AI4Life

Deliverable D1.4

Data Management Plan

Project Title	Artificial Intelligence For Image Data Analysis In The Life Sciences
Project Acronym	AI4Life
Project Number	101057970
Project Start Date	01.09.2022
Project Duration	36 Months

WP N° & Title	WP1: Project Coordination
WP Leaders	EMBL, Euro-BioImaging ERIC
Deliverable Lead Beneficiary	Number - Name
Dissemination Level	PU
DOI	10.5281/zenodo.7785654
Contractual Delivery Date	28.02.2023 (M6)
Actual Delivery Date	31.03.2023
Authors	Anna Kreshuk, Matthew Hartley
Contributors	Arrate Muñoz Barrutia
Reviewers	Rachel Robinson-Lehtinen

www.ai4life.eurobioimaging.eu
ai4life@eurobioimaging.eu
@AI4LifeTeam

Change Log

Version	Date	Author	Description of changes
v0.1	28.01.2023	Anna Kreshuk	Initial draft
v0.2	06.03.2023	Arrate Muñoz Barrutia	Edits and suggestions
v0.3	30.03.2023		Approved for submission

Acronyms and Abbreviations

AI	Artificial Intelligence
DL	Deep Learning
BMZ	BioImage Model Zoo

www.ai4life.eurobioimaging.eu
ai4life@eurobioimaging.eu
@AI4LifeTeam

Table of contents

www.ai4life.eurobioimaging.eu

ai4life@eurobioimaging.eu

@AI4LifeTeam

Executive Summary

The European Commission requires a Data Management Plan (referred to as "the project DMP") as a mandatory component that should be customized to suit the specific project's activities. This plan should detail the relevant project repositories and how they are managed. WP1, as part of the governance structure, has developed the Data Management Plan for AI4Life. It is responsible for overseeing the quality assurance of the project outcomes. AI4Life will not create any new primary data but will instead integrate, annotate and link existing resources. Therefore, the responsibility for data access committees and processes for each dataset remains with primary data producers.

This is version 0.1 of the AI4Life Data Management Plan (DMP), delivered in month 6 of the project. It provides information on the types of data that will be collected or created, as well as how they will be managed, processed, and shared. It also describes the standards and methodology that will be developed and used for data collection, processing, and sharing. This document follows the template and guidelines provided by the European Commission for Horizon Europe.

The project DMP is a live document that will be reviewed and updated periodically to ensure it remains up to date.

www.ai4life.eurobioimaging.eu
ai4life@eurobioimaging.eu
@AI4LifeTeam

1. Introduction

The AI4Life project is part of the European Union's Horizon Europe research and innovation programme, led by the project coordinator Euro-BioImaging and participated by ten partners, four of them being European Research Infrastructures themselves. The project started in September 2022 and will continue until September 2025.

AI4Life aims at bringing state-of-the-art AI-based image analysis to life scientists by establishing and supporting innovative services that target both researchers in the life sciences and computational methods developers in the AI and computer vision fields.

More specifically, the objectives of AI4Life are

- **Objective 1: Democratised availability of AI-based image analysis methods** as a FAIR service accessible through the AI4Life service landscape and computationally powered by the European Open Science Cloud (EOSC) infrastructure.
- **Objective 2: Establish standards** for the submission, storage and FAIR access of reference data, reference annotations (ground-truth), trained AI models, and trainable AI methods.
- **Objective 3: Simple model deployment, sharing, and dissemination** of AI-based methods as a new developer-facing service of the BioImage Model Zoo (BMZ).
- **Objective 4: Organise Open Calls and Challenges** for outstanding image analysis.
- **Objective 5: Empower common image analysis platforms with AI tools.**
- **Objective 6: Organise outreach and training events.**

This Deliverable is part of Work Package 1 (Project Coordination). It contributes to most Objectives of the AI4Life project in the following manner:

- (O1) Defining the standard ways to access, store and share data which serves as inputs and is produced as outputs of the AI-based image analysis methods executed through AI4Life services
- (O2) Defining the data management paths for data, encapsulated models and methods
- (O3) Defining the actions that can be performed on data, models and methods through the novel developer services
- (O4) Defining how the data acquired through Open Calls and models received through challenges are managed

www.ai4life.eurobioimaging.eu
ai4life@eurobioimaging.eu
@AI4LifeTeam

The goal of this Deliverable is to establish the first version of the Data Management Plan as a starting point for the continuously evolving blueprint on how the AI-relevant data and encapsulated models will be ingested, organized and distributed throughout the AI4Life project lifetime. Note that many of the required standards for data and model definitions will be developed in the course of the project and cannot be defined at month 6 – this is why the DMP will be updated frequently.

2. Description of work

In the following, we provide the summary of the data we expect to encounter in the AI4Life project and describe the data flow and how the Open Science and FAIR principles will be implemented in the project.

2.1. Data Summary

Will you re-use any existing data and what will you re-use it for?
Guidance:
State the reasons if re-use of any existing data has been considered but discarded.
What types and formats of data will the project generate or re-use?
What is the purpose of the data generation or re-use and its relation to the objectives of the project?
What is the expected size of the data that you intend to generate or re-use?
What is the origin/provenance of the data, either generated or re-used?
To whom might your data be useful ('data utility'), outside your project?

The data in AI4Life comes in two forms: imaging datasets, potentially enriched with additional annotations, and encapsulated deep learning models, presented as model metadata and model weights obtained through pretraining.

Models: based on the foundation defined in [Ouyang et al., 2022], the model metadata specification will be developed in AI4Life WP???. The format of the model weights is defined by the deep learning framework where the training was performed, usually PyTorch or Tensorflow, which in some cases can also be converted to shared formats like ONNX. Other model weight formats might arise in the course of the project, developed by external industry researchers in companies such as Google, Meta or OpenAI. The models will include those submitted by the community, either ad hoc or related to Challenges performed by WP7. Providing the means for storing and sharing pretrained models is part of Objectives 1, 2, 3, 4 and 5. Model metadata size is negligible, model weights can occupy up to a few GBs of storage. We expect to host hundreds to thousands of pretrained models by the end of the project. The model repository will be one of the main lasting contributions of AI4Life, serving life science researchers and image analysis method developers long after the project lifetime.

Imaging datasets: We will develop metadata standards for AI-ready imaging datasets, including annotations produced by human experts or through human curation. We will

not generate data ourselves, but make existing data AI-ready through correct metadata attribution as well as through the development of libraries and APIs for automated data access and handling. All data will be stored at the BioImage Archive at EMBL-EBI. All data will be related to imaging in the life sciences. Internal storage will be based on the OME-NGFF family of formats [Moore et al., 2022]. Making existing imaging data AI-ready, providing space and means of access for such data, and development of metadata standards contribute to Objectives 1, 2, 4, 5. The BioImage Archive's scalable storage systems ensure that space is not a constraint for the size of datasets. However, since AI4Life will address AI-related data, mostly accompanied by human-curated annotations, we do not expect a single dataset to exceed tens of Gigabytes. The raw data will be generated by imaging facilities in life science institutes around Europe, not specifically for AI4Life, but following research questions of individual groups. AI-related metadata and annotations will be added by the data producers in collaboration with AI4Life WP2, 4, and 6

Collection and sharing of AI-ready datasets is one of the main goals of the AI4Life project. The collection will catalyze the development of AI-based image analysis algorithms, both in life science and computer science research institutes as well as by researchers in industry. In addition to method training, it will be used to validate and compare newly proposed methods.

2.2. Data Flow

Where open calls result in datasets that will be made publicly available, these will be either uploaded to the BioImage Archive directly by the data generators, or, where sample images and smaller datasets have been uploaded to Zenodo for evaluation, ingested into the BioImage Archive.

Submission of annotations for existing data will be uploaded directly to the BioImage Archive, where they will be assigned accession identifiers and linked to the datasets they annotate.

Model weights for AI models submitted to the BioImage Model Zoo (BMZ) are held in Zenodo, and uploaded during the submission process. Model metadata is held in GitHub, and storage of this metadata is managed by the BMZ submission flow.

2.3. Open Science and FAIR sharing

2.1. Making data findable, including provisions for metadata: Will data be identified by a persistent identifier?

Yes, for models, Zenodo provides a persistent identifier. For datasets, the BioImage Archive ingestion procedures assign a persistent identifier. For data annotations, the identifiers will include versioning.

2.1. Making data findable, including provisions for metadata: Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

For models: the first version of the metadata has been proposed in [Ouyang et al., 2022]. This specification standard will be further developed in WP5. There is no disciplinary standard to be followed, although we will consult the communities developing biology-related algorithm ontologies and the communities developing standards for models specific to medical image analysis.

For datasets: will use standard imaging metadata formats and vocabularies, particularly the community standard REMBI model for imaging datasets, the OME model for imaging data and the OME-NGFF format for cloud-ready accessible data. For image annotations, format specifications will be produced and adopted as part of the project.

2.1. Making data findable, including provisions for metadata: Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

Yes, the metadata for both models and datasets include tags related to the specimen and the imaging modality. Additionally, models include tags on architecture and datasets include tags on the available annotation types.

2.1. Making data findable, including provisions for metadata: Will metadata be offered in such a way that it can be harvested and indexed?

Yes. Metadata will be presented in a machine-readable format for programmatic consumption, e.g. harvesting and indexing.

2.2.Making data accessible - Repository: Will the data be deposited in a trusted repository?

Yes, all data will be deposited in the BioImage Archive (for datasets) and Zenodo (for models)

2.2. Making data accessible - Repository: Have you explored appropriate arrangements with the identified repository where your data will be deposited?

Yes – BioImage Archive and EMBL-EBI are part of the AI4Life consortium

2.2. Making data accessible - Repository: Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

Yes, both the BioImage Archive and Zenodo assign permanent identifiers to all datasets that are resolvable to digital objects.

2.2. Making data accessible - Data:

Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.

All data will be made openly available under permissive licenses. Some third-party models in the BioImage Model Zoo might carry restrictive licenses for commercial reuse.

2.2. Making data accessible - Data:

If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Embargo before publication will be provided following standard operational procedures established in the BioImage Archive that allow data to be embargoed until publication release.

www.ai4life.eurobioimaging.eu
ai4life@eurobioimaging.eu
@AI4LifeTeam

2.2. Making data accessible - Data:
Will the data be accessible through a free and standardized access protocol?
Yes, data will be accessible via the standard HTTP(S), FTP, Aspera and Globus protocols.

2.2. Making data accessible - Data:
If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?
No restrictions on use.

2.2. Making data accessible - Data:
How will the identity of the person accessing the data be ascertained?
No need to identify, all data will be openly available, no clinical data will be considered, and all data access will be anonymous.

2.2. Making data accessible - Data:
Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?
No, data access should be defined by the institutions producing the data. AI4Life will not accept any data that cannot be made openly available.

2.2. Making data accessible - Metadata:
Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?
Yes, metadata description will be made openly available along with libraries and APIs for programmatic access to datasets and models.

2.2. Making data accessible - Metadata:
How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?
The BioImage Archive makes data available as long as it remains scientifically useful. The EMBL-EBI archives have never removed data, and many archives have existed for multiple decades.

2.2. Making data accessible - Metadata:
Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?
Yes, all libraries and APIs will be made openly available.

2.3. Making data interoperable:
What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?
We will use standard imaging metadata formats and vocabularies, particularly the community standard REMBI model for imaging datasets, the OME model for imaging data and the OME-NGFF format for cloud-ready accessible data. For image annotations, format specifications will be produced and adopted as part of the project.

2.3. Making data interoperable:
In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?
Yes, we will openly publish all newly developed format specifications and ontologies.

2.3. Making data interoperable:
Will your data include qualified references[1] to other data (e.g. other data from your project, or datasets from previous research)?

Yes, the models will link to their training data and vice versa.

2.4. Increase data re-use:
How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

Datasets acquired through the open calls within the project will be linked to any models or other analyses generated. These will include descriptions of analysis, e.g. computational notebooks.

2.4. Increase data re-use:
Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?

Yes. Data will use permissive standard licenses (e.g. CC0, CC-BY).

2.4. Increase data re-use:
Will the data produced in the project be useable by third parties, in particular after the end of the project?

Yes, this is one of the major goals of the project. Data will remain available as part of the BioImage Archive's permanent collections, linked to all related project outputs.

2.4. Increase data re-use:
Will the provenance of the data be thoroughly documented using the appropriate standards?

Yes, the REMBI metadata model that underpins the BioImage Archive includes provenance information about how data were generated.

2.4. Increase data re-use:
Describe all relevant data quality assurance processes.

All datasets generated through the open calls within the project will be curated by the BioImage Archive team and other consortium members. This will include ensuring that the data are suitably prepared and formatted for use in AI methods development.

2.4. Increase data re-use:
Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects.

In addition to datasets, AI4Life will generate AI-based models which are also addressed within this DMP.

3. Conclusion

The AI4Life Data Management Plan describes the consortium strategy to organize and manage data, in the form of AI-ready imaging datasets and pretrained AI models. The current plan will be reviewed repeatedly during the course of the project to account for user feedback after Open Calls and Challenges and user and developer feedback continuously collected by WP3.

AI4Life has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement number 101057970.

www.ai4life.eurobioimaging.eu
ai4life@eurobioimaging.eu
@AI4LifeTeam