# Plan Overview

*A Data Management Plan created using DMPonline*

**Title:** Co-designed Citizen Observatories Services for the EOS-Cloud

**Creator:**Fernando Aguilar

**Principal Investigator:** Jaume Piera

**Data Manager:** Lara Lloret, Fernando Aguilar

**Affiliation:** Other

**Funder:** European Commission

**Template:** Horizon 2020 Template

**ORCID iD:** 0000-0001-5818-9836

**Project abstract:**

The EU-funded COS4CLOUD project aims to facilitate open science and citizen science initiatives by designing and implementing services. The project will design and prototype these new services using deep machine learning, automatic video recognition, and other cutting-edge technologies. COS4CLOUD hopes to make it easier for citizen science platforms to share data using improved networks in a user-friendly way. The project will use the experiences of platforms like: Artportalen, Natusfera, iSpot, as well as other environmental quality monitoring platforms like FreshWater Watch, KdUINO, OdourCollect, iSpex and CanAir.io. The project will integrate citizen science in the European Open Science Cloud to service the scientific community and society at large. This report summarises the work of WP1 on the Data Management considerations and plan for the COS4CLOUD project. This document describes the types of data that will be generated or collected during the project, the standards that will be used and the ways in which the data may be exploited and shared including the data security and ethical aspects.

**ID:** 75548

**Start date:** 01-11-2019

**End date:** 28-02-2023

**Last modified:** 01-03-2023

**Grant number / URL:** 863463

the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

# Co-designed Citizen Observatories Services for the EOS-Cloud - Initial DMP

## 1. Data summary

**Provide a summary of the data addressing the following issues:**

- **State the purpose of the data collection/generation**
- **Explain the relation to the objectives of the project**
- **Specify the types and formats of data generated/collected**
- **Specify if existing data is being re-used (if any)**
- **Specify the origin of the data**
- **State the expected size of the data (if known)**
- **Outline the data utility: to whom will it be useful**

The data managed within the context of the project is related to the following project objectives:
**O1**. Integrate Citizen Science in the European Open Science landscape through the development of a Minimum Viable Ecosystem for Citizen Science Observatories integrated to the EOSC.
Citizen Science projects generally use applications to enhance the collaboration of individuals capable of generating new data from different disciplines. The objectives is to facilitate the use of this data by scientist and any other stakeholders.

**O3**. Increase the **quantity and quality of the data available from citizen science under the FAIR** data principles (findable, accessible, interoperable & reusable) and extend them with added principles.

The way in which the data is gathered is different along the different Citizen Science Observatories. Cos4Cloud aims to improve the quality of the data adopting best practices on data management, and apply the FAIR principles for the data produced.

The different Citizen Science Observatories that are connected with specific **platforms** like mobile apps have been designed at architectural level differently so that the data models, formats, granularity, etc. may differ. The above mentioned Cos4Bio and Cos4Env aim at integrating the data from different sources using a common layer, which can be difficult. As stated earlier, the project will not produce any data itself, but some derived data can be produced, so this Data Management Plan does not refer to the data produced by the platforms, but due to the heterogeneity of these data, it is interesting to describe the information produced by them. In any case, the data will try to enhance the FAIRness of the data produced by the Citizen Science Observatories to be connected with the European Open Science Cloud, and the derived data will be treated as FAIR and published in the proper repositories.

The following list provides some details about the different platforms in terms of formats and type of gathered information. However, due to the need of citizen involvement, the volume of data to be generated is difficult to calculate.

The observatories

This subsection briefly introduces the different observatories emphasizing in the data part. A detailed version of the platforms and the strategic plan for the exploitation and dissemination of the results can be found in D7.3.

**Artportalen (https://www.artportalen.se/)**

Artportalen is a Citizen Science Observatory to report biodiversity observations in Sweden. Observations consist of four obligatory fields: taxa, location, date and reporter. Additional information

can be uploaded, for example activity (breeding, migrating etc), observation method, determination method or habitat.

Artportalen has received more than 83,000,000 (as of April 2021) observations of birds, plants, insects, fungi and many other taxa along with the 1,300,000 associated media files. Nearly 4,000,000 of the observations that Artportalen has received have been validated by expert validators or committees. Media files can be also connected to textual metadata.

**Natusfera (https://natusfera.gbif.es/)**

Natusfera is a mobile application as well as a web platform to produce biodiversity data, which is validated by experts. It is also planning to incorporate environmental data.

Biodiversity data usually includes media files (pictures in different formats jpg, png, ... and in some cases audio records, mostly in mp3). Thanks to the mobile phones, the media files can be georeferenced, and the information can be sorted in CSV and Darwin core standard.

Some of the data validated within the context of Natusfera is published on the GBIF network.

i**Spot (https://www.ispotnature.org/)**

iSpot is a Citizen Science Observatory on biodiversity. The platform encompasses a network of over 68,000 global nature observers who have crowdsourced the identification of 30,000 taxa, through over 1,500,000 images of more than 750,000 observations of different species (Birds, Amphibians and Reptiles, Fish Fungi and Lichens,etc.). It may include media data like images, as well as georeference and other metadata values.

**Pl@ntNet (https://plantnet.org/)**

Pl@ntNet is a tool (web + mobile app) to help to identify plants with pictures. It is organized in different thematic and geographical floras. Users choose their region or area of interest from a list and select "World Flora" if their region is not available.

Images are in jpg format. Metadata and user accounts are in semi-structured format (json). Taxonomic repositories are based on the International Code of Botanical Nomenclature (Shenzen code)

Pl@ntNet data currently includes:

(i) about 300M plant observations, each observation being composed of one or more images and some metadata such as date, species names (collected and generated), organ tags, geo-location (for about 50% of them), author username (for about 25% of them), image quality ratings.

(ii) 30 botanical taxonomic repositories totalling several hundred thousand species, each associated with their scientific name, synonyms, common names, URLs to external resources and (generated) statistics

(iii) a database of nearly 2,3M user accounts, each associated with username, email, avatar (optional) and statistics (generated)

The number of users and observations is expected to double in 1-2 years.

Regarding data publication:

1. A part of Pl@ntNet's observations is publicly visible in Pl@ntNet applications (about 10M observations). This part corresponds to the observations for which the users explicitly agreed to make them public with their author name (under a cc-by-sa licence).
2. A part of Pl@ntNet's observations is shared through GBIF. About 800K observations are shared with author names and photographs (under a cc-by-sa licence). About 10M of them are shared without the images and the author names (only the species name and location is shared under cc0 licence).
3. Some subsets of Pl@ntNet data built for researchers are publicly available on various platforms (zenodo, GitLab, kaggle, etc.).

**FreshWater Watch (https://freshwaterwatch.thewaterhub.org/)**

FreshWater Watch (FWW) is a global citizen-science project, started in 2012, investigating the health of the world’s freshwater ecosystems. The main parameters measured are nitrates, phosphates, bank vegetation, wildlife presence, pollution sources, water level, water colour, presence of algae, and

turbidity.  FWW data are not managed within the context of the project.

**KdUINO (https://monocle-h2020.eu/Sensors_and_services/KdUINO)**

KdUINO is a low-cost open-source monitoring system to measure water transparency. Citizens build their own buoy with sensors and put it in the sea. It is possible to leave the KdUINO in the water for a long time. The buoy collects data on transparency, measured using the sensors on the KdUINO. It thus gives continuous transparency measurements in real time and provides coverage for a large coastal zone, something not possible using traditional radiometers due to their cost.

It is currently being upgraded to gather information on different colour bands (RGB). A do-it-yourself (DIY) version is being developed that will have better usability, as well as being lighter and more portable, under the MONOCLE framework.

**OdourCollect (https://odourcollect.eu/)**

Odour pollution is the second most common reason for environmental complaints in the world, after noise, and it can be a sign of greater environmental problems. OdourCollect is a free app that aims to tackle odour pollution by empowering affected citizens to build collaborative odour maps. The app promotes a driving force of change, encouraging dialogue among citizens, local authorities, industries and experts.

Any citizen can act as an observer and report georeferenced observations on the odour episode, which are open data and can be used to build collaborative odour complaint maps and identify potential odour emitting sources.

Odour observations can be validated by experts to gather data in a particular area where a community is affected by odour pollution and with the aim of co-designing local solutions with relevant stakeholders.

**iSpex (http://ispex.nl/en/)**

iSpex is an innovative way to measure aerosols and water colour based on a mobile app and a small optical add-on containing a spectrometer and a polarizer. This instrument measures properties of small particles in the sky: aerosols. It measures PM 2.5 values and water colour. The idea is based on the Spectropolarimeter for Planetary Exploration (SPEX), sized down to allow as many people as possible to use the instrument.

The app and add-on are currently in development, and they will be fully operational in 2021. Additionally, the DDQ team is upgrading the add-on and sensor capabilities to monitor air and water quality properties.

**CanAir.io (https://canair.io/)**

CanAirIO is a Colombian Citizen Science Observatory to monitor air quality with mobile and fixed sensors for measuring air quality (Particle Material PM2.5) with mobile phones (mobile measurement) or Wi-Fi (fixed measurements) with low-cost technology and open source code.

This observatory aims to build a citizen network, an air-quality map that will allow us to know what we are breathing and how we can improve quality of life. With the data collected, citizens can independently validate official air-quality numbers: what can be measured can be improved. This knowledge empowers citizens to demand better air quality policies from governments.

The main purpose of the project is to collect air quality data and publish it, for activists, academics, and people in general. The lifecycle of data starts in the sensors of the community, then they can choose whether these data will be mobile or fixed (the sensors have these modes), and whether these data will be shared in one of the servers: one server (InlfuxDB) for fixed stations, and a second server (Firebase) for mobile stations. The data then can be accessed via Grafa (CSV) or InfluxDB API. Regarding the volume, generated data can be exported in JSON format and ~ 2 GB are produced per year.

Potential interest of the data

Due to the heterogeneity of the data and its type (biodiversity, environmental), they might be useful for different stakeholders at different levels:

- Researchers: The data collected have scientific value with different purposes. In fact, the number of people involved can not be substituted by any automatic systems like sensors or any other

instruments. For example, people collecting images or any other media in the field for georeferenced species is a powerful added value of Citizen Science. Furthermore, for environmental data, the resolution at temporal or spacial level can be increased significantly and efficiently.

- Administrations: the use of this data properly analized by scientists can support administrations like government, river basin authorities or any other public rulers in taking decisions and propose policies to improve the citizens' life quality.
- Commercial - Companies: The volume of information created will be potentially useful for companies to create added value. For example, tourisms actors knowing the best places for richest biodiversity or agriculture companies better understanding the details in a specific place in terms of environment conditions.
- Citizens: apart from the involvement in citizen science activities, the citizens can benefit from the data produced. Thanks to the data produced, the citizen can select places with better air quality or better environment conditions or visit areas with interesting species.

## 2. FAIR data

### 2.1 Making data findable, including provisions for metadata:

- **Outline the discoverability of data (metadata provision)**
- **Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?**
- **Outline naming conventions used**
- **Outline the approach towards search keyword**
- **Outline the approach for clear versioning**
- **Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how**

Each platform will manage their data according to their rules although Cos4Cloud will encourage data publication fulfilling all the FAIR criteria.  Some of the platforms like Natusfera or Pl@ntnet are already using metadata standards to describe their datasets, such as EML, Darwin Core, which are the most common metadata formats to describe both Environmental and Biodiversity data. Persistent identifiers are also already being used in some cases (DOIs).

EML

Ecological Metadata Language (EML) is a metadata specification particularly developed for the ecology discipline. It is based on prior work done by the Ecological Society of America and associated efforts (Michener et al., 1997, Ecological Applications). Sponsored by ecoinformatics.org, EML Version 2.2.0 was released in 2019. Some platforms such as Natusfera are already using this standard being a proper format to describe environmental data

Darwin Core

A body of standards, including a glossary of terms (in other contexts these might be called properties, elements, fields, columns, attributes, or concepts) intended to facilitate the sharing of information about biological diversity by providing reference definitions, examples, and commentaries. Sponsored by Biodiversity Information Standards (TWDG), the current standard was last modified in October 2009. The platforms collecting biodiversity  data are already using this standard.

DOIs:

DOIs are  persistent identifiers or handles used to identify objects uniquely, standardized by the International Organization for Standardization (ISO). For instance @PlantNet, for the part of the data

that is public, is already using a global persistent identifier constructed by concatenating a persistent URL with the internal identifier, e.g.: https://identify.plantnet.org/fr/weurope/observations/1009854020
Targetting at specific publication systems like data repositories or data portals, the derived datasets suitable to be reused will be published with a persistent identifier. Furthermore, the different Citizen Science Observatories will be encouraged to identify their datasets.

**2.2 Making data openly accessible:**

- **Specify which data will be made openly available? If some data is kept closed provide rationale for doing so**
- **Specify how the data will be made available**
- **Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?**
- **Specify where the data and associated metadata, documentation and code are deposited**
- **Specify how access will be provided in case there are any restrictions**

Some derived data produced within the different platforms could be published in the Cos4Cloud test beds environments or externally. As already mentioned this will depend on each platform, since they will decide which data will be openly published. Just as an example of some possible publication scenario, includes target publication platforms like GBIF that are accessible via standard protocols like HTTP. There already exist some APIs to provide machine-actionable features. The derived data produced will try to keep the same mechanisms of publications, targetting at the most effective platforms for each scientific community.
Regarding the software developed within the project,  the documentation will be published as defined in the deliverables D2.2 y D2.7 concerning the initial plan and definition of the software life cycle management process and procedures.

**2.3 Making data interoperable:**

- **Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.**
- **Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?**

Since the platforms are responsible for their data, they should provide the mechanisms to make them interoperable. Communities publishing data derived of the use of the different platforms will be encouraged to use Open Repositories compliant with OAI-PMH  and supporting basic metadata standards as Dublin Core. Also, specific or community-based metadata standards like EML (Environment) or Darwin Core (Biodiversity) will be suggested for being used.

**2.4 Increase data re-use (through clarifying licenses):**

- **Specify how the data will be licenced to permit the widest reuse possible**
- **Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed**

- **Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why**
- **Describe data quality assurance processes**
- **Specify the length of time for which the data will remain re-usable**

Methods for data quality assurance depends also on the platforms. The publication of derived data (datasets, neural network weights, etc...) will be stored using digital resources in an Open Repository where the FAIR principles are applied and its re-use is promoted. If possible, both Platform and derived data will be published in community and interest data portals, like GBIF for Biodiversity, which supports perfectly the FAIR principles. The derived data will adopt Creative Common like licence or any other licence enabling the proper re-use, always respecting the original licence if needed. The collected data stored in the project test beds will keep the data origin licences.

Regarding software, the different Platforms have been developed adopting diverse licences types. Although they will keep their original licence, the project will stimulate the adoption of open licences to enhance the Open Science characteristics of the EOSC. The developments and any other software product created within the context of the project will be available as an Open Source resources under open licences like Creative Commons, Apache or MIT (to be defined).

## 3. Allocation of resources

**Explain the allocation of resources, addressing the following issues:**

- **Estimate the costs for making your data FAIR. Describe how you intend to cover these costs**
- **Clearly identify responsibilities for data management in your project**
- **Describe costs and potential value of long term preservation**

The costs of making the derived data under the project context FAIR will be covered by the project itself. The platforms participating in the project are in charge to ensure the data management features during the entire project lifetime. Each platform will be responsible for the preservation of the data produced by them after the end of the Cos4Cloud project.

To ensure the preservation of the derived data, they will be stored in Community Data Portals or any other solution or repository provided by the EOSC.

Security of data will be defined by each involved platform and it will be strictly related to the proper platform. For such reason, the Data Management Plan will be updated in case of need to reflect any data security issues that may arise.

## 4. Data security

**Address data recovery as well as secure storage and transfer of sensitive data**

Security of data will be defined by each involved platform and it will be strictly related to the proper

platform. For such reason, the Data Management Plan will be updated in case of need to reflect any data security issues that may arise.

## 5. Ethical aspects

**To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former**

Ethical aspects and related policy will be defined and described, if needed, by the community in charge of each platform. Within the project personal data collecting or processing is not foreseen. In case it should be needed the project will adhere to the law as laid down in the European Directive 95/46/EEC as well as the relevant national laws and regulations, including the **General Data Protection Regulation** (**GDPR**) (EU) regulation 2016/679.

As already mentioned, Cos4Cloud project is not producing any new data, but will post-process already collected data from existing registries.  Cos4Cloud has established a Data Protection Officer (DPO). The designated DPO is José López Calvo , the CSIC DPO (
Contact: Delegado de protección de datos. Consejo Superior de Investigaciones Científicas, C/ Serrano 117, 28006, Madrid. E-mail: delegadoprotecciondatos [at ] csic.es). He will be in charge of confirming that all data collection and processing are carried out according to EU and national legislation. Cos4Cloud will keep on file the procedures that will be implemented for data processing in planned and future use cases, making sure that they comply with national and EU legislation, i.e. the General Data Protection Regulation (GDPR). The ethical aspects and related policies will be continuously monitored and evaluated for existing and new use cases and the ethics requirements will be updated accordingly. New citizen science observatories joining the project will be warned on the need to fulfil the GDPR.

**Ethical aspects and related policies will be continuously monitored and evaluated and this DMP will be updated accordingly. The management of personal data will follow the procedures available in the Cos4Cloud ethical guidelines available in the deliverables 9.1, 9.2, 9.3 and 9.4.**

## 6. Other

**Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)**

---

# Co-designed Citizen Observatories Services for the EOS-Cloud - Detailed DMP

## 1. Data summary

### State the purpose of the data collection/generation

The data managed within the context of the project is related to the following project objectives:
**O1**. Integrate Citizen Science in the European Open Science landscape through the development of a Minimum Viable Ecosystem for Citizen Science Observatories integrated to the EOSC.
Citizen Science projects generally use applications to enhance the collaboration of individuals capable of generating new data from different disciplines. The objectives is to facilitate the use of this data by scientist and any other stakeholders.

**O3**. Increase the **quantity and quality of the data available from citizen science under the FAIR** data principles (findable, accessible, interoperable & reusable) and extend them with added principles.

The way in which the data is gathered is different along the different Citizen Science Observatories. Cos4Cloud aims to improve the quality of the data adopting best practices on data management, and apply the FAIR principles for the data produced.

The different Citizen Science Observatories that are connected with specific **platforms** like mobile apps have been designed at architectural level differently so that the data models, formats, granularity, etc. may differ. The above mentioned Cos4Bio and Cos4Env aim at integrating the data from different sources using a common layer, which can be difficult. As stated earlier, the project will not produce any data itself, but some derived data can be produced, so this Data Management Plan does not refer to the data produced by the platforms, but due to the heterogeneity of these data, it is interesting to describe the information produced by them. In any case, the data will try to enhance the FAIRness of the data produced by the Citizen Science Observatories to be connected with the European Open Science Cloud, and the derived data will be treated as FAIR and published in the proper repositories.

The following list provides some details about the different platforms in terms of formats and type of gathered information. However, due to the need of citizen involvement, the volume of data to be generated is difficult to calculate.

### Explain the relation to the objectives of the project

**O1**. Integrate Citizen Science in the European Open Science landscape through the development of a Minimum Viable Ecosystem for Citizen Science Observatories integrated to the EOSC.
Citizen Science projects generally use applications to enhance the collaboration of individuals capable of generating new data from different disciplines. The objectives is to facilitate the use of this data by scientist and any other stakeholders.

**O3**. Increase the **quantity and quality of the data available from citizen science under the FAIR** data principles (findable, accessible, interoperable & reusable) and extend them with added principles.

The way in which the data is gathered is different along the different Citizen Science Observatories. Cos4Cloud aims to improve the quality of the data adopting best practices on data management, and apply the FAIR principles for the data produced.

The different Citizen Science Observatories that are connected with specific **platforms** like mobile apps have been designed at architectural level differently so that the data models, formats, granularity, etc. may differ. The above mentioned Cos4Bio and Cos4Env aim at integrating the data from different sources using a common layer, which can be difficult. As stated earlier, the project will not produce any data itself, but some derived data can be produced, so this Data Management Plan does not refer to the data produced by the platforms, but due to the heterogeneity of these data, it is interesting to describe the information produced by them. In any case, the data will try to enhance the FAIRness of the data produced by the Citizen Science Observatories to be connected with the European Open Science Cloud, and the derived data will be treated as FAIR and published in the proper repositories.

The following list provides some details about the different platforms in terms of formats and type of gathered information. However, due to the need of citizen involvement, the volume of data to be generated is difficult to calculate.

**Specify the types and formats of data generated/collected**

The observatories

This subsection briefly introduces the different observatories emphasizing in the data part. A detailed version of the platforms and the strategic plan for the exploitation and dissemination of the results can be found in D7.3.

**Artportalen (https://www.artportalen.se/)**

Artportalen is a Citizen Science Observatory to report biodiversity observations in Sweden. Observations consist of four obligatory fields: taxa, location, date and reporter. Additional information can be uploaded, for example activity (breeding, migrating etc), observation method, determination method or habitat.

Artportalen has received more than 83,000,000 (as of April 2021) observations of birds, plants, insects, fungi and many other taxa along with the 1,300,000 associated media files. Nearly 4,000,000 of the observations that Artportalen has received have been validated by expert validators or committees. Media files can be also connected to textual metadata.

**Natusfera (https://natusfera.gbif.es/)**

Natusfera is a mobile application as well as a web platform to produce biodiversity data, which is validated by experts. It is also planning to incorporate environmental data.

Biodiversity data usually includes media files (pictures in different formats jpg, png, ... and in some cases audio records, mostly in mp3). Thanks to the mobile phones, the media files can be georeferenced, and the information can be sorted in CSV and Darwin core standard.

Some of the data validated within the context of Natusfera is published on the GBIF network.

i**Spot (https://www.ispotnature.org/)**

iSpot is a Citizen Science Observatory on biodiversity. The platform encompasses a network of over 68,000 global nature observers who have crowdsourced the identification of 30,000 taxa, through over 1,500,000 images of more than 750,000 observations of different species (Birds, Amphibians and Reptiles, Fish Fungi and Lichens,etc.). It may include media data like images, as well as georeference and other metadata values.

**Pl@ntNet (https://plantnet.org/)**

Pl@ntNet is a tool (web + mobile app) to help to identify plants with pictures. It is organized in different thematic and geographical floras. Users choose their region or area of interest from a list and select "World Flora" if their region is not available.

Images are in jpg format. Metadata and user accounts are in semi-structured format (json). Taxonomic repositories are based on the International Code of Botanical Nomenclature (Shenzen code)

Pl@ntNet data currently includes:

(i) about 300M plant observations, each observation being composed of one or more images and some metadata such as date, species names (collected and generated), organ tags, geo-location (for about 50% of them), author username (for about 25% of them), image quality ratings.

(ii) 30 botanical taxonomic repositories totalling several hundred thousand species, each associated with their scientific name, synonyms, common names, URLs to external resources and (generated) statistics

(iii) a database of nearly 2,3M user accounts, each associated with username, email, avatar (optional) and statistics (generated)

The number of users and observations is expected to double in 1-2 years.

Regarding data publication:

1. A part of Pl@ntNet's observations is publicly visible in Pl@ntNet applications (about 10M observations). This part corresponds to the observations for which the users explicitly agreed to make them public with their author name (under a cc-by-sa licence).
2. A part of Pl@ntNet's observations is shared through GBIF. About 800K observations are shared with author names and photographs (under a cc-by-sa licence). About 10M of them are shared without the images and the author names (only the species name and location is shared under cc0 licence).
3. Some subsets of Pl@ntNet data built for researchers are publicly available on various platforms (zenodo, GitLab, kaggle, etc.).

**FreshWater Watch (https://freshwaterwatch.thewaterhub.org/)**

FreshWater Watch (FWW) is a global citizen-science project, started in 2012, investigating the health of the world’s freshwater ecosystems. The main parameters measured are nitrates, phosphates, bank vegetation, wildlife presence, pollution sources, water level, water colour, presence of algae, and turbidity. FWW data are not managed within the context of the project.

**KdUINO (https://monocle-h2020.eu/Sensors_and_services/KdUINO)**

KdUINO is a low-cost open-source monitoring system to measure water transparency. Citizens build their own buoy with sensors and put it in the sea. It is possible to leave the KdUINO in the water for a long time. The buoy collects data on transparency, measured using the sensors on the KdUINO. It thus gives continuous transparency measurements in real time and provides coverage for a large coastal zone, something not possible using traditional radiometers due to their cost.

It is currently being upgraded to gather information on different colour bands (RGB). A do-it-yourself (DIY) version is being developed that will have better usability, as well as being lighter and more portable, under the MONOCLE framework.

**OdourCollect (https://odourcollect.eu/)**

Odour pollution is the second most common reason for environmental complaints in the world, after noise, and it can be a sign of greater environmental problems. OdourCollect is a free app that aims to tackle odour pollution by empowering affected citizens to build collaborative odour maps. The app promotes a driving force of change, encouraging dialogue among citizens, local authorities, industries and experts.

Any citizen can act as an observer and report georeferenced observations on the odour episode, which are open data and can be used to build collaborative odour complaint maps and identify potential odour emitting sources.

Odour observations can be validated by experts to gather data in a particular area where a community is affected by odour pollution and with the aim of co-designing local solutions with relevant stakeholders.

**iSpex (http://ispex.nl/en/)**

iSpex is an innovative way to measure aerosols and water colour based on a mobile app and a small optical add-on containing a spectrometer and a polarizer. This instrument measures properties of small particles in the sky: aerosols. It measures PM 2.5 values and water colour. The idea is based on

the Spectropolarimeter for Planetary Exploration (SPEX), sized down to allow as many people as possible to use the instrument.

The app and add-on are currently in development, and they will be fully operational in 2021. Additionally, the DDQ team is upgrading the add-on and sensor capabilities to monitor air and water quality properties.

**CanAir.io (https://canair.io/)**

CanAirIO is a Colombian Citizen Science Observatory to monitor air quality with mobile and fixed sensors for measuring air quality (Particle Material PM2.5) with mobile phones (mobile measurement) or Wi-Fi (fixed measurements) with low-cost technology and open source code.

This observatory aims to build a citizen network, an air-quality map that will allow us to know what we are breathing and how we can improve quality of life. With the data collected, citizens can independently validate official air-quality numbers: what can be measured can be improved. This knowledge empowers citizens to demand better air quality policies from governments.

The main purpose of the project is to collect air quality data and publish it, for activists, academics, and people in general. The lifecycle of data starts in the sensors of the community, then they can choose whether these data will be mobile or fixed (the sensors have these modes), and whether these data will be shared in one of the servers: one server (InlfuxDB) for fixed stations, and a second server (Firebase) for mobile stations. The data then can be accessed via Grafa (CSV) or InfluxDB API. Regarding the volume, generated data can be exported in JSON format and ~ 2 GB are produced per year.

### Specify if existing data is being re-used (if any)

Nothing

### Specify the origin of the data

Question not answered.

### State the expected size of the data (if known)

Question not answered.

### Outline the data utility: to whom will it be useful

Potential interest of the data

Due to the heterogeneity of the data and its type (biodiversity, environmental), they might be useful for different stakeholders at different levels:

- Researchers: The data collected have scientific value with different purposes. In fact, the number of people involved can not be substituted by any automatic systems like sensors or any other instruments. For example, people collecting images or any other media in the field for georeferenced species is a powerful added value of Citizen Science. Furthermore, for environmental data, the resolution at temporal or spacial level can be increased significantly and

efficiently.

- Administrations: the use of this data properly analized by scientists can support administrations like government, river basin authorities or any other public rulers in taking decisions and propose policies to improve the citizens' life quality.
- Commercial - Companies: The volume of information created will be potentially useful for companies to create added value. For example, tourisms actors knowing the best places for richest biodiversity or agriculture companies better understanding the details in a specific place in terms of environment conditions.
- Citizens: apart from the involvement in citizen science activities, the citizens can benefit from the data produced. Thanks to the data produced, the citizen can select places with better air quality or better environment conditions or visit areas with interesting species.

## 2.1 Making data findable, including provisions for metadata [FAIR data]

### Outline the discoverability of data (metadata provision)

Each platform will manage their data according to their rules although Cos4Cloud will encourage data publication fulfilling all the FAIR criteria.  Some of the platforms like Natusfera or Pl@ntnet are already using metadata standards to describe their datasets, such as EML, Darwin Core, which are the most common metadata formats to describe both Environmental and Biodiversity data. Persistent identifiers are also already being used in some cases (DOIs).

### Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?

DOIs are  persistent identifiers or handles used to identify objects uniquely, standardized by the International Organization for Standardization (ISO). For instance @PlantNet, for the part of the data that is public, is already using a global persistent identifier constructed by concatenating a persistent URL with the internal identifier, e.g.: https://identify.plantnet.org/fr/weurope/observations/1009854020

### Outline naming conventions used

Depending on CO

### Outline the approach towards search keyword

Under the Metadata formats described

### Outline the approach for clear versioning

-

**Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how**

Each platform will manage their data according to their rules although Cos4Cloud will encourage data publication fulfilling all the FAIR criteria.  Some of the platforms like Natusfera or Pl@ntnet are already using metadata standards to describe their datasets, such as EML, Darwin Core, which are the most common metadata formats to describe both Environmental and Biodiversity data. Persistent identifiers are also already being used in some cases (DOIs).

EML

Ecological Metadata Language (EML) is a metadata specification particularly developed for the ecology discipline. It is based on prior work done by the Ecological Society of America and associated efforts (Michener et al., 1997, Ecological Applications). Sponsored by ecoinformatics.org, EML Version 2.2.0 was released in 2019. Some platforms such as Natusfera are already using this standard being a proper format to describe environmental data

Darwin Core

A body of standards, including a glossary of terms (in other contexts these might be called properties, elements, fields, columns, attributes, or concepts) intended to facilitate the sharing of information about biological diversity by providing reference definitions, examples, and commentaries. Sponsored by Biodiversity Information Standards (TWDG), the current standard was last modified in October 2009. The platforms collecting biodiversity  data are already using this standard.

## 2.2 Making data openly accessible [FAIR data]

**Specify which data will be made openly available? If some data is kept closed provide rationale for doing so**

Some derived data produced within the different platforms could be published in the Cos4Cloud test beds environments or externally. As already mentioned this will depend on each platform, since they will decide which data will be openly published. Just as an example of some possible publication scenario, includes target publication platforms like GBIF that are accessible via standard protocols like HTTP. There already exist some APIs to provide machine-actionable features. The derived data produced will try to keep the same mechanisms of publications, targetting at the most effective platforms for each scientific community.

**Specify how the data will be made available**

Publication platforms

**Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?**

Via protocols like HTTP

**Specify where the data and associated metadata, documentation and code are deposited**

**Specify how access will be provided in case there are any restrictions**

-

## 2.3 Making data interoperable [FAIR data]

**Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.**

Since the platforms are responsible for their data, they should provide the mechanisms to make them interoperable. Communities publishing data derived of the use of the different platforms will be encouraged to use Open Repositories compliant with OAI-PMH and supporting basic metadata standards as Dublin Core. Also, specific or community-based metadata standards like EML (Environment) or Darwin Core (Biodiversity) will be suggested for being used.

**Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?**

MEtadata formats: EML, Darwin Core, Dublin Core

## 2.4 Increase data re-use (through clarifying licenses) [FAIR data]

**Specify how the data will be licenced to permit the widest reuse possible**

Methods for data quality assurance depends also on the platforms. The publication of derived data (datasets, neural network weights, etc...) will be stored using digital resources in an Open Repository where the FAIR principles are applied and its re-use is promoted. If possible, both Platform and derived data will be published in community and interest data portals, like GBIF for Biodiversity, which supports perfectly the FAIR principles. The derived data will adopt Creative Common like licence or any other licence enabling the proper re-use, always respecting the original licence if needed. The collected data stored in the project test beds will keep the data origin licences.

Regarding software, the different Platforms have been developed adopting diverse licences types. Although they will keep their original licence, the project will stimulate the adoption of open licences to enhance the Open Science characteristics of the EOSC. The developments and any other software product created within the context of the project will be available as an Open Source resources under

open licences like Creative Commons, Apache or MIT (to be defined).

**Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed**

Depending on COs

**Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why**

EOSC

**Describe data quality assurance processes**

Depending on COs

**Specify the length of time for which the data will remain re-usable**

-

## 3. Allocation of resources

**Estimate the costs for making your data FAIR. Describe how you intend to cover these costs**

The costs of making the derived data under the project context FAIR will be covered by the project itself. The platforms participating in the project are in charge to ensure the data management features during the entire project lifetime. Each platform will be responsible for the preservation of the data produced by them after the end of the Cos4Cloud project.

To ensure the preservation of the derived data, they will be stored in Community Data Portals or any other solution or repository provided by the EOSC.

Security of data will be defined by each involved platform and it will be strictly related to the proper platform. For such reason, the Data Management Plan will be updated in case of need to reflect any data security issues that may arise.

**Clearly identify responsibilities for data management in your project**

-

**Describe costs and potential value of long term preservation**

-

## 4. Data security

**Address data recovery as well as secure storage and transfer of sensitive data**

No sensitive data

## 5. Ethical aspects

**To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former**

Ethical aspects and related policy will be defined and described, if needed, by the community in charge of each platform. Within the project personal data collecting or processing is not foreseen. In case it should be needed the project will adhere to the law as laid down in the European Directive 95/46/EEC as well as the relevant national laws and regulations, including the **General Data Protection Regulation** (**GDPR**) (EU) regulation 2016/679.

As already mentioned, Cos4Cloud project is not producing any new data, but will post-process already collected data from existing registries.  Cos4Cloud has established a Data Protection Officer (DPO). The designated DPO is José López Calvo , the CSIC DPO (
Contact: Delegado de protección de datos. Consejo Superior de Investigaciones Científicas, C/ Serrano 117, 28006, Madrid. E-mail: delegadoprotecciondatos@csic.es). He will be in charge of confirming that all data collection and processing are carried out according to EU and national legislation. Cos4Cloud will keep on file the procedures that will be implemented for data processing in planned and future use cases, making sure that they comply with national and EU legislation, i.e. the General Data Protection Regulation (GDPR). The ethical aspects and related policies will be continuously monitored and evaluated for existing and new use cases and the ethics requirements will be updated accordingly. New citizen science observatories joining the project will be warned on the need to fulfil the GDPR.

**Ethical aspects and related policies will be continuously monitored and evaluated and this DMP will be updated accordingly. The management of personal data will follow the procedures available in the Cos4Cloud ethical guidelines available in the deliverables 9.1, 9.2, 9.3 and 9.4.**

## 6. Other

**Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)**

-

# Co-designed Citizen Observatories Services for the EOS-Cloud - Final review DMP

## 1. Data summary

### State the purpose of the data collection/generation

### Provide a summary of the data addressing the following issues:

- State the purpose of the data collection/generation
- Explain the relation to the objectives of the project
- Specify the types and formats of data generated/collected
- Specify if existing data is being re-used (if any)
- Specify the origin of the data
- State the expected size of the data (if known)
- Outline the data utility: to whom will it be useful

The Cos4Cloud project involves different citizen observatories (CO) from biodiversity and environmental fields, which gather different types of data. These kinds of services have been generating data for years, and they use different kinds of devices like mobile phones, cameras and different types of sensors. The FAIR principles (findable, accessible, interoperable & reusable) have become the way to go in terms of data. In particular, within the European Open Science Cloud context, one of the main objectives is to develop the “web of FAIR data”. The idea of the implementation of the FAIR principles has been traditionally linked with data repositories. These kinds of systems, such as Zenodo, integrate data, metadata and they also assign Persistent Identifiers to uniquely refer to a digital object. Due to the dynamics of the data produced by citizen observatories, this approach of publishing data via traditional repositories is not always valid. Citizen observatories gathering real time data has often provided data via APIs or Databases, which also enables reusability with a more agile approach. This remark is important for Cos4Cloud. While some of the Citizen Observatories publish their curated data in platforms like GBIF, some others provide APIs or access to databases. The FAIR principles themselves aim at ensuring the quality of the data, so it is very important to assess that quality before publishing in traditional repositories. Furthermore, the Citizen Observatories involves different types of users, who control their data and decide what is the best way to be published.

Cos4Cloud can not define the data life cycle of the citizen observatories. The project develops services that enforce FAIRness but each CO has its own standards and flows of data. The Cos4Cloud services reuse data from the COs to create an added value, like validation of biodiversity and environmental observations by expert scientists.

Based on the different types of services in the project context, the following types have been defined:

- **Citizen observatories data**:  Cos4Cloud relies on the participation of a network of eleven citizen observatories and Do it Yourself (DIY) initiatives focused on biodiversity and the environmental domain. These platforms will test the the services developed within the project with their users. The COs (Citizen Observatories) collect, store and manage data using mobile and web applications.  Most of the COs have been running for years and have their own data management procedures.
- **Services data:** To ensure the operation of the services involved as well as the right development of the project, some data is gathered and stored internally. That data is required to support different activities, and in some cases, due to the sensitive characteristics, it is only managed

temporarily or not made public. The data produced under this group is not shared and it is managed under the proper directives like the GDPR.
- **Personal data:** data produced as result of engagement activities that include personal data as names and emails. The data under this group is not shared and it is managed under the proper directives like the GDPR.

This section summarizes the general status of the data from the different involved services. The rest of the sections aim at providing information about how Cos4Cloud is pushing to improve the data Findability, Accessibility, Interoperability and Reusability.

**Explain the relation to the objectives of the project**

The data managed within the context of the project is related to the following project objectives:

***O1****. Integrate Citizen Science in the European Open Science landscape through the development of a Minimum Viable Ecosystem for Citizen Science Observatories integrated to the EOSC.*

Although the data produced by citizen observatories are managed out of the scope of the project, COS4CLOUD is improving the integration of the Citizen Science Observatories in the EOSC, not only for the citizens themselves, but also for scientists and experts to easily get data from them.

**O3**. Increase the **quantity and quality of the data available from citizen science under the FAIR data principles** and extend them with added principles.

The way in which the data is gathered is different along the different Citizen Science Observatories. Cos4Cloud aims to improve the quality of the data by adopting best practices on data management, and apply the FAIR principles to the data produced.

**Specify the types and formats of data generated/collected**

The COs that are connected with specific **platforms** like mobile apps have been designed at architectural level differently so that the data models, formats, granularity, etc. may differ. In order to increase the quality of data produced by the COs and DIY services, tools like Cos4Bio and Cos4Env aim at providing feedback to the applications to improve the service provided and the data published, increasing its quality thanks to the connection with experts.

The following list provides some details about the different platforms in terms of formats and type of gathered information. However, for systems like COs, due to the need for citizen involvement, the volume of data being generated is difficult to calculate.

**1.2 Citizen observatories data (CO)**

This subsection briefly introduces the different observatories emphasizing the data part. A detailed version of the platforms and the strategic plan for the exploitation and dissemination of the results can be found in COS4CLOUD Deliverable 7.3.

**Artportalen (**https://www.artportalen.se/**)**

Artportalen is a Citizen Science Observatory to report biodiversity observations in Sweden. Observations consist of four compulsory fields: taxon, location, date and reporter. Additional information can be uploaded, for example activity (breeding, migrating etc), observation method, determination method or habitat.

Artportalen has received more than 94,000,000 (as of October 2022) observations of birds, plants, insects, fungi and many other taxa along with the 1,300,000 associated media files. Nearly 4,000,000 of the observations that Artportalen has received have been verified by experts or committees. Media files can be also connected to textual metadata.

Observations can be reported from desktop PCs or via mobile devices. The checklist feature enables the user to generate a checklist of species likely to occur at a particular location on a particular date. By confirming that all species that were seen and identified are reported, the observer submits a complete list and the not observed species become important zero-observations. Reporting modules, for example the checklist feature or the Invasive and Alien Species (IAS) reporting module have been built using Artportalen's API. This API has also enabled third-party developers to build apps to read information, report observations or both.

An alarm service enables users to receive a notification when relevant observations are reported. This service is not yet available for all kinds of observations, but there are plans on making this possible.

| | |
|---|---|
| Data manager contact | Johan Liljeblad |
| Data formats | Text data. Observations consist of four compulsory fields: taxon, location, date and reporter. Additional information can be uploaded, for example activity (breeding, migrating etc), observation method, determination method or habitat. |
| Data is accessible | via the observatory portal (https://www.artportalen.se/) and an API is enabled (https://api-portal.artdatabanken.se/docs/services/sos-api-v1/operations/Observations_ObservationsBySearchDwc). |

**Natusfera (**https://natusfera.gbif.es/**)**

Natusfera is a mobile application as well as a web platform to produce biodiversity data, which is validated by experts. It is also planning to incorporate environmental data.

Thanks to the mobile phones, the media files can be georeferenced, and the information can be sorted in CSV and Darwin core standard.

Some of the data validated within the context of Natusfera is published on the GBIF network.

| | |
|---|---|
| Data manager contact | Santiago Martinez |
| Data formats | Minimum of location data in textual form, but the application is oriented to upload multimedia data (photo + location). Biodiversity data usually includes media files (pictures in different formats jpg, png, ... and in some cases audio records, mostly in mp3). |
| Data is accessible | in the Natusfera portal (https://natusfera.gbif.es/observations) |

**iSpot (https://www.ispotnature.org/)**

iSpot is a Citizen Science Observatory on biodiversity. The platform encompasses a network of over 68,000 global nature observers who have crowdsourced the identification of 30,000 taxa, through over 1,500,000 images of more than 750,000 observations of different species (Birds, Amphibians and Reptiles, Fish Fungi and Lichens,etc.). It may include media data like images, as well as georeferences and other metadata values.

| | |
|---|---|
| Data manager contact | Michael Dodd |
| Data formats | The system is oriented to support photos of different groups (Birds, Fish, Invertebrates, etc) with georeferenced metadata. |
| Data is accessible | via portal (https://www.ispotnature.org/communities/global/observations) |

**Pl@ntNet (https://plantnet.org/)**

Pl@ntNet is a tool (web + mobile app) to help identifying plants from pictures. It is organized in different thematic and geographical floras. Users choose their region or area of interest from a list. In case their region is not available they can select "World Flora".

Images are in jpg format. Metadata and user accounts are in semi-structured format (json). Taxonomic repositories are based on the International Code of Botanical Nomenclature (Shenzen code).

Pl@ntNet data currently includes:

(i) about 300M plant observations, each observation being composed of one or more images and some metadata such as date, species names (collected and generated), organ tags, geo-location (for about 50% of them), author username (for about 25% of them), image quality ratings.

(ii) 30 botanical taxonomic repositories totalling several hundred thousand species, each associated with their scientific name, synonyms, common names, URLs to external resources and (generated) statistics

(iii) a database of nearly 2,3M user accounts, each associated with username, email, avatar (optional) and statistics (generated)

The number of users and observations is expected to double in 1-2 years.

Regarding data publication:

1. A part of Pl@ntNet's observations is publicly visible in Pl@ntNet applications (about 10M observations). This part corresponds to the observations for which the users explicitly agreed to make them public with their author name (under a cc-by-sa licence).
2. A part of Pl@ntNet's observations is shared through GBIF. About 800K observations are shared with author names and photographs (under a cc-by-sa licence). About 10M of them are shared without the images and the author names (only the species name and location is shared under cc0 licence).
3. Some subsets of Pl@ntNet data built for researchers are publicly available on various platforms (zenodo, GitLab, kaggle, etc.).

| | |
|---|---|
| Data manager contact | Alexis Joly |
| Data formats | Pl@ntNet helps identify plant species from photographs using visual recognition software supported by artificial intelligence algorithms. |
| Data is accessible | Curated data is available at GBIF.org (https://www.gbif.org/dataset/7a3679ef-5582-4aaa-81f0-8c2545cafc81) |

**FreshWater Watch (**https://freshwaterwatch.thewaterhub.org/**)**

FreshWater Watch (FWW) is a global citizen-science project, started in 2012, investigating the health of the world's freshwater ecosystems. The main parameters measured are nitrates, phosphates, bank vegetation, wildlife presence, pollution sources, water level, water colour, presence of algae, and turbidity.  FWW data are not managed within the context of the project.

| | |
|---|---|
| Data manager contact | James Sprinks |
| Data formats | The main parameters measured are nitrates, phosphates, turbidity, bank vegetation, wildlife, pollution sources, water level, water speed, water colour and presence of algae provided in tabular/textual form. |
| Data is accessible | in the portal, which can be filtered by different criteria (https://www.freshwaterwatch.org/pages/explore-our-data#Dive%20Deeper) |

**KdUINO (**https://monocle-h2020.eu/Sensors_and_services/KdUINO**)**

KdUINO is a low-cost open-source monitoring system to measure water transparency. Citizens build their own buoy with sensors and put it in the sea. It is possible to leave the KdUINO in the water for a long time. The buoy collects data on transparency, measured using the sensors on the KdUINO. It thus gives continuous transparency measurements in real time and provides coverage for a large coastal zone, something not possible using traditional radiometers due to their cost.

It is currently being upgraded to gather information on different colour bands (RGB). A do-it-yourself (DIY) version is being developed that will have better usability, as well as being lighter and more portable, under the MONOCLE framework.

| | |
|---|---|
| Data manager contact | Jaume Piera |
| Data formats | Citizens build their own buoy with sensors and put it in the sea. It is possible to leave the KdUINO in the water for a long time. The buoy collects data on transparency, measured using the sensors on the KdUINO. |
| Data is accessible | through a web data portal (https://monocle-h2020.eu/Data). |

**OdourCollect (**https://odourcollect.eu/**)**

Odour pollution is the second most common reason for environmental complaints in the world, after noise, and it can be a sign of greater environmental problems. OdourCollect is a free app that aims to tackle odour pollution by empowering affected citizens to build collaborative odour maps. The app promotes a driving force of change, encouraging dialogue among citizens, local authorities, industries and experts.

Any citizen can act as an observer and report georeferenced observations on the odour episode, which are open data and can be used to build collaborative odour complaint maps and identify potential odour emitting sources.

Odour observations can be validated by experts to gather data in a particular area where a community is affected by odour pollution and with the aim of co-designing local solutions with relevant stakeholders.

| | |
|---|---|
| Data manager contact | Alex Amo |
| Data formats | Form of textual characteristics of an odour. Any citizen can act as an observer and report geo-localized observations on the odour episode, which are open data and can be used to build collaborative odour complaint maps and identify potential odour emitting sources. |
| Data is accessible | under request or using COS4CLOUD services like MECODA. |

Data manager: Alex Amo

Data Formats: Form of textual characteristics of an odour. Any citizen can act as an observer and report geo-localized observations on the odour episode, which are open data and can be used to build collaborative odour complaint maps and identify potential odour emitting sources.

Data is available under request or using COS4CLOUD services like MECODA.

**iSpex (**http://ispex.nl/en/**)**

iSpex is an innovative way to measure aerosols and water colour based on a mobile app and a small optical add-on containing a spectrometer and a polarizer. This instrument measures properties of small particles in the sky: aerosols. It measures PM 2.5 values and water colour. The idea is based on

the Spectropolarimeter for Planetary Exploration (SPEX), sized down to allow as many people as possible to use the instrument. Additionally, the DDQ team is upgrading the add-on and sensor capabilities to monitor air and water quality properties.

| | |
|---|---|
| Data manager contact | Norbert Schmidt |
| Data formats | Spectrometry data produced by instrument. |
| Data is accessible | Users can access their data thanks to COS4CLOUD developed systems like MOBIS. |

**CanAir.io (**https://canair.io/**)**

CanAirIO is a Colombian Citizen Science Observatory to monitor air quality with mobile and fixed sensors for measuring air quality (Particle Material PM2.5) with mobile phones (mobile measurement) or Wi-Fi (fixed measurements) with low-cost technology and open source code.

This observatory aims to build a citizen network, an air-quality map that will allow us to know what we are breathing and how we can improve quality of life. With the data collected, citizens can independently validate official air-quality numbers: what can be measured can be improved. This knowledge empowers citizens to demand better air quality policies from governments.

The main purpose of the project is to collect air quality data and publish it, for activists, academics, and people in general. The lifecycle of data starts in the sensors of the community, then they can choose whether these data will be mobile or fixed (the sensors have these modes), and whether these data will be shared in one of the servers: one server (InlfuxDB) for fixed stations, and a second server (Firebase) for mobile stations. The data then can be accessed via Grafa (CSV) or InfluxDB API. Regarding the volume, generated data can be exported in JSON format and ~ 2 GB are produced per year.

| | |
|---|---|
| Data manager contact | Antonio Vanegas |
| Data formats | Time series of data quality parameters. |
| Data is accessible | via portal https://canair.io/samples/first_track.html |

**MINKA**

MINKA is developed in Ruby on Rails. Behind the scenes, it uses a PostgreSQL database with the PostGIS plugin and an ElasticSearch database for text searches. It also works on a Redis database for quick key-value operations and MemCache for caching content. For an easy deployment, it relies on Docker Compose.

Minka API and documentation is under development: https://minka-sdg.org:4000/v1/docs/, the observations and images can be downloaded using the python library MECODA-Minka, developed to facilitate access to data. This library is available to download using Pypi repositories: https://pypi.org/project/mecoda-minka/ .

The library is published under GNU General Public License v3 or later (GPLv3+). The installation, use, models and ways to contribute are documented in the README file.

Using the library, it is possible to extract data from the observations collected making requests to the

API, choosing different combinations of arguments, which act as filters, and getting the observations in descending order of ID, with a maximum of 10,000 (API limitation).

| | |
|---|---|
| Data manager contact | Jaume Piera |
| Data formats | Environmental and biodiversity observations |
| Data is accessible | via portal https://minka-sdg.org/users/sign_in |

**Aire Ciudadano**

AireCiudadano is a citizen observatory born from citizens to measure air quality, especially focused on particulate matter 2.5. It takes data related to air quality both in a fixed way, through stations that report through the internet all the time, and in a mobile way, with sensors with internal battery. It was created in 2017 in response to the need of citizens to show the air quality problems in the Colombian city of Bogotá. The project is Open Source and the data produced is open.

Among the data taken, there are spatio-temporal data, with the date and time of the measurement taken and the GPS coordinates. They mainly take PM2.5 values, although optionally they can also take air humidity and temperature values.

The data is stored in JSON format, which can be easily converted to CSV via their portal. They are stored on a cloud server through a commercial provider.

Possible uses of the data include environmental activism, as it can highlight air quality issues in large cities, as well as environmental education in order to raise awareness. In addition, in the academic environment, the data can be used for a variety of studies.

| | |
|---|---|
| Data manager contact | Daniel Bernal |
| Data formats | Air quality measurements |
| Data is accessible | https://aireciudadano.com/ |

### 1.3 Services data

**COS4BIO**

Cos4Bio, a service whose main mission is to create an ecosystem that experts in Biodiversity related to Citizen Science can use to carry out searches and downloads quickly and in a standardized way, generating data sets from different sources of information, such as citizen observatories.

| | |
|---|---|
| Data manager contact | Santiago Martínez |
| Data formats | This service does not generate new data. It works with two different types of data, both related with the reusing of data provided by citizen observatories. The first type is the validation of an observation that may be done by an expert, which is reported to the original CO that generates the data. This data is internal and only shared with the CO. The second type is the integration of observations provided by COs. To improve the interoperability, Cos4Bio uses the Global Biodiversity Information Facility (GBIF) backbone to unify the list of species from different observatories. The Darwin Core standard is also used as the biodiversity standard in GBIF to share biodiversity information. |

**COS4ENV**

Cos4Env service is an online platform that integrates environmental monitoring data from various citizen observatories, potentially an enormous number of data that are of interest to the expert community.

| | |
|---|---|
| Data manager contact | Santiago Martínez |
| Data formats | This service works with two different types of data, both related with the reusing of data provided by citizen observatories. The first type is feedback or comments for a record or set of records. Experts can contribute their comments all in one place. Then the feedback will go directly to the citizen observatory where the environmental measurement came from. This data is internal and only shared with the CO. The second type is the reusing of data from CO thanks to a search engine, where the users can get the environmental data they are interested in using multiple criteria, such as ‘type of environmental measurement’: air quality, water quality, temperature, odour, etc. |

**DUNS**

DUNS (Data Use Notification Service) is a centralised service to (1) register usage of the citizen science observations downloaded from the Cos4Bio and Cos4Env portals and (2) make this information available to the citizen observatory the observation comes from. The aim is to help make citizen observatories aware of how their data is used and

reward their users’ contributions.

| | |
|---|---|
| Data manager contact | Santiago Martínez |
| Data formats | The data produced by DUNS service is only internal, with report data that is sent only to the COs including metrics about the reusing of its data. |

**FASTCAT-Cloud - FASTCAT-Edge**

FASTCAT (Flexible Ai System for CAmera Traps)-Cloud is an open website service able to:

(1) Automatically filter out most unwanted pictures and video streams, keeping images of animals. This saves you time as empty recordings or photos can be removed automatically.

(2) Integrate machine learning technology (a subset of AI – Artificial Intelligence) to automatically identify species, which means that you will see the suggested species names for each image.

(3) Provide you with the ability to obtain counts (number of species recorded or photographed), i.e. how many different species have been sighted this week, or how many times you have photographed a fox in the last 30 days.

| | |
|---|---|
| Data manager contact | Frederic F. Leymarie |
| Data formats | The data generated within this service is originally created by camera traps that can record photos or videos in a wild environment. Therefore, the format is photos or videos that are used internally. The service itself does not preserve the data and images are deleted within 30 days. This data can be connected with external COs to be published and preserved, but the COs will decide how to manage. |

**MOBIS**

The MOBIS (Mobile Observation Integration Service) service offers a nice, user-friendly interface to get valuable data from smartphone sensors and images. For example: you can take air-pollution readings using iSPEX (a smartphone add-on to ‘see’ aerosols in the sky’) and a ‘normal’ picture of lichens using the same app. This means that this innovative service will allow citizen scientists to customize their own project by collecting and combining all sorts of useful information from photographs or from low-cost sensors linked to a mobile website or a native app platform, depending on the needs/wishes of the scientific and citizen community.

| | |
|---|---|
| Data manager contact | Norbert Schmidt |
| Data formats | MOBIS does not generate new data but access to the data produced by a particular user in external services like iSpex or CanAirIO. This includes spectral images or tabular data from air quality parameters (CSV like). Rather than manage data, MOBIS allows to create pipelines to connect to existing data sources and process a particular analysis. |

**MECODA**

MECODA (ModulE for Citizen Observatory Data Analysis) is an online tools repository to facilitate the analysis and visualization of all sorts of citizen science data. MECODA is a platform to provide you with tools to analyze and view all sorts of data. It also allows you to create and share new ones for everyone to use. Anyone can use it. Moreover, if you use MECODA, you can choose the data you want to analyze. Therefore, it provides tools to analyze all sorts of data -biodiversity, environmental, social,

etc.- coming from citizen observatories such as Natusfera, other data sources or your own set.

| | |
|---|---|
| Data manager contact | Ana Alvarez |
| Data formats | MECODA does not generate data itself but reuse data from external sources like citizen observatories, including Natusfera, Odourcollect and others. It provides a set of tools for data analytics. |

**AI-Taxonomist**

This service will develop an application that allows citizen science platforms (or other research projects) to develop automatic identification tools adapted to their needs, e.g. for particular groups of species. The identification tool, based on machine learning, will include a search to find similar results. For instance, photographs of species that potentially resemble the observation made by the citizen scientist.

| | |
|---|---|
| Data manager contact | Alexis Joly |
| Data formats | The service reuses data from the COs in order to provide information of potential identified species (list of species, % of accuracy), which is generated dynamically and temporarily. |

Data manager: Alexis Joly

Data Format: The service reuses data from the COs in order to provide information of potential identified species (list of species, % of accuracy), which is generated dynamically and temporarily.

**AI-GeoSpecies**

Automatically predicting the list of species that are the most likely to be observed at a given location is a key technology for many research domains, as well as practical usage in sustainable landscape management, environmental education, ecotourism, etc.

| | |
|---|---|
| Data manager contact | Alexis Joly |
| Data formats | The service will (dynamically) provide indicators of the occurrence probability of the reported species and allow users to execute probability range queries (e.g.to focus on common or rare species in the location spot). It will also allow for the integration of warning messages conditioned by the probability of observing some particular species (e.g. to signal the presence of invasive species). |

**Pl@ntNet-ID**

This service has developed an Application Programming Interface (API) allowing users to query the Pl@ntNet identification engine. This service is available to developers, researchers and other citizen science observatories interested in plant biodiversity

| | |
|---|---|
| Data manager contact | Alexis Joly |
| Data formats | The users and third-party apps upload an image to the system, which is identified with a list of potential species and its accuracy (%). However, neither the image nor the result is stored, so management is not needed. |

**STA+**

STAplus has extended the SensorThings API data structure to apply it to citizen science. In particular, it has added these aspects: ownership of the observations, citizen science project or campaign, link between observations-author-project.

STAplus standardises citizen science data and make it accessible, interoperable and reusable among different citizen observatories and service.

| | |
|---|---|
| Data manager contact | Joan Masó Pau |
| Data formats | Data is generated dynamically under queries in SensorThings format (based on JSON). The data is reused from the Citizen Observatories connected to the service. |

Data manager contact: Joan Masó Pau

Data format: Data are generated dynamically under queries in SensorThings format (based on JSON). The data are reused from the Citizen Observatories connected to the service.

**Authenix**

A service that facilitates compliance with General Data Protection Regulation (GDPR): it provides unique user IDs to overcome data silos in independent citizen science platforms or services. It collects personal information and stores it for making it available to registered applications after the user gives consent. The data is managed externally, not in the scope of COS4CLOUD and it operates under the current privacy rules (https://www.authenix.eu/PrivacyStatement

Users manage their data and they have complete control.

More info available (https://cos4cloud-eosc.eu/authenix-gdpr/). Some of the applications and services use Authenix and few information is interchanged among the services.

### 1.4 Personal data

**CoNNect - Stakeholders**

Personal data generated by WPs 5, 6, 7, 8 are mainly related to the organization of events like workshops and webinars; organization of co-design and testing activities and  the creation of the Cos4Cloud community.

All anonymised datasets are stored in a Confluence project page.This is the project's online

management platform during the project lifetime.

The non-anonymous datasets are stored locally by the Data Controllers and not shared with others, with the exception of project generated contact lists which are stored in a strict access-controlled Confluence project page. Only those who need this information to perform their activities can access it with a one-time-use password provided by ICM-CSIC. Access to the non-anonymous data is managed by ICM-CSIC under the current privacy protection laws and GDPR and is not going to be published, except for some aggregate statistics about the participation that are published in the Cos4Cloud website. The data gathered include name, email, institution, age range, educational level, etc.

The analysis of the data for statistical and communication purposes will be carried out on de-identified or anonymised data. Anonymized information is intended to be published in a proper platform. Open systems like Zenodo could be the target repository.

The legal aspects that impact data collection are described in section 5 “Ethical aspects”.At the end of the project all non-anonymous data will be deleted unless explicit consent was given to ICM-CSIC to store the data after the project. Pictures, videos and audios used in communication and dissemination activities will be stored up to 4 years after the end of the project and shared through the project website, newsletters and social media.

The personal data is collected always with explicit consent and following current privacy protection laws and GDPR. The data is collected through:

**1- Registration form to events:** the registration is managed in platforms like Eventbrite or Google Forms.

**2- Post-event surveys:** anonymized data is collected through a Google Form after each event organized by Cos4Cloud.

**2- Website - Join Cos4Cloud community:** when an user decides to join Cos4Cloud community, they have to fill in the following form:

https://cos4cloud-eosc.eu/please-fill-this-form-out-to-join-our-community/

**3- Mailchimp:** When an user decides to join Cos4Cloud newsletter, they have to fill in the following form:

https://cos4cloud-eosc.us4.list-manage.com/subscribe?u=f5c3b8a4c4745402b7c4ffa59&id=e3c25589b8

This information is stored in a Mailchimp platform controlled by ICM-CSIC.

**4- Pictures and videos for use in communication activities.** During Cos4Cloud activities pictures and videos of the participants are always taken with explicit consent from participants..

Workshop contents allowed to be published like slides are being shared in slideshare portal (https://es.slideshare.net/GestorCos4CloudEurop)

| | |
|---|---|
| Data manager contact | Sonia Liñan |
| Data formats | Slides, spreadsheets, tables of statistics. |
| Data is accessible | Slides available at https://es.slideshare.net/GestorCos4CloudEurop<br><br>Aggregated and anonymized data available at: https://cos4cloud-eosc.eu/ |

**Specify if existing data is being re-used (if any)**

Question not answered.

**Specify the origin of the data**

Question not answered.

**State the expected size of the data (if known)**

Question not answered.

**Outline the data utility: to whom will it be useful**

Due to the heterogeneity of the data and its type (biodiversity, environmental), they might be useful for different stakeholders at different levels:

- **Researchers:** The data collected have scientific value for different purposes. In fact, the number of people involved can not be substituted by any automatic systems like sensors or any other instruments. For example, people collecting images or any other media in the field for georeferenced species is a powerful added value of Citizen Science. Furthermore, for environmental data, the resolution at temporal or spatial level can be increased significantly and efficiently.
- **Administrations:** the use of this data properly analyzed by scientists can support administrations like government, river basin authorities or any other public rulers in taking decisions and proposing policies to improve the citizens' life quality.
- **Commercial - Companies:** The volume of information created will be potentially useful for companies to create added value. For example, tourism actors knowing the best places for richest biodiversity or agriculture companies better understanding the details in a specific place in terms of environment conditions.

**Citizens:** apart from the involvement in citizen science activities, the citizens can benefit from the

data produced. Thanks to the data produced, citizens can select places with better air quality or better environment conditions or visit areas with interesting species.

## 2.1 Making data findable, including provisions for metadata [FAIR data]

### Outline the discoverability of data (metadata provision)

Each platform will manage their data according to their rules although Cos4Cloud will encourage data publication by fulfilling all the FAIR criteria.  Some of the platforms like Natusfera or Pl@ntnet are already using metadata standards to describe their datasets, such as EML, Darwin Core, which are the most common metadata formats to describe both Environmental and Biodiversity data. Persistent identifiers are also already being used in some cases (DOIs), for example, those publishing their curated data in GBIF. for example. AFFOUARD A, JOLY A, LOMBARDO J, CHAMP J, GOEAU H, BONNET P (2020). Pl@ntNet observations. Version 1.2. Pl@ntNet. Occurrence dataset https://doi.org/10.15468/gtebaa accessed via GBIF.org on 2022-10-18. https://www.gbif.org/occurrence/2644182894

Apart from that, there are different services developed under the context of Cos4Cloud that increase the findability, in particular:

- COS4BIO: allows to find data from biodiversity citizen observatories. It helps experts to validate data from COs.
- COS4ENV: allows to find data from environmental citizen observatories. It helps experts to validate data from COs.
- MOBIS: connects to different services like iSpex or CanAirIO to reuse the data, find the data from different sources and integrate it.
- MECODA: create easy access to the datasets from citizen observatories and make them available in an open data analysis tool.

### Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?

Question not answered.

### Outline naming conventions used

Question not answered.

### Outline the approach towards search keyword

Question not answered.

**Outline the approach for clear versioning**

Question not answered.

**Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how**

Question not answered.

## 2.2 Making data openly accessible [FAIR data]

**Specify which data will be made openly available? If some data is kept closed provide rationale for doing so**

Accessibility from FAIR principles refers to the use of particular protocols to access both metadata and data, as well as the use of proper metadata terms. Some derived data produced within the different platforms could be published in the Cos4Cloud test beds environments or externally. As already mentioned this will depend on each platform, since they will decide which data will be openly published. Just as an example of some possible publication scenarios, includes target publication platforms like GBIF that are accessible via standard protocols like HTTP. There already exist some APIs to provide machine-actionable features. The derived data produced will try to keep the same mechanisms of publications, targeting at the most effective platforms for each scientific community.

Thanks to the features implemented in Cos4Cloud services, the accessibility is improved this way:

- **COS4BIO:** uses the Darwin Core standard to share biodiversity information and improve accessibility.
- **COS4ENV:** uses an adaptation of the Darwin Core standard to share environmental data and improve accessibility.
- **MECODA**: facilitates the access of different data sources from citizen observatories to be analyzed. It includes mechanisms to easily get the data and process it.
- **APIs**: there are different services that implement Application Programming Interfaces which facilitate the access to data and its management. For example, AI-Taxonomist, AI-GeoSpecies and Pl@ntNet-ID allows machine-actionable ways of querong from data. STA+ extends the original model to query data from different services using one single endpoint.

**Specify how the data will be made available**

Question not answered.

**Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?**

Question not answered.

**Specify where the data and associated metadata, documentation and code are deposited**

Question not answered.

**Specify how access will be provided in case there are any restrictions**

Question not answered.

## 2.3 Making data interoperable [FAIR data]

**Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.**

Since the different citizen observatories have been designed differently, their data models can be diverse. This way it is needed to develop mechanisms of making them interoperable, in particular for similar data groups. In particular, the following services support the interoperability in Cos4Cloud:

- **COS4BIO**: to improve the interoperability, Cos4Bio uses the Global Biodiversity Information Facility (GBIF) backbone to unify the list of species from different observatories. The Darwin Core standard is also used as the biodiversity standard in GBIF to share biodiversity information.
- **COS4ENV**: harmonize data from different environmental data observatories.
- **MOBIS**: offers a way to connect to different citizen observatories from one single service.
- **MECODA**: allows the integration of different citizen observatories data sources to perform different types of analysis.
- **STA+**: The data model proposed by STA+ harmonize the data produced by different citizen observatories.

**Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?**

Question not answered.

## 2.4 Increase data re-use (through clarifying licenses) [FAIR data]

**Specify how the data will be licenced to permit the widest reuse possible**

One of the particularities of the data produced under a citizen observatory project is the variety of data producers. The data gathered by citizens need to be properly treated, and the credit given in a fair manner. That is why, Cos4Cloud suggests providing data using tracking for the users. In particular, the DUNS service  (Data Use Notification Service) is a centralized service to register usage of the citizen science observations downloaded from the Cos4Bio and Cos4Env portals and make this information available to the citizen observatory the observation comes from. The aim is to help make citizen observatories aware of how their data is used and reward their users’ contributions.

**Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed**

Question not answered.

**Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why**

Question not answered.

**Describe data quality assurance processes**

Question not answered.

**Specify the length of time for which the data will remain re-usable**

Question not answered.

## 3. Allocation of resources

### Estimate the costs for making your data FAIR. Describe how you intend to cover these costs

The main data products related to the project are produced by the COs themselves. Those services have been living before the start of the project and they have their own mechanisms of sustainability, including the proper management of data and its preservation. The platforms participating in the project are in charge to ensure the data management features during the entire project lifetime. Each platform will be responsible for the preservation of the data produced by them after the end of the Cos4Cloud project.

To ensure the preservation of the derived data, they will be stored in Community Data Portals or any other solution or repository provided by the EOSC. For example, any relevant Cos4Cloud results / materials / documents will be uploaded to Zenodo, as catch-all repository supported at the EOSC.

### Clearly identify responsibilities for data management in your project

Question not answered.

### Describe costs and potential value of long term preservation

Question not answered.

## 4. Data security

### Address data recovery as well as secure storage and transfer of sensitive data

Security of data is defined by each involved platform and it will be strictly related to the proper platform. For such reasons, the Data Management Plan will be updated in case of need to reflect any data security issues that may arise.

## 5. Ethical aspects

### To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

Ethical aspects and related policy will be defined and described, if needed, by the community in charge of each platform. The personal data gathered by the project have adhered to the law as laid down in the European Directive 95/46/EEC as well as the relevant national laws and regulations, including the **General Data Protection Regulation** (**GDPR**) (EU) regulation 2016/679.

As already mentioned, Cos4Cloud project is not producing any new scientific data, but will post-process already collected data from existing registries.  The new data produced is a consequence of the activities developed by the engaged work packages, which has been managed according to the current laws. Cos4Cloud has established a Data Protection Officer (DPO), José López Calvo , the CSIC DPO.

Contact: Delegado de protección de datos.
Consejo Superior de Investigaciones Científicas, C/ Serrano 117, 28006, Madrid.
E-mail: delegadoprotecciondatos [at ] csic.es).

He will be in charge of confirming that all data collection and processing are carried out according to EU and national legislation. Cos4Cloud will keep on file the procedures that will be implemented for data processing in planned and future use cases, making sure that they comply with national and EU legislation, i.e. the General Data Protection Regulation (GDPR). The ethical aspects and related policies will be continuously monitored and evaluated for existing and new use cases and the ethics requirements will be updated accordingly. New citizen science observatories joining the project will be warned on the need to fulfill the GDPR.

Within the context of the project, Work Package 9 has released different Deliverables related to ethics. In particular, D9.2 describes the ethics review process within Cos4Cloud and gives general guidelines on where responsibilities lie and how they can be fulfilled. It describes concrete procedures for assessing the need for ethics issue mitigation and provides tools for the practical work of managing ethics issues on the task level of the project

## 6. Other

**Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)**

Question not answered.