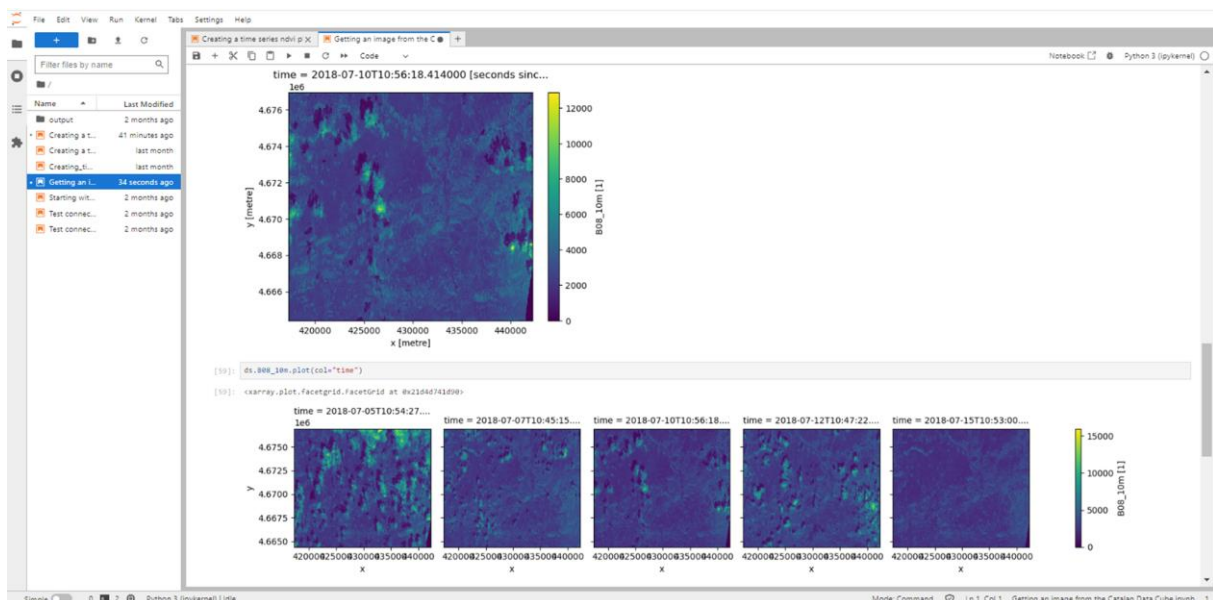


Figure 14: Visualisation of the NIR band for a certain date in the array, and for the whole array of date values.

4.2.3 Rasdaman Data Cube

An instance of Rasdaman data cube²¹ has been deployed on the FAIRiCUBE²² platform through the user credentials assigned to Aston University. By now, data exported to FAIRiCUBE include input data (LULC time series) and output data (Index of Terrestrial Connectivity timeseries for six LULC types, ecological corridors and integral index of connectivity for forests from Graphab). Datasets uploaded do not include raw Sentinel-2 images for Catalonia due to current storage space limitations on uploads per user in FAIRiCUBE. All data deployed may be accessed through WCS (GetCapabilities, DescribeCoverage, GetCoverage and ProcessCoverages which can be run as an interactive WCPS console) or WMS (GetCapabilities, GetMap) GET

²¹ <https://www.rasdaman.com/>

²² <https://fairicube.nilu.no/>

requests²³. However, WCS responses are based in the Coverage implementation Schema that works fine for 2D retrievals but are currently resulting in XML metadata outputs that validate incorrectly if the time metadata is added.

Apart from the usual WMS and WCS queries, FAIRiCUBE is able to run powerful WCPS queries (Figure 15: and Figure 16:) mostly to carry out the spatial analysis. However, it has been revealed that Cloud Optimized GeoTIFFs datasets do not significantly reduce time to execute WCPS queries respect to conventional GeoTIFFs (since grid domain remains the same and all queries still iterate over all pixels, except from null values) and are even slightly more time-consuming than non-compressed GeoTIFF files. So far, Graphab COG timeseries outputs for only one index for the whole Catalonia (only forests) use from 1.84 GB to 7.37 GB in the FAIRiCUBE database, whereas MiraM on outputs sizes are a bit more 100 MB in total. In the Rasdaman instance, data access is restricted by 200 MB, which can complicate the following processing.

```
for $c in (output_connectivity_miramon_ict_forests)
return encode(
```

```
  clip($c[time("2022")] - $c[time("1987")], MultiPolygon (((392419.15593403 4643816.15482667,
374154.87447849 4612164.32804365, 366503.6214363 4580078.42818932, 366503.6214363 4559592.8152054,
365022.73375072 4543796.6798925, 369712.21142173 4536639.05607884, 397849.07744784 4533677.28070767,
457578.21409975 4543303.05866397, 495340.85008216 4567244.06824759, 503979.3615814 4587482.86661724,
500038.32775318 4624998.68798539, 494847.22085363 4640548.00808402, 483000.11936896 4649680.1494118,
453629.18027153 4653875.99785429, 421543.2084172 4651407.85171164, 405500.33049003 4649433.33479753,
392419.15593403 4643816.15482667)))
```

```
), "image/png",
```

```
"{ \"colorMap\":
  { \"type\": \"intervals\",
    \"colorTable\": {
      \"-1\\\": [236, 0, 0, 255],
      \"-0.7\\\": [236, 36, 0, 255],
      \"-0.5\\\": [236, 83, 0, 255],
      \"-0.4\\\": [236, 155, 0, 255],
      \"-0.1\\\": [236, 202, 0, 255],
      \"0\\\": [0, 0, 0, 0],
      \"0.1\\\": [113, 227, 230, 255],
      \"0.2\\\": [28, 204, 0, 255],
      \"0.3\\\": [34, 182, 0, 255],
      \"0.4\\\": [0, 156, 25, 255],
      \"1\\\": [0, 255, 0, 255]
```

```
    }
  }
}
```

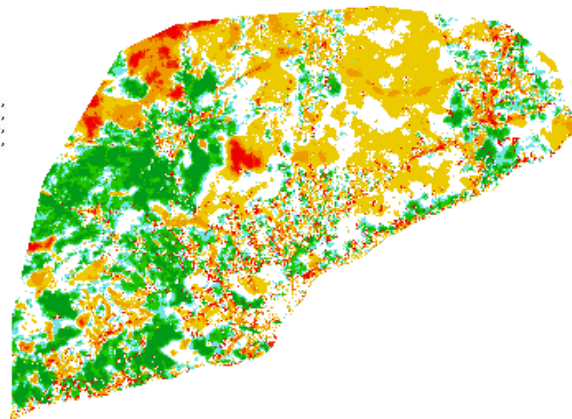


Figure 15: Example of WCPS query in FAIRiCUBE to visualise dynamics in Index of Terrestrial Connectivity (1987-2022) based on the bounding box, covering central Catalonia.

²³ Example of WCS request to Fairicube with user credentials required:

https://fairicube.rasdaman.com/rasdaman/ows?&SERVICE=WCS&VERSION=2.1.0&REQUEST=DescribeCoverage&COVERAGEID=output_connectivity_graphab_iic_forests&outputType=GeneralGridCoverage

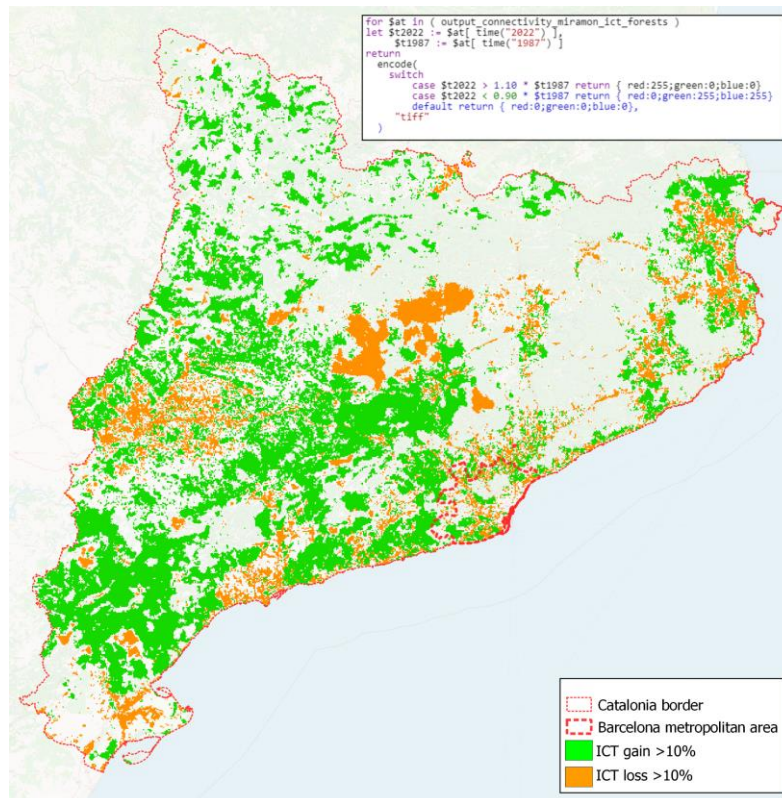


Figure 16: Gain and loss in Index of Terrestrial Connectivity (1987-2022).

One of the main disadvantages is that a FAIRiCUBE instance cannot be accessed directly from frontend GUI tools, for example open.Dash (the one selected by AD4GD as the frontend of the project pilots). Rasdaman capabilities to visualise graphic outputs of any OGC API queries are limited as the FAIRiCUBE instance is currently under testing (Figure 17:). A colour table (or legend) can be defined as a list of RGB definitions (Figure 15:) via the WCPS request. There is a possibility to upload user-specified styles of raster data visualisation through the Rasdaman GUI, but there is no documentation and clear understanding on which formats can be used there.

At the same time, datacubes do not allow to directly ingest vector data to extract bounding boxes or CSV files (GBIF occurrences). Therefore, to implement the data extraction from Rasdaman and include other data formats, additional libraries (Leaflet or Mapbox) are required to be implemented in the GUI tool.

Footprint of layer

NOTE: When a non-spatial axis for a 3D+ coverage is omitted from the WMS GetMap request, then according to the WMS 1.3 standard the top-most spatial slice is selected.

Latitude:

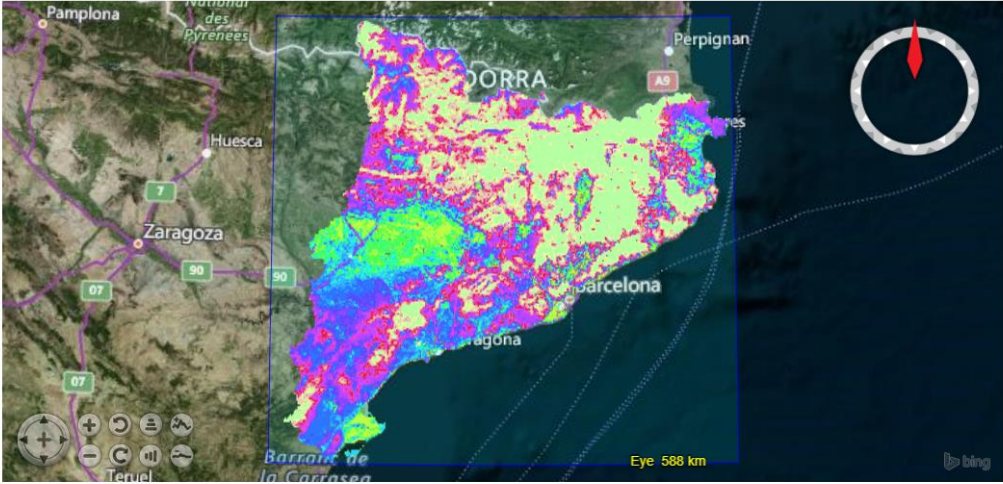
Longitude:

time:

min:"1987-01-01T00:00:00.000Z" - max:"2022-01-01T00:00:00.000Z" - number of values: 8

Display WMS layer on WebWorldWind globe:

Layer's original bounding box in EPSG:4326 CRS is: minLon=0.06, minLat=40.51, maxLon=3.34, maxLat=42.89



GML layer descriptions document

```

<Layer queryable="0" cascaded="0" opaque="0" noSubsets="0" fixedWidth="0" fixedHeight="0">
  <Name>output_connectivity_miramon_ict_forests</Name>
  <Title>output_connectivity_miramon_ict_forests</Title>
  <Abstract> <![CDATA[<ows:AdditionalParameters xmlns:ows="http://www.opengis.net/ows/2.0"><ows:AdditionalParameter><ows:Name>sizeInBytes</ows:Name><ows:Value>14598084</ows:Value></ows:AdditionalParameter><ows:AdditionalParameter><ows:Name>axislist</ows:Name><ows:Value>time,E,N</ows:Value></ows:AdditionalParameter></ows:AdditionalParameters>]]</Abstract>
  5. <CRS>EPSG:25831</CRS>
  <EX_GeographicBoundingBox>
    <westBoundLongitude>0.06391193862943229</westBoundLongitude>
    <eastBoundLongitude>3.338787747599169234</eastBoundLongitude>
    <southBoundLatitude>40.512481119638777112</southBoundLatitude>
  10. <northBoundLatitude>42.88509522613726</northBoundLatitude>
  </EX_GeographicBoundingBox>
  <BoundingBox CRS="CRS:84" minx="0.06391193862943229" miny="40.512481119638777112" maxx="3.338787747599169234" maxy="42.88509522613726"/>
  <BoundingBox CRS="EPSG:25831" minx="260085" miny="4488705" maxx="527625" maxy="4748055"/>
  <Dimension name="time" units="d">"1987-01-01T00:00:00.000Z", "1992-01-01T00:00:00.000Z", "1997-01-01T00:00:00.000Z", "2002-01-01T00:00:00.000Z", "2007-01-01T00:00:00.000Z", "2012-01-01T00:00:00.000Z", "2017-01-01T00:00:00.000Z", "2022-01-01T00:00:00.000Z"</Dimension>
  <Style xmlns="http://www.opengis.net/wms"><Name>color</Name><Title>color</Title><Abstract>color <![CDATA[<rasdaman><default>true</default></wcp:QueryFragment>image.Stretch($c, 0, 2.481)</wcp:QueryFragment><ColorTable><ColorTableType>ColorMap</ColorTableType
          
```

Figure 17: The fragments of FAIRiCUBE platform to access Index of Terrestrial Connectivity timeseries.

4.3 FUTURE WORK

Within the air quality pilot there is a mixture of both kinds of data. There is both spatially unstructured IoT data coming, for example, from Sensor.Community sensors and raster data such as the output of the CAMS air quality model or spatially interpolated versions of the IoT data. This mixture highlights the challenges of working with datacubes and IoT data together. Existing tools tend to focus on one side or the other which can make it difficult to achieve basic operations such as plotting, comparison and analysis of the two data types. The Kriging interpolation method that we are using solves this in a way by simply transforming all the data into a datacube, however this inevitably loses some context and information about the original data, so it is not a one size fits all solution.



One of the important issues in implementing data cubes, despite being powerful and fast to retrieve multidimensional data, is the incorporation of OGC queries into GUI tools, e.g. open.Dash, which can be solved by JavaScript libraries (Leaflet or Mapbox).

Regarding the Pilot 2 and bearing in mind the important dependency of connectivity computations (especially MiraMon) on time and hardware capabilities, it seems to be more viable to execute connectivity computations not in data cubes, but as an external software packaged as a container and deployed on the HPC cluster (for further details see Deliverable 6.1, section 6.7). Even though data preprocessing for Pilot 2 does not heavily depend on time and computational resource, it cannot be implemented on FAIRiCUBE, at least the entire preprocessing workflow, as preprocessing is mostly focused on retrieval, filtering of vector data and its rasterisation.



5 CONCLUSIONS AND RECOMMENDATIONS

During the first half of this project, significant progress in integrating socio-economic data, IoT data, low-cost sensor data, and multidimensional environmental data into the Green Deal Data Space has been made. For the first ones, a comprehensive mapping of relevant socio-economic datasets and resources available for integration in the GDDS has been performed, as well as a basic demonstration of its possible impact particularized for pilot 1. It is planned that more data integrations and improvements will be done in the second part of the project.

The integration of IoT & low-cost sensors was discussed and described for each pilot, identifying for pilot 1 the water level and temperature retrieved from the Wasserportal, the usage of camera traps in the context of pilot 2, and for pilot 3 datasets from Citizen-Science initiatives (i.e. Sensor.Community and PurpleAir). The deployment of three sensors for determining the oxygen concentration were also procured as part of the project for the water quality pilot. The tools to incorporate all these data in being developed, including a model for the integration of heterogeneous IoT data to SensorThings API (STA) server with STAplus plugin to make it aligned with the FAIR principles, and IonBeam which is responsible of building scalable data processing pipelines out of modular and reusable software components.

The usage of multidimensional data, and that of data cubes in particular, was tested in pilot 2, where two data cube technologies are being tested: Open Data Cube and FAIRiCUBE Rasdaman. In case of pilot 3, data from CAMS regional model analysis and the ECMWF fifth reanalysis is incorporated, where both datasets are provided via the Copernicus Climate Data Store and the Copernicus Atmosphere Data Store. In this matter, Kriging interpolation is in development, i.e. a geostatistical method that allows to optimally interpolate observations from arbitrary spatio-temporal coordinates to a structure grid, essentially a data cube. In parallel to this, several standards are being analysed to get the best option to retrieve multidimensional data in data cubes in terms of feasibility, usability and standardisation.



6 REFERENCES

- [1] AD4DG Project 101061001, "Deliverable D6.1 Pilot Technical implementation planning, implementation and assessment," 2024.
- [2] Kotsev, A., Minghini, M., Cetl, V., Penninga, F., Robbrecht, J., & Lutz, M, "INSPIRE—A Public Sector Contribution to the European Green Deal Data Space. A vision for the technological evolution of Europe's Spatial Data Infrastructures for 2030," 2021.
- [3] AD4DG Project 101061001, "Deliverable D3.1 Heterogeneous IoT Data Integration Model," 2024.
- [4] Schramm, Matthias et al., "The openEO API—Harmonising the Use of Earth Observation Cloud Services Using Virtual Data Cube Functionalities," *Remote Sensing*. 13 1125. 10.3390/rs13061125, 2021.
- [5] K. Ivanov, "Urban air pollution regression-kriging mapping of the city of Sofia. Vol 2, pages 114-120.," 2020.

ANNEX I - RELEVANT IDENTIFIED SOURCES OF SOCIO-ECONOMIC DATA

Data Hosts & Owners

Data Hosts & Owners	Categories (Catalogues)	Databases/Datasets	URL/Domain	API	Accessibility - Limitations	Data Privacy	IPR
Eurostat	General and regional statistics	https://ec.europa.eu/eurostat/web/main/data/database#Data%20navigation%20tree	https://ec.europa.eu/eurostat/web/main/home	- SDMX RESTful API (official sdmx-rest 2.1 specifications official sdmx-rest 3.0 specifications)	https://ec.europa.eu/eurostat/web/main/help/accessibility	This site is managed by the European Commission, Directorate-General for Communication. More information: https://european-union.europa.eu/privacy-policy-european-union-website_en	Eurostat has a policy of encouraging free re-use of its data, both for non-commercial and commercial purposes.
	Economy and finance			- Eurostat statistics web services (replacing JSON/Unicode web services)			
	Population and social conditions			- RSS (in the context of the API, RSS provides the users with a very simple way to be automatically informed about the last changes carried out to data products and code lists published by Eurostat)			
	Industry, trade and services			More information: https://wikis.ec.europa.eu/display/EUROSTATHELP/API+for+data+access			
	Agriculture and fisheries						
	External trade						
	Environment and energy						
	Transport						
	Health						
	Education						
Employment							
Science and technology							

Data Hosts & Owners	Categories (Catalogues)	Databases/Datasets	URL/Domain	API	Accessibility - Limitations	Data Privacy	IPR
EU Open Data Portal	Economy and Finance Agriculture, fisheries, forestry and food Environment Energy Public Sector Health Population and Society Regions and cities Transport	1.576.392 Datasets	data.europa.eu	<p>The REST API is one of the ways to access the EU Open Data Portal. The application used by the EU Open Data Portal for the API is CKAN.</p> <p>DCAT Application Profile for Data Portals in Europe (DCAT-AP). All metadata of data.europa.eu is stored as RDF triples (RDF) and can be queried using SPARQL query language at this endpoint. Datasets and catalogues published as RDF triples on the portal follow the DCAT-AP specification.</p> <p>data.europa.eu provides the following APIs to read our metadata:</p> <p>Search: https://data.europa.eu/api/hub/search/</p> <p>SPARQL: https://data.europa.eu/sparql</p> <p>Registry: https://data.europa.eu/api/hub/repo/</p> <p>Use cases: https://data.europa.eu/en/export-use-cases</p> <p>MQA: https://data.europa.eu/api/mqa/cache/</p> <p>SHACL metadata validation: https://data.europa.eu/api/mqa/shacl/</p> <p>More information: https://data.europa.eu/api/provider-manual/api-documentation/</p>	https://data.europa.eu/gitlab.io/data-provider-manual/accessibility/	<p>This site is managed by the European Commission, Directorate-General for Communication.</p> <p>More information: https://european-union.europa.eu/privacy-policy-european-union-website_en</p>	<p>The Commission's reuse policy is implemented by Commission Decision 2011/833/EU of 12 December 2011 on the reuse of Commission documents. Unless otherwise noted (e.g. in individual copyright notices), the reuse of the editorial content on this website owned by the EU is authorised under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence. This means that reuse is allowed, provided appropriate credit is given and any changes are indicated.</p> <p>Reproduction is authorised, provided the source is acknowledged, save where otherwise stated.</p> <p>Where prior permission must be obtained for the reproduction or use of textual and multimedia information (sound, images, software, etc.), such permission shall cancel the above-mentioned general permission and shall clearly indicate any restrictions on use.</p> <p>More information: https://data.europa.eu/en/copyright-notice</p>

Data Hosts & Owners	Categories (Catalogues)	Databases/Datasets	URL/Domain	API	Accessibility - Limitations	Data Privacy	IPR
OECD Data	Development Economy Education Environment Energy Finance Health Innovation & Technology Society	2093 Databases Other sites/resources: OECD Better Life Index OECD iLibrary OECD Observer OECD Insights blog OECD Development Centre FATF - Financial Action Task Force IEA - International Energy Agency ITF - International Transport Forum NEA - Nuclear Energy Agency	https://data.oecd.org/	https://data.oecd.org/api/	The use of OECD datasets and the API is governed and limited by specific Terms and Conditions (https://www.oecd.org/termsandconditions/)	<p>The OECD is committed to protecting the personal data of users of its websites, which are an essential tool for advancing our mission and programme of work.</p> <p>This privacy policy describes the types of personal data we collect, how we use it, who has access and for how long. It also provides information on the rights provided to website users under our data protection rules. Some activities supported by our websites are further subject to specific data protection notices (e.g. recruitment).</p> <p style="text-align: right;">https://www.oecd.org/privacy/</p>	<p>The OECD encourages the use of its data, publications and multimedia products (sound, image, software, etc.), collectively, the "Material". Unless otherwise stated, the Material is the intellectual property of the OECD and protected by copyright or other similar rights.</p> <p style="text-align: center;">More information: https://www.oecd.org/termsandconditions/</p>

Data Hosts & Owners	Categories (Catalogues)	Databases/Datasets	URL/Domain	API	Accessibility - Limitations	Data Privacy	IPR
OECD Data Explorer (New)	Development Economy Education Environment Energy Finance Health Innovation & Technology Society	NOTE: Not all data is available on this platform yet, as it is being progressively migrated from OECD.Stat.	https://data-explorer.oecd.org/	https://data.oecd.org/api/	The use of OECD datasets and the API is governed and limited by specific Terms and Conditions (https://www.oecd.org/termsandconditions/)	https://www.oecd.org/privacy/	The OECD encourages the use of its data, publications and multimedia products (sound, image, software, etc.), collectively, the "Material". Unless otherwise stated, the Material is the intellectual property of the OECD and protected by copyright or other similar rights. More information: https://www.oecd.org/termsandconditions/
World Bank Open Data	Energy Earth Observation for Sustainable Development Finance Geospatial Road software tools Transport Digital Development	More Resources: Open Data Catalog DataBank Microdata Library Global Data Facility World Development Indicators Living Standards Measurement Study Global Consumption Database	https://data.worldbank.org/	This is the interface of choice for creating custom data visualisations, live combinations with other data sources (mashups), and more. data.worldbank.org is built using data through the Data API. More information: https://datahelpdesk.worldbank.org/knowledgebase/topics/125589	https://www.worldbank.org/en/access-to-information/overview#1 https://www.worldbank.org/en/access-to-information Note: The Open Data Catalog lists all data adhering to the Open Data terms of use. https://data.worldbank.org/restricted-data	https://www.worldbank.org/en/about/legal/privacy-notice	Unless indicated otherwise in the data or indicator metadata, you are free to copy, distribute, adapt, display or include the data in other products for commercial or noncommercial purposes at no cost under a Creative Commons Attribution 4.0 International License, with the additional terms below. https://data.worldbank.org/summary-terms-of-use Legal: https://www.worldbank.org/en/about/legal

Data Hosts & Owners	Categories (Catalogues)	Databases/Datas ets	URL/Domain	API	Accessibility - Limitations	Data Privacy	IPR
THE WORLD BANK Data Catalogue	<p><u>Contributing Catalogues:</u></p> <p>Microdata Library</p> <p>EnergyData.info</p> <p>Open Finances Library Network</p> <p>The Data Catalog is designed to make World Bank's development data easy to find, download, use, and share. It includes data from the World Bank's microdata (https://microdata.worldbank.org/index.php/home), finances (https://finances.worldbank.org/) and energy data platforms (https://energydata.info/), as well as datasets from the open data catalogue (aforementioned).</p>	<p>6913</p> <p>Note: The Open Data Catalog lists all data adhering to the Open Data terms of use. This catalogue lists all Open Data and other publicly available data disseminated through World Bank sites, but may have some restrictions on use.</p>	<p>https://datacatalog.worldbank.org/home</p>	<p>This is the interface of choice for creating custom data visualisations, live combinations with other data sources (mashups), and more. data.worldbank.org is built using data through the Data API.</p> <p>More information: https://datahelpdesk.worldbank.org/knowledgebase/topics/125590</p>	<p>https://www.worldbank.org/en/access-to-information/overview#1</p> <p>The World Bank Group makes data available using a combination of licences based on the Access to Information policy. As much as possible, we distribute data according to open data standards and licensed under the Creative Commons Attribution license (CC-BY 4.0). Many datasets are also available under other licences.</p> <p>For more details: https://datacatalog.worldbank.org/public-licenses</p>	<p>https://www.worldbank.org/en/about/legal/privacy-notice</p>	<p>Unless indicated otherwise in the data or indicator metadata, you are free to copy, distribute, adapt, display or include the data in other products for commercial or noncommercial purposes at no cost under a Creative Commons Attribution 4.0 International License, with the additional terms below.</p> <p>https://data.worldbank.org/summary-terms-of-use</p> <p>Legal: https://www.worldbank.org/en/about/legal</p>

Data Hosts & Owners	Categories (Catalogues)	Databases/Datasets	URL/Domain	API	Accessibility - Limitations	Data Privacy	IPR
EnergyData.info	Energy Sector Environment Sustainability	979	https://energydata.info/	ENERGYDATA.INFO is powered by the open-source data portal platform CKAN, and it is developed publicly on GitHub. More information on CKAN and GitHub can be accessed at http://ckan.org/ and https://github.com/ https://datahelpdesk.worldbank.org/knowledgebase/topics/125589	ENERGYDATA.INFO has been developed as a public good available to governments, development organisations, private sector, non-governmental organisations, academia, civil society and individuals to share data and analytics that can help achieving the United Nations’ Sustainable Development Goal 7 of ensuring access to affordable, reliable, sustainable and modern energy for all. ENERGYDATA.INFO is designed according to open data and open source standards and principles. Access to data and information available on ENERGYDATA.INFO is free, subject to the terms of this agreement. Terms of Use You are encouraged to use the content in ENERGYDATA.INFO to benefit yourself and others in creative ways. The majority of content in ENERGYDATA.INFO is licensed under the Creative Commons Attribution licence (CC BY 4.0). Works in ENERGYDATA.INFO that are not subject to the CC BY licence are clearly labelled with the specific licence terms under which those works may be used. By using ENERGYDATA.INFO, you agree to be bound by these Terms of Use for ENERGYDATA.INFO. More information: https://energydata.info/terms	https://www.worldbank.org/en/about/legal/privacy-notice	Intellectual Property Rights – Licences You are encouraged to use the content in energydata.info (the “Works”) to benefit yourself and others in creative ways. The World Bank Group owns the copyright to the Works unless specifically noted to the contrary. The majority of the Works are licensed under the Creative Commons Attribution licence (CC BY 3.0 IGO). Works that are not subject to the CC BY licence are clearly labelled with the specific licence under which those works may be used. By using energydata.info, you agree to be bound by these Terms of Use for energydata.info. https://energydata.info/terms



Data Hosts & Owners	Categories (Catalogues)	Databases/Datasets	URL/Domain	API	Accessibility - Limitations	Data Privacy	IPR
ECB Eurosystem (ECB Data Portal)	Economy Finance Monetary Policies Environmental Protection Climate Inflation	https://data.ecb.europa.eu/data/datasets	https://data.ecb.europa.eu	The ECB SDMX 2.1 RESTful web service offers programmatic access to the statistical data and metadata disseminated via the ECB Data Portal. More information: https://data.ecb.europa.eu/help/api/overview	https://data.ecb.europa.eu/help/account/how-can-i-use-ecb-data-portal-account Accessibility: https://www.ecb.europa.eu/stats/ecb_statistics/accessing-our-data/html/index.el.html	The ECB is obliged to comply with the Regulation (EU) 2018/1725 (EUDPR) when it processes personal data. More information: https://www.ecb.europa.eu/services/data-protection/html/index.en.html	Free use of the information obtained directly from ECB, subject though to specific conditions. More information: https://www.ecb.europa.eu/services/using-our-site/disclaimer/html/index.en.html#cx
United Nations Economic Commission for Europe (UNECE) Data Portal	SDGs Transport (Traffic Safety etc.) Economy	https://w3.unece.org/PXWeb/en	https://w3.unece.org/PXWeb/en	YouTube API Services to manage videos	This database is maintained by the Statistical Division of the UNECE Secretariat. It provides free access to data structured to facilitate easy retrieval of statistics by users. It is organised by subject or policy areas. Multi-dimensional tables present data by country, by various socio-economic classifications related to the context, and by time period. More information: https://unece.org/terms-and-conditions-use-united-nations-websites	https://unece.org/privacy-notice	The data in the UNECE Statistical Database are available free of charge. Users are free to copy, reproduce and redistribute the data for both commercial and non-commercial purposes, provided UNECE is acknowledged as the source. UNECE does not take any responsibility for the use of the data.



Data Hosts & Owners	Categories (Catalogues)	Databases/Datas ets	URL/Domain	API	Accessibility - Limitations	Data Privacy	IPR
European Environment Agency (Charts, Maps, Indicators, Datahub)	Air pollution Biodiversity Climate Environmental Health Impacts Health Sustainability Transport	229	https://www.eea.europa.eu/en/datahub	https://www.eea.europa.eu/code/api	https://www.eea.europa.eu/en/accessibility?size=n_10_n&filters%5B0%5D%5Bfield%5D=is_sued.date&filters%5B0%5D%5Btype%5D=any&filters%5B0%5D%5Bvalues%5D%5B0%5D=Al%20time	https://www.eea.europa.eu/en/privacy	https://www.eea.europa.eu/en/legal-notice Unless otherwise indicated, the European Environment Agency (EEA) is the owner of copyrights and database rights in this website and its contents, and EEA materials are published under the CC-BY licence. The sources and owner(s) of the content are clearly indicated for each content.
International Monetary Fund (IMF) Data	Economy Finance Climate change	https://data.imf.org/?sk=388DFA60-1D26-4ADE-B505-A05A558D9A42&slid=1479329132316	https://www.imf.org/en/Data	Developers can use Data Services to make applications with the ability to import data from the repository databases in the SDMX formats 2.0 (as ASP.NET, WCF, or RESTFUL Services), SDMX 2.1 and JSON Restful. The specifications can be accessed through the API tab located at the top of each dataset. The IMF DataMapper API can be used to retrieve time series as used in the DataMapper. Endpoints are available to get lists of indicators, countries, regions and analytical groups used in these time series. The current version of the API is v1. https://www.imf.org/external/datamapper/api/help	https://datahelp.imf.org/knowledgebase/topics/71762-accounts-and-access	https://www.imf.org/external/privacy.htm	The International Monetary Fund (the "IMF") maintains websites and mobile apps (the "Sites") to provide direct access to its information, documents, data and materials (collectively, "Content") to those who choose to access and use the Sites ("You" or "Users"). The IMF maintains the Sites for informational purposes only. https://www.imf.org/external/terms.htm

Data Hosts & Owners	Categories (Catalogues)	Databases/Datasets	URL/Domain	API	Accessibility - Limitations	Data Privacy	IPR
International Energy Agency IEA	Energy Environment Climate	63	https://www.iea.org/data-and-statistics		<p>Access to data via account</p> <p>Access your data online through our Web Data Service (WDS) platform. Many IEA data services are available online through our Web Data Service (WDS) platform. To access the platform, go to your products page, accessible from your User Profile (icon at the top right of the main menu).</p> <p>More information: https://www.iea.org/help-centre/accessing-iea-products-and-services</p> <p>IEA Open Use Terms</p> <p>The IEA makes much of its content available under open Creative Commons licences. Unless one of the below exceptions to the Open Use Terms applies, all text content, reports, articles, commentaries, standalone graphs, figures and infographics produced by and/or sourced to the IEA and hosted on the IEA Websites are licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence.</p> <p>https://www.iea.org/terms</p>	https://www.iea.org/privacy	https://www.iea.org/terms/rights-requests



<p style="text-align: center;">Consortium of European Social Science Data Archives (CESSDA) DC Data Catalogue</p>	<p>This is the CESSDA Data Catalogue (CDC) version 3.5.0 released on 2024-01-30. It contains descriptions of the more than 40,000 data collections held by CESSDA's Service Providers (SPs), originating from over 20 European countries. The CDC is a one-stop shop for searching and discovering European social science data.</p>	<p style="text-align: center;">41198</p>	<p style="text-align: center;">https://datacatalogue.cessda.eu/</p>	<p style="text-align: center;">REST API https://api.tech.cessda.eu/</p>	<p>The CDC is a portal for discovering data and detailed dataset descriptions are provided. For information and procedures on how to access data, there is a link ['Access data'] from each study to the study information on the website of the data provider (Publisher). Refer to the information found there to see which access conditions apply to data created in the particular study.</p> <p>There is also an OAI-PMH compliant endpoint, so that data aggregators can easily harvest the entire contents of the data catalogue. Record identifiers are unified with those used in the User Interface. It is located at https://datacatalogue.cessda.eu/oai-pmh/v0/oai.</p> <p>https://datacatalogue.cessda.eu/accessibility-statement/</p>	<p style="text-align: center;">https://www.cessda.eu/Privacy-Policy</p>	<p style="text-align: center;">Article 16 of the STATUTES of CESSDA ERIC</p> <p>Intellectual Property</p> <ol style="list-style-type: none"> 1. The term "intellectual property" shall in these Statutes be understood in accordance with Article 2 of the Convention Establishing the World Intellectual Property Organisation (WIPO) signed on 14 July 1967. 2. With respect to questions of intellectual property, the relations between Members, Observers and Service Providers shall be governed by applicable national as well as relevant international rules and regulations. 3. Intellectual property that Members or Service Providers contribute to CESSDA ERIC shall remain the property of the intellectual property holder. 4. If the intellectual property originates from CESSDA ERIC-funded work (direct contribution or in kind), such property shall belong to CESSDA ERIC. CESSDA ERIC may relinquish its rights fully or partially in favour of the Member, Observer or Service Provider that has created the intellectual property rights.
--	--	--	--	---	--	--	--

Data Hosts & Owners	Categories (Catalogues)	Databases/Datas ets	URL/Domain	API	Accessibility - Limitations	Data Privacy	IPR
Climate Trace	Environment Energy Climate change Pollution	https://climatetrace.org/data	https://climatetrace.org/	https://api.climatetrace.org/v4/swagger/index.html	Climate TRACE emissions data is free and publicly available for download and via API. https://climatetrace.org/data	https://climatetrace.org/privacy	https://climatetrace.org/terms
Global Biodiversity Information Facility (GBIF)	Environment Biodiversity Climate change Health	92,716	https://www.gbif.org/	https://techdocs.gbif.org/en/openapi/	To the greatest extent possible, GBIF is an open-access facility. All users, whether GBIF Participants or others, should have equal access to data in databases affiliated with or developed by GBIF. https://www.gbif.org/terms/data-user	https://www.gbif.org/terms/privacy-policy	https://www.gbif.org/terms
Intergovernmental Panel on Climate Change Data Distribution Centre (IPCC DDC)	Socio-Economic Data: Population and human development, Economic Conditions, Land cover/land use, Water, Agriculture/food, Energy, Biodiversity. Environmental Data	Old site: https://www-devel.ipcc-data.org/guidelines/pages/approvedDatasetLinks.html	https://www.ipcc-data.org/		https://www.ipcc-data.org/guidance.html	https://www.ipcc-data.org/privacy-policy.html	https://www.ipcc-data.org/copyright.html

<p style="text-align: center;">Climate Policy Database</p>	<p>Energy Environment Transport</p>	<p>https://climatepolicydatabase.org/policies</p>	<p>https://climatepolicydatabase.org/</p>	<p>The Climate Policy Database is updated periodically. The latest version of the database can be downloaded here or accessed through a Python API.</p>	<p style="text-align: center;">publicly-accessible portal</p> <p style="text-align: center;">More information: https://newclimate.org/about-us/legal-notice</p>	<p>Data protection</p> <p>Insofar as the possibility of entering personal or business data (email addresses, names, addresses etc.) exists in the Internet offer the disclosure of this data by the user expressly takes place on a voluntary basis. The use of all offered services are permitted – if and so far technically possible and reasonable – without specification of any personal data or under specification of anonymized data or an alias. The use of the contact data published in the framework of the imprint or similar information like mailing addresses, telephone and fax numbers as well as addresses of third parties for marketing purposes of not expressly requested information is not allowed. Legal steps against the senders of spam</p>	<p>Copyright</p> <p>Copyright 2014-2024 NewClimate – Institute for Climate Policy and Global Sustainability gGmbH. All rights reserved. All content (text, pictures, graphics, audio-, video- and animation files, their layout etc.) on the website newclimate.org are subject to the copyright protection and other protection laws. This legal protection also applies to databases and similar facilities. This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License for non commercial use. Any commercial reproduction, in full or in part, is prohibited without the prior written consent of climatepolicy.net. Parts of the website furthermore contain images that are protected by the copyright of third parties. Unless otherwise specified, all trademarks on the website are protected by copyright. The sole mentioning of a trade mark on this website should not lead to the assumption that it is not protected by the rights of a third party. The copyright for published material created by</p>
						<p>Insomuch as the possibility of entering personal or business data (email addresses, names, addresses etc.) exists in the Internet offer the disclosure of this data by the user expressly takes place on a voluntary basis. The use of all offered services are permitted – if and so far technically possible and reasonable – without specification of any personal data or under specification of anonymized data or an alias. The use of the contact data published in the framework of the imprint or similar information like mailing addresses, telephone and fax numbers as well as addresses of third parties for marketing purposes of not expressly requested information is not allowed. Legal steps against the senders of spam</p>	<p>Copyright 2014-2024 NewClimate – Institute for Climate Policy and Global Sustainability gGmbH. All rights reserved. All content (text, pictures, graphics, audio-, video- and animation files, their layout etc.) on the website newclimate.org are subject to the copyright protection and other protection laws. This legal protection also applies to databases and similar facilities. This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License for non commercial use. Any commercial reproduction, in full or in part, is prohibited without the prior written consent of climatepolicy.net. Parts of the website furthermore contain images that are protected by the copyright of third parties. Unless otherwise specified, all trademarks on the website are protected by copyright. The sole mentioning of a trade mark on this website should not lead to the assumption that it is not protected by the rights of a third party. The copyright for published material created by</p>

Data Hosts & Owners	Categories (Catalogues)	Databases/Datasets	URL/Domain	API	Accessibility - Limitations	Data Privacy	IPR
						<p>mails in case of violation of this prohibition are expressly reserved.</p> <p>More information: https://newclimate.org/about-us/legal-notice</p>	<p>climatepolicy.net remains the property of the author of the pages.</p> <p>More information: https://newclimate.org/about-us/legal-notice</p>

Data Hosts & Owners	Categories (Catalogues)	Databases/Datas ets	URL/Domain	API	Accessibility - Limitations	Data Privacy	IPR
<p style="text-align: center;">UN Department of Economic and Social Affairs Statistics - SDG Indicators Database</p>	<p>Socio-Economic Data</p>	<p>https://unstats.un.org/sdgs/dataportal/database</p>	<p>https://unstats.un.org/sdgs/dataportal/</p>	<p>https://unstats.un.org/sdgs/iaeg-sdgs/sdmx-working-group/</p> <p style="text-align: center;">and</p> <p>https://unstats.un.org/sdgs/UNSDGAPIV5/swagger/index.html</p>	<p>The United Nations reserves the right to deny in its sole discretion any user access to this Site or any portion thereof without notice.</p>	<p>By accessing this site, certain information about the User, such as Internet protocol (IP) addresses, navigation through the Site, the software used and the time spent, along with other similar information, will be stored on United Nations servers. These will not specifically identify the User. The information will be used internally only for web site traffic analysis. If the User provides unique identifying information, such as name, address and other information on forms stored on this Site, such information will be used only for statistical purposes and will not be published for general access. The United Nations, however, assumes no responsibility for the security of this information.</p>	<p>None of the materials provided on this web site may be used, reproduced or transmitted, in whole or in part, in any form or by any means, electronic or mechanical, including photocopying, recording or the use of any information storage and retrieval system, except as provided for in the Terms and Conditions of Use of United Nations Web Sites, without permission in writing from the publisher.</p> <p>https://www.un.org/en/about-us/copyright</p>



Data Hosts & Owners	Categories (Catalogues)	Databases/Datasets	URL/Domain	API	Accessibility - Limitations	Data Privacy	IPR
World Resources Institute (WRI)	Climate Energy Environment Economy	135 datasets https://datasets.wri.org/dataset More resources https://www.wri.org/data/data-products	https://www.wri.org/data	<p>WRI maintains two distinct APIs that power several web applications and research efforts.</p> <p>Application Programming Interfaces (API)</p> <ul style="list-style-type: none"> Resource Watch API (RW API) provides a set of common back-end services used to power most applications maintained by WRI. These applications use the RW API for user management (which allow users to keep the same login credentials cross multiple products), front-end configuration, and some tabular data storage. The RW API makes use of a “microservice” infrastructure that allows users to develop custom logic to interact with the API’s services. <p>Website: api.resourcewatch.org</p> <p>GitHub repository: github.com/resource-watch</p> <ul style="list-style-type: none"> GFW Data API provides access to GFW-maintained data and analysis services and powers all the applications on the GFW platform. In general terms, it is composed of four main pieces: dataset metadata, tile caches, analysis services, and downloads. <p>Website: data-api.globalforestwatch.org</p> <p>GitHub repository: github.com/wri/gfw-data-api</p>	<p>Data should be licensed openly for free, allowing anyone to use, share, and adapt our work.</p> <p>Data should be easily accessible and downloadable, thoroughly described, machine-readable, and maintained over time, to enable reuse.</p> <p>Data should be complete and primary, such that others are able to test and examine our work.</p> <p>More information: https://www.wri.org/data/open-data-commitment</p>	https://www.wri.org/about/privacy-policy	https://www.wri.org/research/permissions-licensing

