AI4 | eosc

Data Management Plan

Deliverable ID	D1.2 - Data Management Plan
Work Package Reference	WP1
Issue	1.0
Due Date of Deliverable	28/02/2023
Submission Date	13/03/2023
Dissemination Level	PU
Lead Partner	CSIC
Contributors	CSIC, MMIS, PRED, WODR, KIT
Grant Agreement No	101058593
Call ID	INFRA-2021-EOSC-01

Funded by
the European Union

Prepared by	Reviewed by	Approved by
F. Aguilar I. Heredia V. Kozlov	M. David A. Calatrava	Project Management Board

Issue	Date	Description	Author(s)
0.1	10/02/2023	DMP structure and draft	F. Aguilar
0.2	20/02/2023	Use Case data description	I. Heredia, V. Kozlov
0.3	13/03/2023	Reviewed and approved	PMB

Funded by
the European Union

Table of contents

Funded by
the European Union

1. Introduction

The primary objective of the initial Data Management Plan (DMP) for the AI4EOSC project is to ensure compliance with GDPR regulations and to provide a comprehensive overview of how the project will intend to manage the data generated within the scope of AI4EOSC in accordance with the relevant principles of making data findable, accessible, interoperable, and reusable (FAIR). The DMP outlines the type of data that the project will generate, how it will be made accessible for verification and re-use, and how it will facilitate potential re-use of the collected and processed data.

As this document is prepared during the initial stages of the project, the datasets referred to are in some cases external to the AI4EOSC scope and subject to revision. The DMP is a dynamic document that will be updated, revised, and expanded based on the collected datasets and other data products, as well as any changes in the project and its approach. The DMP will be reviewed and updated at the end of each reporting period and as necessary. The document is inspired by the DMP template for projects funded under Horizon Europe.

Within the list of objectives, work packages and tasks of the project, some FAIR-related activities can be found. For example, Task 7.2 aims to support the different use cases to ensure that the different digital objects potentially reusable along the workflow, adopt the FAIR principles: data, models, predictions, metadata, publications, software, etc. This will support explicitly the activities described in this DMP.

FAIR principles are part of the EOSC vision of Science, as such, AI4EOSC will be inspired by this spirit trying to extend this idea to its activities related to data management. To make this DMP a living document, it will be updated during the project and the OpenAIRE service ARGOS[1] will be used to link the dataset derived from the project.

2. Data Summary

The scientific data used, processed or produced within the context of the AI4EOSC project are basically given by the use cases defined and managed in WP6. Since one of the objectives of the AI4EOSC platform is to facilitate the access and use of computational resources for the application of artificial intelligence techniques, the data used by the different use cases are sometimes not generated within the context of the project itself, but reused from existing sources. In some cases, new data will be gathered on demand for improving and increasing the volume of the learning set. However, the use of artificial intelligence techniques can result in new models that can be packaged as a data product for publication and reuse. Also, certain outputs of the models may correspond to derived data, which can, in some cases, be interesting for publication as a product. Therefore, as a summary of the data, on the one hand, the use

[1] https://argos.openaire.eu/

cases involved in the project will reuse data that serve as input for the artificial intelligence models. These data may be private, or published in external sources, so it is not the responsibility of the AI4EOSC project how it is published and how to improve their FAIRness level. In any case, these data will be properly referenced by the project, to maintain the intellectual property. On the other hand, some data will be generated within the project scope, so it will need to be properly managed.. For this type of data, the different practices to be adopted to ensure that they are published within the FAIR principles will be defined within the Data Management Plan.

During the course of the project new use cases may join the project, so the data management plan will be updated to take them into account.

UC1 – AGROMETEOROLOGY

This Use Case is about usage of radar imagery together with in-situ measurements and numerical weather predictions (NWP) outputs to generate utilizing AI techniques for improving farmers activity. timely and precise warnings based on forecasts of high impact weather (such as thunderstorms), for farmers.

Thunderstorms can cause damage through a variety of means:

- high winds can break and damage buildings, equipment, material and plants.
- hail can cause leaf damage reducing yield or destroying plants, machines, cars, buildings, etc.

The farmer needs to check the meteorological forecast before planning any work activities and make sure to have a way of receiving weather information while working, especially at remote locations. In the current state, a farmer listens to the weather forecast about thunderstorms in the coming day. As an enhancement, he wants to know when exactly they will hit his area so that he or she can get everyone to safety.

The data used for model input comes from weather radars, which can form 2D and 3D maps stored in HDF5 files. As they come from external sources, data management is only carried out internally to preprocess and feed the defined algorithms. The expected data volume is about 300GB per year, in addition to the historical data already available. Therefore, for data processing, a volume of about 2-3TB is expected depending on the historical data to be used.

It is not expected to collect new data within the context of the project, but relying on external sources. However, the model generated with the application of artificial intelligence algorithms, together with the data and metadata for training and the model metadata may be susceptible to be considered as a data product, thus it may be beneficial for training similar models and for reproducibility purposes. This data product can be beneficial for different stakeholders such as farmers, public administrations or local governments.

Funded by
the European Union

UC2 – INTEGRATED PLANT PROTECTION

Use case 2 aims to enhance capabilities of currently used disease detection methods based on mathematical model calculations, with new possibilities of ML/DL-based models developed and scaled on the AI4EOSC platform.

ML/DL will be based on a network of meteorological data from ground stations, the results of existing mathematical models, and ground observations. At the same time, they would be enhanced with greater terrain coverage and spatial precision by using satellite data. Current precision based on ground data is about 30 km. Supporting spatial datasets of photos in visible frequencies as well as other spectra and indices, such as NDVI (normalized difference vegetation index) images.

The data used for model input combines data acquired through the use of cameras (in-situ and users' mobile phones) and data from weather stations. Photo data is stored in standard formats such as JPEG or PNG and station data in tabular formats (CSV, JSON).

Much of the data necessary for building the learning set has already been collected. Agricultural advisors have been participating for nearly 3 years in Poland in a pilot program of pests signalling, in which their task is to systematically visit the same fields and enter information about observed pests and diseases into the system in the form of descriptive data and photos.

Descriptive data includes parameters such as:

- Type of disease.
- Plant development stage.
- The stage of development of the disease (if it occurred).
- Percentage of plant infestation (determined visually).
- Whether it requires protection treatment.
- Date of observation.

However, due to the need to enlarge the collection of images, it was decided that under the project an additional collection of images would be created, increasing the volume of the learning set. For this purpose, measures were taken to clarify the guidelines for taking photographs so that the collection would contain the largest possible number of samples suitable for feeding the learning set.

The learning set will consist of pre-processed photos depicting selected sugar beet diseases along with information describing the objects in the images.

Different scenarios of disease and pest identification could be developed in further work by different actors. Both data and models are potentially exploitable in the future. Gathered data contains more information about different plants and development stages. For instance, a single picture aggregates the information about the plant stage, plant genus, disease severance, disease occupancy and exact date and place of observation. Initially trained models also will be available for further development and training for other types of threads that could be applied to plants protection systems. The estimate is roughly 300GB per season before preliminary selection in JPG or PNG picture formats.

An example of a service beneficiary are scientific institutions or other entities interested in observations showing, e.g. the development of the plant in different regions of the country (due to the fact that each observation is described by the developmental stage of the plant).

The planned number of the eDWIN[2] platform users (where the outputs will be integrated) are of the order of 100,000 in Poland (Farmers and Advisors). The results will also be integrated with other agriculture platforms in Poland. Developed AI models can be used in other countries and platforms. The important feedback is on improving the food production quality and safety by reducing usage of the pesticides.

UC3 – AUTOMATED THERMOGRAPHY

Use case 3 leverages thermal UAV-based imaging combined with artificial intelligence to identify “hot spots” (thermal anomalies) in urban settings and contributes to improving the efficiency of energy-related systems. In this instance, the general idea can be implemented within two scenarios:

1. Detecting thermal hotspots on building rooftops caused by thermal bridges. It supports urban planners and building owners in pinpointing retrofitting potential.
2. Detecting thermal hotspots caused by urban features (cars, manholes, streetlamps, buildings, etc.). It supports district heating network operators in their search for pipeline leakages by automatically removing common false alarms from the list of potential suspects.

The data are acquired by UAS (Unmanned Aircraft System) and attached “Duo Pro R” - a dual camera made by the companies FLIR and DJI for the simultaneous acquisition of dual thermal infrared (TIR) and ordinary Red/Green/Blue (RGB) images. Apart from start and landing, the flights are automated to follow a designated pattern, supervised by an experienced UAV pilot. Due to the acquisition time frame (nighttime) and height settings (large distances to the ground), people aren’t recognisable in either RGBs nor TIRs, meaning privacy concerns are negligible and official permits easily obtained. The aforementioned applies specifically to Germany.

Types and formats

Images taken by UAS are TIRs and RGBs:

- TIRs: in RJPEG (FLIR’s proprietary) format, 640x512 pixels.
- RGBs: in JPEG format, 4000x3000 pixels.

[2] https://www.edwin.gov.pl/

A subset of annotated data is published on Zenodo [RefUC3.1]

Although some of the data for model input has already been collected, new datasets may be collected on demand during the project lifetime. The data will be used to improve the models, so there is no concrete number of flights defined yet. The volume of data gathered by flight is around 200 GB and the format is the same as above: TIRs and RGBs images.

The data gathered can be beneficial for urban planners (to detect hotspots of retrofit needs in urban districts), operators of infrastructure (e.g. photovoltaics, district heat network) and service providers, but also AI researchers to test and benchmark their AI models for urban image processing.

Apart from the datasets collected to feed the models, the models themselves can be considered as another data product. The format of the model resulting from training with artificial intelligence techniques is not yet defined, but a format compatible with AI4EOSC developments will be chosen to facilitate its reuse. The beneficiaries of this data product may be users with similar use cases, who can build on an already trained model to achieve better results.

3. FAIR DATA

This section details the practices to be carried out during the project that will facilitate the adoption of the FAIR principles by the data described in the previous section. As some of the data is acquired and managed outside the context of the project, only those data collected for the different use cases and for which the responsibility lies within the project have been considered for this section. In addition, the models produced with the application of Machine and Deep Learning techniques have been considered as data products, which allow them to be used for the training of new networks.

To facilitate the adoption of FAIR principles and other good practices, WP7 has as one of its objectives "Provide the means to automatically verify the FAIRness of data within the project", through different services and techniques. Therefore, the management of project data will be in coordination with WP7. In particular, Task 7.2 aims to support the different use cases to ensure that the different digital objects potentially reusable along the workflow, adopt the FAIR principles: data, models, predictions, metadata, publications, software, etc. These components will be identified for the diverse use cases and the correct formats, protocols, standards and tools will be suggested. The final goal of this task is to connect all the elements with FAIR assessment tools in order to get some metrics about the FAIRness level in an automatic way.

There are two types of results that can be published for reusing during the project: collected data for feeding the models and the output results of the models themselves. There is data that has been gathered before the starting of the project and out of the scope of AI4EOSC. AI4EOSC will only manage and publish data gathered directly by AI4EOSC members. Zenodo will be used as default repository, although some disciplinary repositories will be explored throughout the project. Zenodo will host the collected data. The models and dataset metadata will be gathered in the AI4EOSC Exchange Database, based on database management systems.

Findability

To facilitate the findability of the collected data, it is necessary to define an appropriate metadata standard and to assign a persistent identifier to the datasets. Zenodo supports different generic metadata types such as Dublin Core[3] or Datacite[4], which will be used to describe the data published under the AI4EOSC context. The repository assigns DOIs for each published dataset and provides a versioning system that allows the traceability of the changes.For the generated models, the AI4EOSC Exchange Database system will allow the exchange of trained models within the project platform. Use of metadata standards and the publication of models for findability outside the platform are ongoing.

Accessibility

The data generated and managed within the project consortium that feeds the Machine and Deep Learning models will be published in Zenodo whenever possible. Zenodo provides manual and automatic access through well-documented APIs and the basic metadata describes the access methods. This data will be published in the different formats described in the previous sections. Although they are collected with the intention of feeding artificial intelligence models, their use may be different and the software for access diverse in each case.

The model and dataset metadata is going to be gathered in the AI4EOSC Exchange Database, based on MySQL or similar alternatives. The interaction with the Exchange Database will be done via the Exchange API. The read access to this information will be open to everyone, both via the API and the Dashboard. The write access to this information will only be granted to the owner of the module.

During the project, the procedure to publish both data and models will be evaluated in order to check if new ways or dedicated repositories could be used.

Interoperability

In order to ensure smooth interoperability and facilitate reusability, the research datasets, accompanying metadata, and documentation will comply with international standards. The consortium will prioritise the use of open file formats and adopt standard metadata formats such as Dublin Core or Datacite, which are supported by

[3] https://www.dublincore.org/specifications/dublin-core/
[4] https://schema.datacite.org/

Zenodo. The feasibility of implementing controlled vocabularies will also be considered.

During the lifetime of the project, we intend to develop a provenance framework that will standardise the metadata produced by the modules. This will facilitate the searching, reproducibility and reuse of the modules by external researchers.

The framework will consider using some of the following tools:

- ProvONE[5]: a PROV Extension Data Model for Scientific Workflow Provenance.
- Prov-ML[6]: a W3C PROV extension for provenance data representation for scientific ML.
- Metaclip[7]: a PROV extension for describing climate products, developed by IFCA and Predictia.
- ML-Schema[8]: a set of schemas for describing data mining and machine learning algorithms, datasets, and experiments.

Reusability

To ensure data quality for the use cases, during the second part of the project we plan to host some data validation/exploration tools (like CleanLab, FastDup, deepchecks and kangas) along with data labelling tools (like LabelStudio, LabelImg, refinery, VIA or Biigle).

These are tools that will be useful across projects and will be accessed through the AI4EOSC Dashboard.

All the data, metadata and model outputs generated in the project will have a CC-BY licence by default, although more restrictive licences can be considered if needed in specific cases. We may also consider Open & Responsible AI Licenses (OpenRAIL[9]), a new initiative for licensing AI artifacts. No embargo periods are planned.

4. Allocation of resources

The data processed within the project to feed the artificial intelligence models are external data that do not require additional resources from the project. For processing, they are temporarily stored within the computational systems in order to be used. The resources required have been quantified and are managed by the Workload Management System (Nomad).

[5] http://jenkins-1.dataone.org/jenkins/view/Documentation%20Projects/job/ProvONE-Documentation-trunk/ws/provenance/ProvONE/v1/provone.html

[6] https://arxiv.org/abs/1910.04223

[7] http://www.metaclip.org/

[8] https://github.com/ML-Schema/core/wiki

[9] https://www.licenses.ai/ai-licenses , https://huggingface.co/blog/open_rail

Funded by the European Union

5. Data Security

According to WP6 Gathering Technical Aspects[10], those responsible for each of the three use cases state that in no case are there any privacy issues concerning the data. Specifically, in UC1, the measurements taken do not contain personal information, as is the case in UC2 where the available data present images of an ecological environment (e.g. plants) and meteorological time series. In UC3, the data corresponds to images captured in outdoor drone flights over two German cities, which a priori may lead into a disclosure of personal information. However, it has been stated that the data collection has been carried out at a sufficient height to respect the privacy issues in force in each case.

There is no plan to collect any personal or sensitive data within the project, but, in any case, the data used or produced will be 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.

6. Ethical Aspects

As already mentioned, AI4EOSC will process already collected data from existing registries or sources to feed the models and will also produce new data to improve the models. The new data produced is a consequence of the development of the use cases, which has been managed according to the current laws. AI4EOSC and CSIC as coordinator has established a Data Protection Officer (DPO), which is the corresponding CSIC DPO.

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

In case any sensitive data would be gathered under the project, the DPO will be in charge of confirming that all data collection and processing are carried out according to EU and national legislation. AI4EOSC 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.

[10] WP6 Gathering Technical Aspects