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 DMP

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)
