Deliverable D5.1

Annotation standards and software, libraries and reference examples

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	WP5: Data, model and computing standards
WP Leaders	EMBL-EBI, KTH
Deliverable Lead Beneficiary	EMBL-EBI
Dissemination Level	PU
Contractual Delivery Date	31.08.2023 (M12)
Actual Delivery Date	
Authors	Matthew Hartley, Teresa Zulueta Coarasa
Contributors	
Reviewers	Arrate Muñoz Barrutia, Beatriz Serrano-Solano, Anna Kreshuk, Florian Jug, Wei Ouyang, Dorothea Dörr

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

Change Log

Version	Date	Author	Description of changes
v0.1	17.08.23	Matthew Hartley, Teresa Zulueta Coarasa	Initial draft
v0.2	23.08.23	Beatriz Serrano-Solano	Edits and suggestions
v0.3	27.08.23	Arrate Muñoz Barrutia	Edits and suggestions
v0.4	28.08.23	Matthew Hartley, Teresa Zulueta Coarasa	Final draft approved for submission

Acronyms and Abbreviations

AI	Artificial Intelligence
D	Deliverable
FAIM	Formats, Accessibility, Incentives, Metadata
FAIR	Findable, Accessible, Interoperable, Reusable
JSON	Javascript Serialised Object Notation
ML	Machine Learning
OME	Open Microscopy Environment
PU	Public
RI	Research Infrastructure
v	Version

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

Table of contents

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

Executive Summary

AI4Life aims to democratise access to modern AI methods in biological imaging by bridging the gap between life science and computer vision researchers to broaden the reach of cutting-edge techniques. A key part of the widening access is improving accessibility to the rich variety of annotated imaging data that is necessary to train, evaluate, and benchmark AI models.

The work described in this deliverable establishes standards for large-scale sharing of these annotated imaging datasets. These standards were developed with input from community experts in AI, imaging, data management, and software development. We have worked to implement them programmatically, such that they are ready for immediate use, and have created reference examples and documentation to support this use.

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.

Supporting AI services across life sciences requires coordinated standards for how models, data and annotations are stored and processed. In particular, developing infrastructure and support systems for AI model development and application requires a well-established standard model format. Ensuring that this support is sustainable requires processes for developing, evolving, and maintaining these standards across the community. The aim of WP5 is to develop standardised formats, tools, and processes to ensure that the reference data, annotations, and models that underpin AI are FAIR.

This deliverable focuses on **Objective 5.1:** Develop standards, tools, and processes for FAIR AI reference data and annotations. This objective specifically aims to create standardised ways to represent image annotations. These annotations are a critical part of making image data "AI-ready", so it can be used for model training, benchmarking, and evaluation.

2. Description of work

2.1. FAIR AI workshop

To support FAIR and open access to image annotations, we organised a virtual workshop with expert participants from AI, image analysis, tool development, image generation, and other domains. The virtual workshop ran on the 24th and 25th of January, 2023, with 46 total participants.

Over four sessions, we discussed the following topics:

- What are the important types of annotation to record, and what extra information/metadata should accompany them?
- What are the useful ways (formats, metadata) to share and present annotation data?
- How should we support/allow/encourage the sharing and archival of annotations?

- What are our community recommendations to accelerate AI methods development through sharing AI image annotations? What are the missing pieces?

The workshop participants made several recommendations:

1. Reduce the diversity of annotation formats to a small selection tailored to appropriate use cases (e.g., OME-NGFF for raster image data, geo-JSON for 2D vector annotations).
2. Provide annotation-specific metadata, for example, who created the image annotations, how the images were annotated (expert human, algorithm, crowdsourced), etc.
3. Make images and annotations accessible through standard protocols, preferably via an API for programmatic consumption.
4. Encourage generation and public sharing of annotations through systematic recognition of annotator credit.

A short document summarising the workshop outcomes is available here: https://doi.org/10.5281/zenodo.7681687, which we summarise with the acronym FAIM (Formats, Accessibility, Incentives, Metadata). A detailed description of the outcomes of the workshop was published on the AI4Life website and newsletter to reach a wider audience.

2.2. Annotation schema specification + code library

To implement the community recommendations, we developed schema (formal implementations of the community guidance on necessary metadata) using the LinkML language. These schema, together with worked examples of metadata adherent to the schema, full documentation and code examples, are publicly available on GitHub at https://github.com/BioImage-Archive/bia-faim-models, with a version of record archived at https://doi.org/10.5281/zenodo.8246386.

2.3. Reference examples

To demonstrate the utility of images accompanied by well-described annotations, we have compiled a collection of "AI-ready" datasets. These reference examples are available here as part of the BioImage Archive's collections https://bit.ly/ai-ready-bioimage-data.

2.4 Conversion and packaging code

To support the use of the recommended formats, in particular the emerging community standard OME-Zarr, we created software tools designed to enable the conversion of

individual images and segmentations into a single OME-Zarr package. Packing images and segmentations together will facilitate provenance tracking and data reuse. This code and usage instructions are available on Github https://github.com/BioImage-Archive/bia-images-masks-to-omezarr, with a version archived at https://doi.org/10.5281/zenodo.8248712.

3. Conclusion

The image data annotation standards that we have developed are community driven and meet a critical need for the life sciences imaging AI community. Full schemas, reference examples, and supporting software code enable immediate use by scientists and developers. These standards will help both data producers and consumers and accelerate the development of AI tools for bioimage analysis.

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