

MIFA: Metadata, Incentives, Formats and Accessibility guidelines to improve the reuse of AI datasets for bioimage analysis

Received: 15 November 2023

Accepted: 19 August 2025

Published online: 15 September 2025

 Check for updates

Teresa Zulueta-Coarasa ¹, Florian Jug ², Aastha Mathur³, Josh Moore ⁴, Arrate Muñoz-Barrutia ^{5,6}, Liviu Anita¹, Kolawole Babalola ¹, Peter Bankhead ⁷, Perrine Gilloteaux ⁸, Nodar Gogoberidze ⁹, Martin L. Jones ¹⁰, Gerard J. Kleywegt ¹, Paul Korir¹, Anna Kreshuk ¹¹, Aybüke Küpcü Yoldaş¹, Luca Marconato^{12,13,14}, Kedar Narayan ^{15,16}, Nils Norlin ¹⁷, Bugra Oezdemir ³, Jessica L. Riesterer ¹⁸, Craig Russell¹, Norman Rzepka ¹⁹, Ugis Sarkans ¹, Beatriz Serrano-Solano ³, Christian Tischer ²⁰, Virginie Uhlmann ^{1,21}, Vladimír Ulman ^{22,23} & Matthew Hartley ¹ ✉

Artificial intelligence (AI) methods are powerful tools for biological image analysis and processing. High-quality annotated images are key to training and developing new algorithms, but access to such data is often hindered by the lack of standards for sharing datasets. We discuss the barriers to sharing annotated image datasets and suggest specific guidelines to improve the reuse of bioimages and annotations for AI applications. These include standards on data formats, metadata, data presentation and sharing, and incentives to generate new datasets. We are sure that the Metadata, Incentives, Formats and Accessibility (MIFA) recommendations will accelerate the development of AI tools for bioimage analysis by facilitating access to high-quality training and benchmarking data.

Imaging is an essential tool in molecular, cell and developmental biology. Advances in microscopy have endowed scientists with the ability to investigate biological processes across different scales, from imaging molecules at unprecedented resolution^{1–6}, to recording whole organisms over time^{7–13}. Such microscopy experiments can generate image data amounting to terabytes in size and, in many cases, quantitative automated analysis must be undertaken to extract meaningful conclusions. AI, machine learning (Supplementary Box 1)¹⁴ and deep learning methods have emerged as essential tools for automated bioimage analysis^{14–18}. Deep learning models can identify the most representative features for specific image-related tasks (for example, segmentation). They utilize these features to convert input images into desired outputs, such as segmentation masks, while simultaneously learning more complex features¹⁹. These features can often be learned on a subset of the data and then shown to generalize well to the entire dataset. Therefore,

deep learning is particularly suitable for analyzing large microscopy datasets while minimizing human interaction.

The two most common paradigms in machine learning are supervised and unsupervised learning¹⁷. Supervised approaches take advantage of human knowledge in the form of a ground-truth annotated training set to learn the relationship between input data and the desired output. Examples of annotations include class labels, bounding boxes or segmentation masks (Box 1 and Fig. 1). On the other hand, unsupervised approaches find underlying patterns in input data without being trained with exemplary outputs. Given that currently the most successful machine learning models for bioimage analysis are supervised, having access to annotated datasets according to the Findability, Accessibility, Interoperability and Reusability (FAIR) principles²⁰ is crucial for model development. Furthermore, most deep learning models are domain specific, and their performance depends on the

A full list of affiliations appears at the end of the paper. ✉ e-mail: matthewh@ebi.ac.uk

BOX 1

Annotation types

Bounding boxes: rectangles or rectangular prisms completely enclosing a structure of interest within an image.

Class labels: tags that identify specific features, patterns or classes in images. They can be given for a whole image or for individual structures within it.

Counts: number of objects, such as cells, found in an image.

Derived annotations: additional analytical data extracted from the images. For example, the image point spread function, the signal-to-noise ratio and focus information.

Geometrical annotations: polygons and shapes that follow the outline of a region of interest in the image. These can be geometrical primitives, 2D polygons and 3D meshes.

Graphs: graphical representations of the morphology, connectivity or spatial arrangement of biological structures in an image. Graphs, such as skeletons or connectivity diagrams, typically consist of nodes and edges, where nodes represent individual elements or regions and edges represent the connections or interactions between them.

Point annotations: X, Y and Z coordinates of a point of interest in an image (for example, an object's centroid or biological landmarks such as particle positions in cryoEM, cryoET or spatial transcriptomics).

Segmentation masks: an image, the same size as the source image, with the value of each pixel representing some biological identity or background region.

Tracks: annotations marking the movement or trajectory of objects within a sequence of bioimages.

data on which the model is trained²¹. A strategy to cover a larger domain and to make models more reusable and generalizable is to use large, heterogeneous, annotated datasets for training²².

Despite how crucial training sets are for model training, testing, validation and benchmarking, access to high-quality curated images is scarce. Generating well-annotated datasets is time consuming, requiring experts to manually annotate on the order of hundreds to thousands of images or to curate machine-created annotations. Therefore, the quantity of datasets to train AI models, the number of annotated images and the density of annotations within a dataset can be insufficient for a model to generalize effectively beyond the training set. Additionally, due to the lack of a central repository for curated AI-ready annotated datasets, the training sets that are published are scattered in different storage locations, hindering their findability²². This also results in datasets being published using diverse formats, with varying degrees of metadata and unclear licenses, making reuse difficult. As scientists from different fields use different vocabularies to describe their data, dataset reuse becomes challenging when model developers lack a proper understanding of the biological domain, highlighting the importance of clear metadata.

Many community efforts have been put in place to improve dataset development, annotation, general metadata standardization and data findability. Consortia such as QUAREP-LiMi²³ or the

Open Microscopy Environment (OME) work to provide standards for bioimage data²⁴ and metadata²⁵ formats. Furthermore, metadata standards have emerged such as the REcommended Metadata for Biological Images (REMBI) guidelines for microscopy images from multiple modalities²⁶, and the Minimum Information about Highly Multiplexed and histology images²⁷. Valuable large annotated datasets and datasets curated specifically for AI purposes have been recently made available^{28–32}. In addition, useful datasets have been published as part of competitions such as the 2018 Data Science Bowl³³ or the Cell Tracking Challenge^{34,35}, and in resources like the Broad Bioimage Benchmark Collection³⁶, the Electron Microscopy Public Image Archive³⁷ or the BioImage Archive³⁸. The ‘papers with code’ initiative hosts datasets related to machine learning publications (<https://paperswithcode.com/datasets/>). Additionally, journals such as *Data in Brief* (<https://www.sciencedirect.com/journal/data-in-brief/>) and *Scientific Data* (<https://www.nature.com/sdata/>) focus on publishing articles that provide access to research data, crediting the data owners. However, to the best of our knowledge, there are no largely adopted guidelines for bioimage annotation metadata and formats in the field. Developers of AI methods for bioimage analysis would strongly benefit from a specialized repository that provides easy access to images with the corresponding standardized annotations and metadata.

To improve accessibility and reusability of annotated image datasets, we describe a series of recommendations on four main topics: Metadata, Incentives, Formats and Accessibility (MIFA). Metadata are essential to enable data reuse and search within repositories; in the first section, we present a standard for metadata for biological image annotations. Then we discuss ways to incentivize the production of AI-ready datasets by both crediting annotators and organizing events focused on data annotation and dataset curation. The third section considers different annotation file formats that address key accessibility considerations, such as being cloud-ready to support analysis of large datasets without download, handling multidimensional data and being compatible with different community tools. Finally, we examine ways to improve data presentation and retrieval from repositories.

The devil is in the metadata: toward a recommended metadata standard

Standardizing metadata enhances the findability and reusability of data. We recommend including four main categories of accompanying metadata to maximize the reuse of an AI dataset (Fig. 2). These include general study metadata (providing experimental and investigative context), metadata related to the images, metadata related to the annotations and metadata related to versioning. Supplementary Table 1 contains metadata entries as well as two examples.

The study-level metadata must include a brief description detailing the biological application of the dataset to facilitate the reuse of data by researchers from different fields. These metadata must also identify the authors and publications related to the dataset. Furthermore, to increase findability and interoperability, persistent identifiers, ontologies and controlled vocabularies³⁹ should be used when possible. The lack of clear copyright permissions is one of the main impediments developers encounter when reutilizing datasets. Therefore, the type of license the images and the annotations are under must be clearly specified, and open licenses that encourage reuse, such as CC0 (ref. 40) and CC BY⁴¹ are preferred. Finally, the study-level metadata should include pointers to any AI models trained using the dataset stored in specialized repositories such as the BioImage Model Zoo⁴² and vice versa.

The REMBI guidelines²⁶ should be used for the dataset image metadata, covering information about the biological material used in imaging, organisms and sample preparation, together with technical details of imaging methods and image analysis among others.

The attributes of the annotation metadata module are summarized in Table 1.

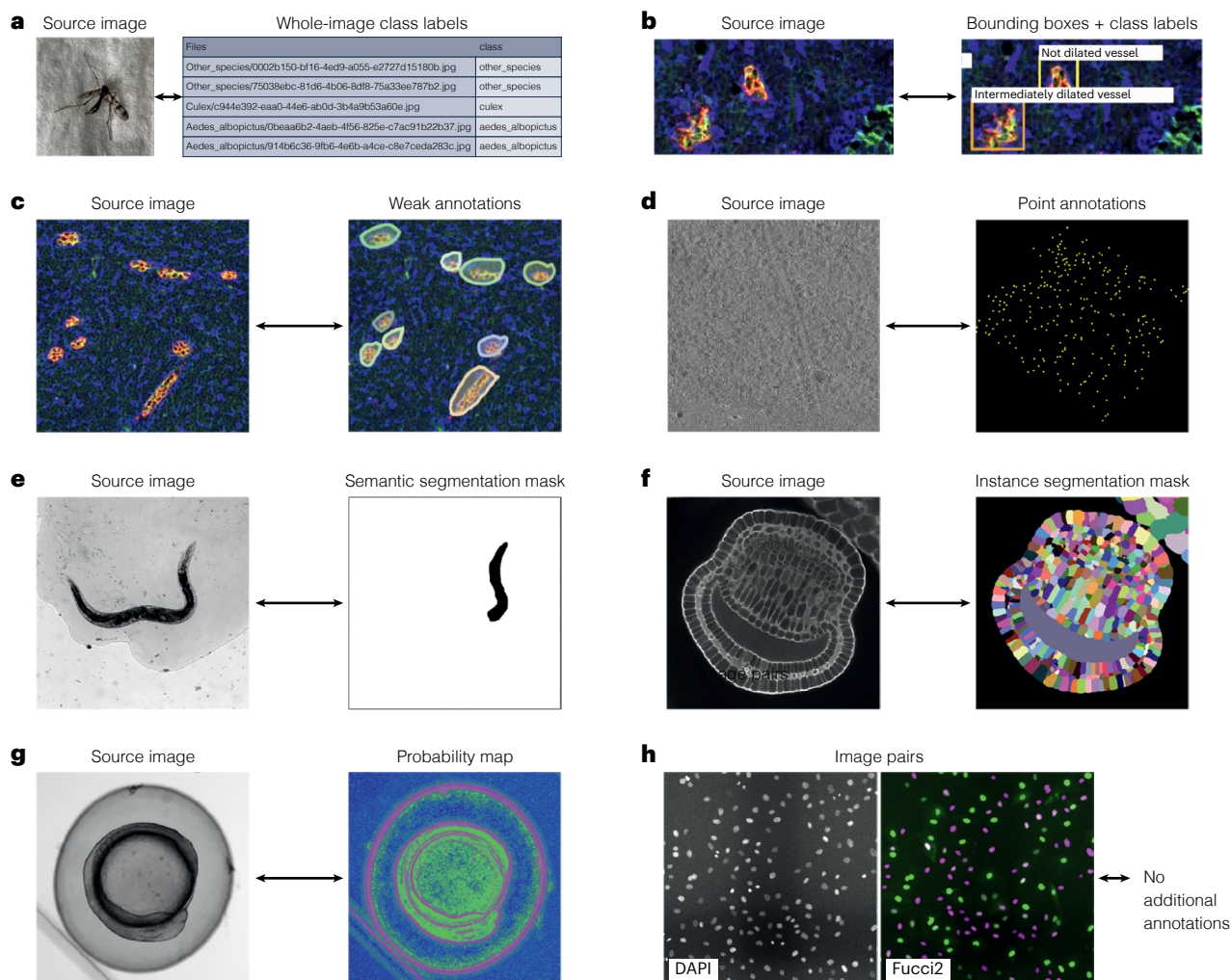


Fig. 1 | Diverse annotation types belonging to AI-ready datasets stored at the BioImage Archive and EMPIAR. a, Whole-image class labels designating mosquito genera (accession code: [S-BIAD249](https://www.mosquitoalert.com/en/), <http://www.mosquitoalert.com/en/>). Unknown image size. **b**, Bounding boxes and class labels indicating high endothelial venules with different degrees of dilation (accession: [S-BIAD463](#))⁵⁶. Unknown image size. **c**, Weak manual annotations roughly noting high endothelial venules (accession: [S-BIAD463](#))⁵⁶. Unknown image size. **d**, Point annotations noting the coordinates of the centers of ribosomes in *Saccharomyces cerevisiae* (Electron Microscopy Public Image Archive accession: [EMPIAR-11658](#))⁵⁷.

e, Semantic segmentation mask of a *Caenorhabditis elegans* head (accession: [S-BIAD300](#))⁵⁸. Unknown image size. **f**, Instance segmentation mask for all the cells in an image of *Utricularia gibba* (accession: [S-BSST734](#))⁵⁹. Image size, 208.1 × 208.1 μm. **g**, Probability map indicating the likelihood of each pixel belonging to a boundary (accession: [S-BIAD531](#))⁶⁰. Image size, 1.6 × 1.6 mm. **h**, Pair of images showing the same cells expressing nuclear (left) and cell cycle progression (right) markers. The association between the images is a form of annotation, and these pairs of images alone can be used for supervised training (accession: [S-BSST323](#))⁶¹. Unknown image size.

The final module is related to versioning metadata. Incremental versions of the dataset must have accompanying metadata with timestamps and a pointer to the previous version. A textual description of the changes between versions should also be included and, if a new version is based on a previous one, the new version should preserve the credit to the original annotators. In all cases, the identity of the creators of the original images, licensing of those images, and a link to the original images must be maintained.

Credit where credit is due: incentivizing production and sharing of AI-ready datasets

Encouraging sharing annotated image datasets through open-access data repositories requires both practical recognition of those sharing data by the scientific community, for example, as part of researcher assessment, and mechanisms to support the attribution of shared datasets to individuals and groups.

One measure to support recognition of the contribution of annotators is including their name or, ideally, a unique researcher identifier,

such as ORCID, as part of annotation metadata. This allows annotation outputs to be directly linked to individuals in a persistent way, to maintain credit for initial annotation in subsequent dataset versions and to track dataset creation as part of contributions to researcher assessment. Furthermore, data repositories can then use this information to build mechanisms to support the recognition of frequent contributors, those involved in the creation of datasets that are frequently accessed and reused by the scientific community or other objective measures of the contribution that may be developed in the future.

Widening recognition of the value of open annotated image data requires participation by the whole community. Routes toward this include organizing annotation challenges, conducting dataset workshops and providing training to support the use of annotation standards. Given the importance of annotations for the development of AI methods and, therefore, for the advancement of the bioimaging community, the involvement of funding agencies and data journals in these events is key. These efforts could help incentivize and reward the valuable contributions of annotators, foster a sense of community,

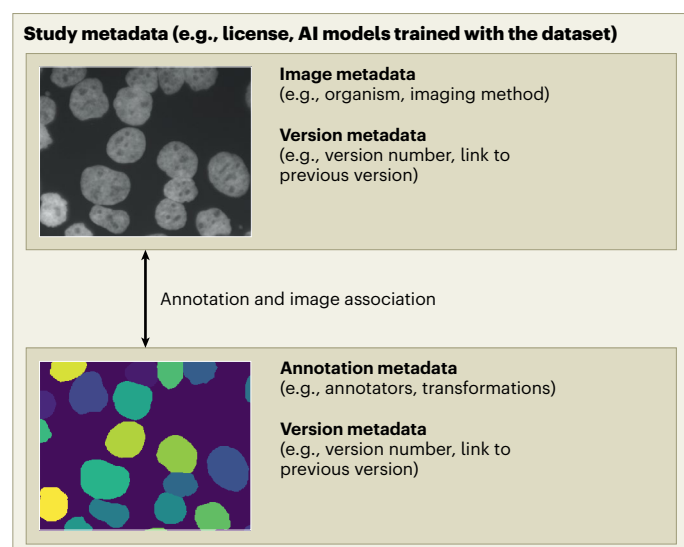


Fig. 2 | Metadata modules for AI image datasets. To make AI-ready datasets reusable, metadata must include general information about the study, the images, the annotations and the version for the image and annotation modules. Furthermore, the association between the annotations and the corresponding images must be clear. Images are from a BioImage Archive study with accession number [S-BIAD634](#) (ref. 62).

and promote and support the creation of high-quality annotations for bioimage analysis.

Formats: next step, next generation

Bioimage annotations that are part of AI-ready datasets come in a plethora of formats, from widely used CSV tables or TIFF segmentation masks to specialized formats such as AnnData for annotated data matrices⁴³ or YOLO for two-dimensional (2D) object detection⁴⁴. An important distinction to make when choosing a format is between raster-based and vector/coordinate-based formats. Raster files are pixel-based images and, therefore, raster formats are better to unambiguously assign pixels to objects. Vector files contain mathematically defined shapes, making vector-based formats the better choice to handle sparse and overlapping annotations (especially for very large images, like pathology slides).

Format considerations are one aspect of the overall issue of user accessibility. Because many imaging modalities, such as light-sheet⁴⁵ or volume electron microscopy EM⁴⁶, can create very large datasets, providing data in cloud-ready formats that support easy online access to subsets of data without the need for very large downloads is important. To achieve this, formats should also be compatible with community tools for data visualization and annotation. Some imaging techniques, like confocal microscopy, can generate image sequences in time and space. Therefore, formats that can handle multidimensional data will be necessary.

Although it is important to give annotators flexibility when creating their data, there is a need to (a) select a manageably small set of common formats that both data generators and consumers can easily read, write and edit and (b) work toward ways to combine images and annotations into a single container format. For raster-based segmentations, OME-Zarr is a cloud-ready, scalable and interoperable open format⁴⁷. Importantly, for AI-ready datasets, annotations such as segmentation masks can be packaged with the image data, allowing provenance tracking. Demonstrating its growing role in providing a common distribution format for the biological imaging community, there are several OME-Zarr tools and libraries for image and annotation visualization, and tools for format conversion⁴⁷; however, work is still needed for OME-Zarr to support vector-based annotations.

The development of parsers to work with the outputs of commercial platforms for data processing would also improve data accessibility. For example, in the context of spatial omics profiling, the ‘spatialdata-io’ effort (<https://spatialdata.scverse.org/projects/io/>) provides an open-source, community-maintained set of readers from several commercial platforms that enables the conversion of the data to a modular in-memory representation⁴⁸, and the subsequent conversion on disk to OME-Zarr. The process ensures that transformations and spatial information are parsed and saved correctly.

Careful selection of formats for vector-based annotations is recommended. A widely used option to store polygons and shapes is GeoJSON⁴⁹, a JavaScript Object Notation (JSON), multi-language format that supports several well-defined geometrical 2D data primitives such as ‘polygon’ and ‘multipolygon’, alongside ‘linestring’, ‘linearring’, ‘point’ and ‘multipoint’. Furthermore, for applications where performance is a concern, GeoJSON can be readily converted into other formats, such as GeoParquet (<https://geoparquet.org/>). For three-dimensional (3D)

Table 1 | Annotation metadata module

Attribute	Comments
Authors and contact	People involved in creating or curating annotations. Include contact information, such as email, ORCID, GitHub account or Google Scholar link.
Annotation overview	Short description of the annotation and how it was generated.
Annotation type	Annotation type, for example, class labels, segmentation mask or object counts (Box 1).
Annotation method	Crowdsourced or expertly annotated. Produced by a human or software (for example, synthetic data or silver-standard annotations obtained by combining the results of benchmarked algorithms ³³). Software used and protocols used for consensus and quality assurance.
Annotation confidence level	Confidence in annotation accuracy (for example, self-reported confidence, the variance among several annotators ³⁵ , or the number of years of experience of the annotator generating those particular annotations). It can also be noted here if these are weak annotations: rough, imprecise annotations that are fast to generate. Weak annotations are used, for example, to detect an object without providing accurate boundaries.
Annotation criteria	Rules used to generate annotations. For example, when counting cells in an image, at what point a dividing cell is considered two different cells?
Annotation coverage	Which images or regions of interest from the dataset were annotated, and what percentage of the data has been annotated from what is available? If any images or regions of interest have not been annotated, what are the reasons?
Source image association	Association between annotations and the source images from which they originated. If the annotation refers to a dataset separate from that containing the annotations, it should include the unique identifier for that dataset.
Transformations	Any coordinate transformation required to link the images to the annotations in the same common coordinate system, such as rotations and translations.
Spatial information	Spatial information for non-pixel annotations (for example, counts of items in a region of interest), including physical measurements or the region that has been annotated. If coordinates for non-rasterized annotations are stored in units other than pixels (for example, micrometers), the relevant information must be detailed in this field.
Last modification time	Date and time when the annotation was last modified.

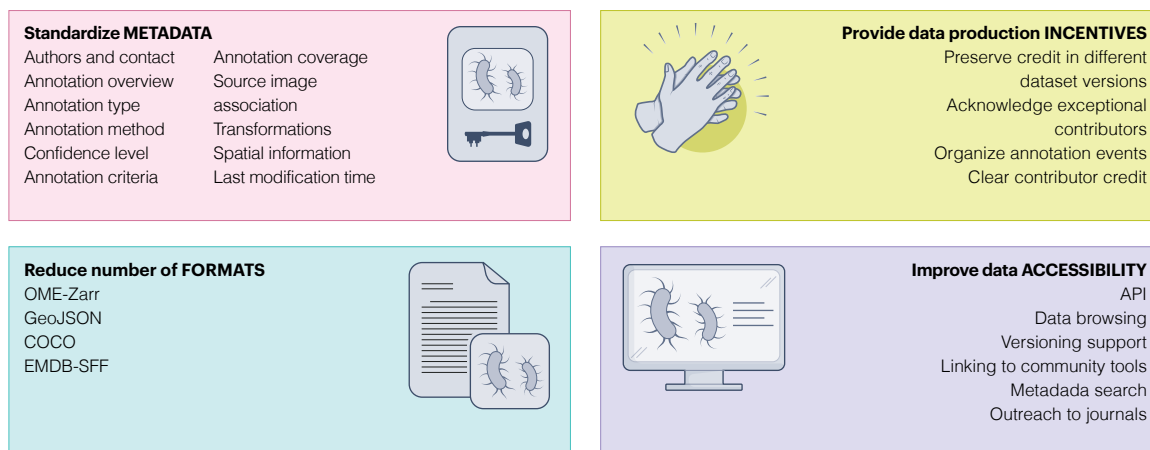


Fig. 3 | MIFA recommendations for FAIR AI data sharing. Our recommendations can be summarized in four principles: standardizing metadata, incentivizing dataset production, reducing format diversity and increasing data accessibility.

vector annotations, there are a range of formats describing meshes and point clouds. For example, the biological domain EMDB-SFF, the Segmentation File Format from the Electron Microscopy Data Bank, allows the storage of shape primitives (cones, cuboids and cylinders), surface meshes and 3D segmentation masks while also supporting structured textual annotation using ontological terms and archive accessions.

Additionally, ‘dataset’ formats can aggregate multiple annotations for a set of images. One example is the format from the Common Objects in Context (COCO) dataset⁵⁰, a JSON format that supports multiple annotation types such as segmentations and categorical classes, and is language agnostic. COCO was developed primarily for 2D computer-vision tasks and has been widely adopted by the computer-vision community beyond the specific initial dataset for which it was created. This format is starting to be used by the bioimaging community. However, changes will be required to address specific needs in the field, such as handling segmentations of very large 2D images (for example, those commonly used in histopathology) or data in three or higher dimensions.

Improving findability, accessibility and presentation of AI-ready datasets

Having a central repository for AI-ready bioimage datasets, such as the BioImage Archive, will greatly help to make data more accessible. Suggestions to increase findability within data repositories include allowing metadata search, supported by ontologies and enabling users to find images that are similar to a supplied example. Standardizing metadata according to the guidelines described above will be key to enable search. More generally, image archives should establish relationships with scientific journals to explore the possibility of including links to specific datasets in papers. Once shared, making annotated image data discoverable by a range of different data consumers is critical. Progress in the use of AI methods within biological imaging is critically dependent on being able to mix members of different communities, including imaging scientists, research biologists and AI specialists. Presenting AI-ready data in forms that are understandable by all these groups is crucial to ensure meaningful interpretation of annotations. Application programming interface (API) access is important for data integration use cases, as well as for allowing direct data querying, transformation and model training. A further advantage to this approach is that it supports standardized access to data across multiple different storage systems, without requiring homogeneous data storage.

To guarantee robust quality control, it is essential for datasets to have unique identifiers. Repeated rounds of annotation may be created with newer models, for larger subsets of the data, for different features within the same dataset or to fix problems with the annotations.

Therefore, metadata versioning should be allowed. To ensure that the quality of the annotations is maintained, a system to indicate when a dataset needs updates and to submit proposals for annotation improvements should be implemented. However, for this to work, the community will need to agree on protocols that allow experts to validate the annotations and who the reviewing experts would be in each case. A simpler alternative would be to describe the changes between two versions on the metadata so users can make an informed decision when using a set of annotations.

Discussion

As funders and scientific publishers strongly endorse open data sharing, and data deposition in repositories becomes a mandatory part of the publication process^{51,52}, scientific communities need to create guidelines to ensure that the archived data are actually useful⁵³. The MIFA guidelines aim to improve FAIR sharing of annotated bioimages by recommending four guiding principles: standardization of metadata, incentivizing dataset production, reducing the number of data formats used in the field and increasing data accessibility (Fig. 3).

These principles present a high-level roadmap toward FAIRer AI image data. Full implementation of the guidelines will require detailed work on schemas and models (particularly for software developers). Here we offer some suggestions as to how individual stakeholders can start adopting the MIFA recommendations:

Data producers can immediately start using the MIFA metadata model to capture key contextual information about their annotations. Supporting toolkits, training and high-quality examples will help with adoption.

Data repositories should develop deposition pipelines that allow users to include such metadata in their submissions. They can also implement changes to their existing presentation mechanisms according to the MIFA guidelines, for example, to allow data browsing and provide APIs for direct data access.

Software developers can include support for the recommended formats in their development of the workflows, tools and programming languages used by data producers, annotators and AI scientists. Additionally, they can support the inclusion of vector annotation formats in OME-Zarr.

Funders can require that annotated datasets that result from funded projects are shared according to the MIFA guidelines.

The whole community, including data repositories, journals, funders, software developers and scientists can participate in the organization of events to encourage the production of new high-quality annotated datasets⁵⁴.

We hope that the adoption of these guidelines by various stakeholders will help accelerate the generation and reuse of AI methods for better analysis of biological images, and all the downstream benefits to life-sciences research that this will enable. Beyond that, the MIFA guidelines would benefit any annotated dataset of biological images, not only those intended for AI applications. Standardizing annotated images would also help the development of classical image analysis algorithms and would permit data aggregation to create larger datasets. Many aspects of these guidelines also apply to medical or preclinical imaging. However, given the additional complexity of managing patient-identifiable information, differences in imaging modalities and existing medical-specific standards, we chose to focus on biological imaging. Aligning the guidelines with parallel initiatives in diagnostic imaging would be useful further work.

Paradigms such as self-supervised learning enable algorithms to learn image features from unannotated training sets, potentially eliminating the need for image annotation⁵⁵. However, the progress of self-supervised methods for bioimage analysis is limited by the lack of large collections of analogous images with consistent, harmonized metadata, which require substantial human effort to assemble. Curating such collections, for use in unsupervised/self-supervised learning could catalyze the growth of these techniques. For this, a unique metadata set will be required, encompassing the metadata outlined in Supplementary Table 1, excluding the annotations module.

AI models can perform image analysis tasks that would be difficult or impossible to achieve using traditional methods. However, the performance of AI models depends on the quality of the data used to train them. Therefore, providing open, diverse and high-quality annotated datasets will help unlock the potential of AI and push the state of the art in the bioimage analysis field. To achieve this, it is essential to strengthen the synergy between life scientists and AI developers. By incentivizing data producers to openly share expertly annotated datasets with the community, we can accelerate the development of new methods and guide developers toward the new and exciting challenges scientists encounter when analyzing their images. This way we will empower researchers in diverse fields with AI tools that will ultimately lead to new discoveries and improved understanding of living systems.

Data availability

Data availability is not applicable to this article as no new data were created or analyzed for this work. All used images are already publicly accessible, permissibly licensed and referenced by identifier.

References

- Zhang, K., Pintilie, G. D., Li, S., Schmid, M. F. & Chiu, W. Resolving individual atoms of protein complex by cryo-electron microscopy. *Cell Res.* **30**, 1136–1139 (2020).
- Nakane, T. et al. Single-particle cryo-EM at atomic resolution. *Nature* **587**, 152–156 (2020).
- Yip, K. M., Fischer, N., Paknia, E., Chari, A. & Stark, H. Atomic-resolution protein structure determination by cryo-EM. *Nature* **587**, 157–161 (2020).
- Betzig, E. et al. Imaging intracellular fluorescent proteins at nanometer resolution. *Science* **313**, 1642–1645 (2006).
- Rust, M. J., Bates, M. & Zhuang, X. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nat. Methods* **3**, 793–795 (2006).
- Hess, S. T., Girirajan, T. P. K. & Mason, M. D. Ultra-high resolution imaging by fluorescence photoactivation localization microscopy. *Biophys. J.* **91**, 4258–4272 (2006).
- Megason, S. G. In toto imaging of embryogenesis with confocal time-lapse microscopy. *Methods Mol. Biol.* **546**, 317–332 (2009).
- McDole, K. et al. In toto imaging and reconstruction of post-implantation mouse development at the single-cell level. *Cell* **175**, 859–876 (2018).
- Daetwyler, S., Günther, U., Modes, C. D., Harrington, K. & Huisken, J. Multi-sample SPIM image acquisition, processing and analysis of vascular growth in zebrafish. *Development* **146**, dev173757 (2019).
- Chen, B. -C. et al. Lattice light-sheet microscopy: imaging molecules to embryos at high spatiotemporal resolution. *Science* **346**, 1257998 (2014).
- Huisken, J., Swoger, J., Del Bene, F., Wittbrodt, J. & Stelzer, E. H. K. Optical sectioning deep inside live embryos by selective plane illumination microscopy. *Science* **305**, 1007–1009 (2004).
- Udan, R. S., Piazza, V. G., Hsu, C. -W., Hadjantonakis, A. -K. & Dickinson, M. E. Quantitative imaging of cell dynamics in mouse embryos using light-sheet microscopy. *Development* **141**, 4406–4414 (2014).
- Royer, L. A. et al. Adaptive light-sheet microscopy for long-term, high-resolution imaging in living organisms. *Nat. Biotechnol.* **34**, 1267–1278 (2016).
- Moen, E. et al. Deep learning for cellular image analysis. *Nat. Methods* **16**, 1233–1246 (2019).
- Hallou, A., Yevick, H. G., Dumitrascu, B. & Uhlmann, V. Deep learning for bioimage analysis in developmental biology. *Development* **148**, dev199616 (2021).
- Gupta, A. et al. Deep learning in image cytometry: a review. *Cytometry A* **95**, 366–380 (2019).
- Villoutreix, P. What machine learning can do for developmental biology. *Development* **148**, dev188474 (2021).
- Wang, S., Yang, D. M., Rong, R., Zhan, X. & Xiao, G. Pathology image analysis using segmentation deep learning algorithms. *Am. J. Pathol.* **189**, 1686–1698 (2019).
- LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
- Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016). **This publication introduces the FAIR principles, explains their rationale and highlights example implementations.**
- Rutschi, C., Berente, N. & Nwanganga, F. Data sensitivity and domain specificity in reuse of machine learning applications. *Inf. Syst. Front.* <https://doi.org/10.1007/s10796-023-10388-4> (2023).
- Laine, R. F., Arganda-Carreras, I., Henriques, R. & Jacquemet, G. Avoiding a replication crisis in deep-learning-based bioimage analysis. *Nat. Methods* **18**, 1136–1144 (2021). **This Comment highlights important considerations for researchers to ensure reproducibility when publishing studies using deep learning in microscopy, including validation methods, tool selection, data practices and reporting standards.**
- Boehm, U. et al. QUAREP-LiMi: a community endeavor to advance quality assessment and reproducibility in light microscopy. *Nat. Methods* **18**, 1423–1426 (2021).
- Swedlow, J. R. et al. A global view of standards for open image data formats and repositories. *Nat. Methods* **18**, 1440–1446 (2021).
- Linkert, M. et al. Metadata matters: access to image data in the real world. *J. Cell Biol.* **189**, 777–782 (2010).
- Sarkans, U. et al. REMBI: REcommended Metadata for Biological Images—enabling reuse of microscopy data in biology. *Nat. Methods* **18**, 1418–1422 (2021). **This article introduces the REMBI guidelines aimed to maximize the reuse of biological images across diverse imaging communities.**
- Schapiro, D. et al. MITI minimum information guidelines for highly multiplexed tissue images. *Nat. Methods* **19**, 262–267 (2022).
- Schwendy, M., Unger, R. E. & Parekh, S. H. EVICAN—a balanced dataset for algorithm development in cell and nucleus segmentation. *Bioinformatics* **36**, 3863–3870 (2020).

29. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nat. Methods* **18**, 100–106 (2021).
30. Edlund, C. et al. LIVECell—a large-scale dataset for label-free live cell segmentation. *Nat. Methods* **18**, 1038–1045 (2021).
31. Conrad, R. & Narayan, K. CEM500K, a large-scale heterogeneous unlabeled cellular electron microscopy image dataset for deep learning. *Elife* **10**, e65894 (2021).
32. Conrad, R. & Narayan, K. Instance segmentation of mitochondria in electron microscopy images with a generalist deep learning model trained on a diverse dataset. *Cell Syst.* **14**, 58–71 (2023).
33. Caicedo, J. C. et al. Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. *Nat. Methods* **16**, 1247–1253 (2019).
34. Ulman, V. et al. An objective comparison of cell-tracking algorithms. *Nat. Methods* **14**, 1141–1152 (2017).
35. Maška, M. et al. The Cell Tracking Challenge: 10 years of objective benchmarking. *Nat. Methods* <https://doi.org/10.1038/s41592-023-01879-y> (2023).
36. Ljosa, V., Sokolnicki, K. L. & Carpenter, A. E. Annotated high-throughput microscopy image sets for validation. *Nat. Methods* **9**, 637 (2012).
37. Iudin, A. et al. EMPIAR: the Electron Microscopy Public Image Archive. *Nucleic Acids Res.* **51**, D1503–D1511 (2023).
38. Hartley, M. et al. The BioImage Archive—building a home for life-sciences microscopy data. *J. Mol. Biol.* **434**, 167505 (2022).
39. Bard, J. B. L. & Rhee, S. Y. Ontologies in biology: design, applications and future challenges. *Nat. Rev. Genet.* **5**, 213–222 (2004).
40. Creative Commons—CCO. *Creative Commons* <https://creativecommons.org/share-your-work/public-domain/cc0/> (2009).
41. Creative Commons—Attribution 4.0 International—CC BY 4.0. <https://creativecommons.org/licenses/by/4.0/>
42. Ouyang, W. et al. BioImage Model Zoo: a community-driven resource for accessible deep learning in bioimage analysis. Preprint at *bioRxiv* <https://doi.org/10.1101/2022.06.07.495102> (2022).
43. Virshup, I., Rybakov, S., Theis, F. J., Angerer, P. & Alexander Wolf, F. anndata: access and store annotated data matrices. *J. Open Source Softw.* **9**, 4371 (2024).
44. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You only look once: unified, real-time object detection. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* <https://doi.org/10.1109/cvpr.2016.91> (IEEE, 2016).
45. Stelzer, E. H. K. et al. Light sheet fluorescence microscopy. *Nat. Rev. Methods Primers* **1**, 1–25 (2021).
46. Peddie, C. J. et al. Volume electron microscopy. *Nat. Rev. Methods Primers* **2**, 1–23 (2022).
47. Moore, J. et al. OME-Zarr: a cloud-optimized bioimaging file format with international community support. *Histochem. Cell Biol.* **160**, 223–251 (2023).
- This work introduces the cloud-optimized file format OME-Zarr, which aims to improve FAIR data access and unify file standards across fields to support efficient data management and analysis.**
48. Marconato, L. et al. SpatialData: an open and universal data framework for spatial omics. *Nat. Methods* <https://doi.org/10.1038/s41592-024-02212-x> (2024).
49. Butler, H. et al. The GeoJSON Format, RFC 7946. <https://doi.org/10.17487/rfc7946> (2016).
50. Lin, T.-Y. et al. Microsoft COCO: Common Objects in Context. Preprint at <https://arxiv.org/abs/1405.0312> (2014).
51. Data sharing is the future. *Nat. Methods* **20**, 471 (2023).
52. Kaiser, J. & Brainard, J. Ready, set, share! *Science* **379**, 322–325 (2023).
53. Sever, R. We need a plan D. *Nat. Methods* **20**, 473–474 (2023).
54. Uhlmann, V., Hartley, M., Moore, J., Weisbart, E. & Zaritsky, A. Making the most of bioimaging data through interdisciplinary interactions. *J. Cell Sci.* **137**, jcs262139 (2024).
- This article examines key players in the bioimaging field, highlights barriers to interdisciplinary interaction and proposes actions to foster a culture of open data sharing to drive innovation.**
55. Jing, L. & Tian, Y. Self-supervised visual feature learning with deep neural networks: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 4037–4058 (2021).
56. Bekkhus, T. et al. Remodeling of the lymph node high endothelial venules reflects tumor invasiveness in breast cancer and is associated with dysregulation of perivascular stromal cells. *Cancers* **13**, 211 (2021).
57. Rangan, R. et al. CryoDRGN-ET: deep reconstructing generative networks for visualizing dynamic biomolecules inside cells. *Nat. Methods* **21**, 1537–1545 (2024).
58. Galimov, E. & Yakimovich, A. A tandem segmentation-classification approach for the localization of morphological predictors of lifespan and motility. *Aging* **14**, 1665–1677 (2022).
59. Vijayan, A. et al. The annotation and analysis of complex 3D plant organs using 3DCoordX. *Plant Physiol.* **189**, 1278–1295 (2022).
60. Jones, R. A., Renshaw, M. J., Barry, D. J. & Smith, J. C. Automated staging of zebrafish embryos using machine learning. *Wellcome Open Res.* **7**, 275 (2022).
61. Rappez, L., Rakhlin, A., Rigopoulos, A., Nikolenko, S. & Alexandrov, T. DeepCycle reconstructs a cyclic cell cycle trajectory from unsegmented cell images using convolutional neural networks. *Mol. Syst. Biol.* **16**, e9474 (2020).
62. Kromp, F. et al. An annotated fluorescence image dataset for training nuclear segmentation methods. *Sci. Data* **7**, 262 (2020).

Acknowledgements

To improve support for image annotations of AI-related datasets and to develop annotation standards for the community, a workshop was held with 45 community experts from various backgrounds, including data generators, annotators, curators, AI researchers, bioimage analysts and software developers. The workshop sessions resulted in a series of recommendations on four main topics: Metadata, Incentives, Formats and Accessibility (MIFA), which are described above. We are grateful to the FAIR AI workshop participants F. Ballllosera, A. Bhardwaj, J.-M. Burel, A. French, M. Hammer, D. Hensen, K. Ho, S. Jasek, I. Kemmer, J. Kriel, A. Iudin, W. Ouyang, A. Papaleo, A. Rupaningal, C. Strambio De Castillia, B. Wester, S. Weyand and G. Zaki for their insightful inputs and valuable contributions to the discussion.

The workshop was organized in the framework of the AI4Life project, which has received funding from the European Union's Horizon Europe Research and Innovation Programme under grant agreement no. 101057970. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under contract no. 75N91019D00024. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products or organizations imply endorsement by the US Government.

J.M. is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation; 501864659 as part of NFDI4BIOIMAGE). M.L.J. is supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (CC1076), the UK Medical Research Council (CC1076) and the Wellcome Trust (CC1076). G.J.K. and P.K. were supported by EMBL-EBI and the Wellcome Trust (221371/Z/20/Z). N.N. acknowledges support from the Swedish Research Council (2023-05450), Sigurd and Elsa Goljes Memorial Foundation, IngaBritt och Arne Lundberg foundation, Magnus Bergvall Foundation, and Greta and Johan Kock's foundations. V. Ulman was supported by the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (90254). C.T. and V. Ulman are supported by grant nos. 2020-225265 and 2024-342803, respectively, from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation. P.P.-G. is a member of the national infrastructure France-Bioimaging supported by the French national research agency (ANR-24-INBS-0005 FBI BIOGEN). A.M.-B. is supported by Ministerio de Ciencia, Innovación y Universidades, Agencia Estatal de Investigación (MCIN/AEI/10.13039/501100011033/), under grant PID2023-152631OB-I00.

Author contributions

All authors attended the workshop and participated in the plenary and breakout room discussions. T.Z.-C. and M.H. led the writing process. F.J., A.M., J.M. and A.M.-B. (in alphabetical order) acted as chairs of the breakout rooms and summarized discussions. A.M., J.M., A.M.-B., L.A., K.B., P.B., P.G., N.G., M.L.J., G.J.K., P.K., A.K., A.K.Y., L.M., K.N., N.N., B.O., J.L.R., C.R., N.R., U.S., B.S.-S., C.T., V. Uhlmann and V. Ulman (in alphabetical order) provided comments and edited the manuscript.

Competing interests

N.R. is an employee of and owns equity in scalable minds GmbH, which is a company that sells image analysis software and services. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41592-025-02835-8>.

Correspondence should be addressed to Matthew Hartley.

Peer review information *Nature Methods* thanks Yingke Xu and Michele Darrow for their contribution to the peer review of this work. Primary Handling Editor: Rita Strack, in collaboration with the *Nature Methods* team.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© Springer Nature America, Inc. 2025

¹European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK. ²Fondazione Human Technopole, V.le Rita Levi-Montalcini, Milan, Italy. ³Euro-BioImaging ERIC Bio-Hub, European Molecular Biology Laboratory (EMBL) Heidelberg, Heidelberg, Germany. ⁴German BioImaging – Gesellschaft für Mikroskopie und Bildanalyse e.V., Konstanz, Germany. ⁵Universidad Carlos III de Madrid, Madrid, Spain. ⁶Instituto de Investigación Sanitaria Gregorio Marañón, Madrid, Spain. ⁷Edinburgh Pathology, Centre for Genomic and Experimental Medicine, and CRUK Scotland Centre, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK. ⁸Nantes Université, CHU Nantes, CNRS, Inserm, BioCore, US16, SFR Bonamy, Nantes, France. ⁹Imaging Platform, Broad Institute, Cambridge, MA, USA. ¹⁰Electron Microscopy Science Technology Platform, Francis Crick Institute, London, UK. ¹¹Cell Biology and Biophysics Unit, European Molecular Biology Laboratory, Heidelberg, Germany. ¹²European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Heidelberg, Germany. ¹³Division of Computational Genomics and System Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany. ¹⁴Collaboration for joint PhD degree between EMBL and Heidelberg University, Faculty of Biosciences, Heidelberg, Germany. ¹⁵Center for Molecular Microscopy, Center for Cancer Research, National Cancer Institute, NIH, Bethesda, MD, USA. ¹⁶Cancer Research Technology Program, Frederick National Laboratory for Cancer Research, Frederick, MD, USA. ¹⁷Department of Experimental Medical Science, Lund University Bioimaging Centre (LBIC), & Nanolund, Lund University, Lund, Sweden. ¹⁸Cancer Early Detection Advanced Research (CEDAR), Knight Cancer Institute, Oregon Health & Science University, Portland, OR, USA. ¹⁹Scalable Minds GmbH, Potsdam, Germany. ²⁰European Molecular Biology Laboratory, Data Science, Heidelberg, Germany. ²¹BioVisionCenter, Universität Zürich, Zürich, Switzerland. ²²IT4Innovations, VSB – Technical University of Ostrava, Ostrava-Poruba, Czech Republic. ²³CEITEC - Central European Institute of Technology, Masaryk University, Brno, Czech Republic.

✉ e-mail: matthewh@ebi.ac.uk