The BioImage Archive - home of life-sciences microscopy data

Matthew Hartley[1], Gerard Kleywegt[1], Ardan Patwardhan[1], Ugis Sarkans[1], Jason R. Swedlow[2], Alvis Brazma[1],

[1] European Molecular Biology Laboratory, European Bioinformatics Institute, EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK.
[2] Division of Computational Biology, Centre for Gene Regulation and Expression, School of Life Sciences, University of Dundee, Dundee, UK.

Highlights

- The BioImage Archive is a new archival data resource at the European Bioinformatics Institute (EMBL-EBI).
- The BioImage Archive aims to accept all biological imaging data associated with peer-reviewed publications using microscopy that probe biological structure, mechanism and dynamics, as well as other important datasets that can serve as a reference.
- The BioImage Archive aims to maximise the use of valuable microscopy data, to improve reproducibility of published results that rely on image data, and to facilitate development of both novel biological insights from existing data and new image analysis methods.
- The BioImage Archive anchors an ecosystem of related databases, supporting those resources with storage infrastructure, linkage and indexing across databases.
- Across this ecosystem, the BioImage Archive already stores and provides access to over one petabyte of image data from many different imaging modalities and biological domains.
- Future development of the BioImage Archive will support the fast-emerging next generation file formats (NGFFs) for bioimaging data, providing access mechanisms tailored toward unlocking the power of modern AI-based image-analysis approaches.

Abstract

Despite the importance of data resources in genomics and structural biology, until now there has been no central archive for biological data for all imaging modalities. The BioImage Archive is a new data resource at the European Bioinformatics Institute (EMBL-EBI) designed to fill this gap. It accepts bioimaging data associated with publication in any format, from any imaging modality at any scale, as well as reference datasets. The BioImage Archive will improve reproducibility of published studies that derive results from image data. In addition, providing reference datasets to the scientific community reduces duplication of effort and allows downstream analysis to focus on a consistent set of data. The BioImage Archive will also help to generate new insights through reuse of existing data to answer new biological questions, or provision of training, testing and benchmarking data for image analysis tool development. The Archive is available at https://www.ebi.ac.uk/bioimage-archive/.

Introduction

Imaging is a key research tool in the life sciences. “Biological imaging” encompasses a broad range of modern microscopy methods that generate large and complex datasets. Biological imaging comprises a diverse set of subdomains covering different physical scales and technological approaches. Image data provides spatial and temporal information on biological systems across a wide range of scales, together with insight into structures and interactions. In most modalities, biological imaging involves a combination of data acquisition and nontrivial data analysis workflows that together provide results that can be interpreted by biologists.

Such data creates significant opportunities for reuse, potentially enabling new scientific discoveries. To enable this reuse requires supporting access to open image data that follows the FAIR principles[1]. Where similar access has been provided for sequence and structural data, through key resources such as the ENA[2] and PDB[3], the scientific value has been immense. Such biological data resources have become a critical part of the infrastructure for life-sciences research[4]. Bioinformatics databases can be broadly categorized as either deposition databases (archives) or added-value databases[5]. Deposition databases create a persistent scientific record of the data on which published scientific conclusions are based by providing deposition pipelines for data and associated metadata and making those data searchable and accessible by the community. Added-value databases enrich data through expert curation, data integration and further analysis.

Over the past decade, several resources have emerged that have begun to tackle the challenge of publishing bioimaging datasets. In 2014, EMBL-EBI launched EMPIAR, the Electron Microscopy Public Image Archive, in response to community demand for public archiving of raw 2D image data to support the validation of 3D cryo-EM structures[6]. In 2016, a collaboration between the OME consortium and EMBL-EBI resulted in the Image Data Resource (IDR), a platform for bioimage data integration and reanalysis[7]. In parallel, the Systems Science of Biological Dynamics Database (SSBD)[8] began publishing biological imaging datasets that measured temporal changes in various model systems. In addition, several institute- or project-specific resources have emerged that made cell or tissue imaging datasets available (for example, the Allen Cell Explorer, https://www.allencell.org/)[9]. Although these added-value resources met specific demands, there remained a gap in provision for a broader image archive, leading to a community call for the development of a public bioimage archive in 2018[4]**.**

In 2019, EMBL-EBI launched the BioImage Archive to meet this need. The initial launch provided both a central hub to link together IDR and EMPIAR, and direct data deposition through BioStudies, EMBL-EBI’s resource for data integration and data that does not fit into existing specialised archives[10]. The BioImage Archive now operates as a data resource of its own identity, separate from its historic antecedents. It supports rapid direct deposition of novel imaging datasets associated with publications, import of datasets from other biological imaging resources to improve sustainability and integration with other data resources.

In this article we give an overview of the BioImage Archive, highlighting its current collections, how submitters and users interact with the archive and describe future development plans.

Results

Purpose and scope

The primary goals of the BioImage Archive are to:

1. Provide a single home for biological microscopy data and facilitate the discoverability of these data.
2. Maximise the use of expensive microscopy data.
3. Ensure the reproducibility of scientific results based on biological imaging.
4. Enable new insights to be gained from existing data by encouraging their reuse.
5. Accelerate the development of image analysis methods.

To achieve these goals, the Archive needs to provide access to a wide range of bioimage data in a way that facilitates their discoverability and reuse, as the FAIR principles indicate[1]. This requires supporting straightforward direct deposition of new data, indexing, search and retrieval of datasets in reuse and visualisation-friendly formats, and integration with other data resources.

Moving towards meeting all of these goals is a gradual process that is based on iterative development and engagement with the broad bioimaging community. To meet the immediate demands of the community, while building infrastructure and processes to support long-term growth and wide data reuse, the initial focus of the Archive is on:

a) Providing a rapid and straightforward deposition process, such that submitters preparing for publication can quickly receive an accession identifier for the imaging data associated with that publication.
b) Building a diverse collection of image data, while steadily improving the quality of associated metadata.
c) Enhancing reusability of those images through discoverability based on rich metadata and easy access to both whole datasets and individual images.
d) Supporting added-value databases through resource indexing, provision of storage infrastructure, data import/export and linking between resources.

Data deposition

The BioImage Archive accepts biological imaging data associated with publications from any imaging modality, at a large range of scales from Ångströms to centimetres. The archive also accepts "reference" image datasets, where data clearly provide value beyond a single experiment or study. Data can be deposited in any format, through a lightweight submission process involving upload of data files and completion of a web form to supply appropriate metadata. Depositors receive a unique accession identifier for their data, which can be referenced in their publication.

Where a specialised resource exists that can provide added value to a particular type of image data, such as EMPIAR's curation of cryo-EM and volume EM data, or cell or tissue data (Cell IDR and Tissue IDR) the submission is redirected to that resource, and the resulting datasets indexed by the BioImage Archive.

The BioImage Archive's data collections

The BioImage Archive now indexes approximately 1100 individual datasets across its component resources, which together represent more than 1.5 petabyte of data. Collectively these datasets derive from over 5000 different publication authors. Access varies by accession, with the most popular datasets accessed several hundred times. Although newly established, a recent EBI-wide user survey identified the BioImage Archive as one of the EBI's 20 most used data resources (https://www.ebi.ac.uk/about/our-impact/impact-report-2021).

The Archive's collections are divided in three categories:

1. Datasets deposited directly into the BioImage Archive.
2. Data indexed from other imaging resources, particularly EMPIAR.
3. Data imported from other resources, including databases that are now discontinued, for example the Journal of Cell Biology DataViewer[11].

The number of datasets directly deposited to the BioImage Archive is growing rapidly (**Figure 1**). As of this writing, the BioImage Archive collection of directly submitted data currently holds 58 imaging datasets. The sizes of those datasets vary considerably, from scales of tens of megabytes to several terabytes for a single dataset. The number of individual files comprising a dataset also ranges from a single file to 1.3 million.

The majority of accessions are from light and electron microscopy, common imaging modalities. However, as the archive's broad scope would indicate, many less common technologies including Atomic Force Microscopy (AFM), Micro-CT, and Ultrasound imaging are also represented. By data volume the majority of image files are in TIFF format (approximately 75%), but formats reflect the diversity of imaging technologies with over 30 different file types in use.

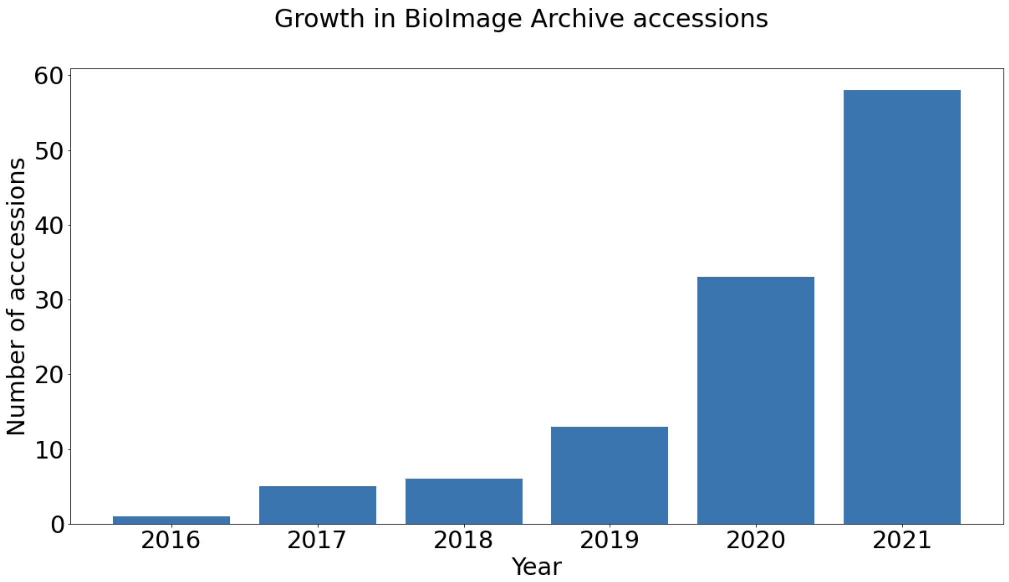


Figure 1. Cumulative growth in image dataset submissions to EMBL-EBI. Data prior to 2019 were deposited into the BioStudies database.

As part of its role in anchoring an ecosystem of bioimaging resources, the BioImage Archive indexes data deposited directly to EMPIAR, allowing search of those datasets through its web portal. Data can then be directly viewed and accessed on the EMPIAR pages. EMPIAR has also experienced very rapid growth: its data holdings passed 1 petabyte in Summer 2021.

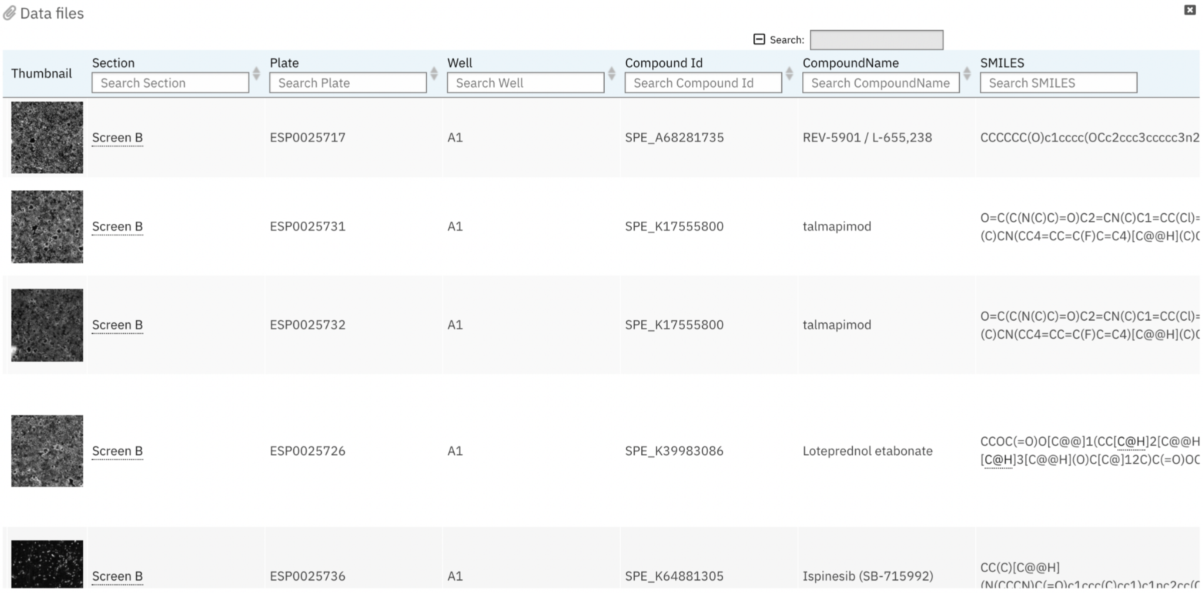
The BioImage Archive holds 424 datasets originally deposited through the Journal of Cell Biology's DataViewer application[11]. This service was discontinued in 2018, and transfer of the depositions allowed the data to be preserved for the future. This illustrates the role of the BioImage Archive in the provision of long-term sustainability for community collections of image data, where such collections are public, established and are likely to be of future scientific value.

Access, downloads and use

The BioImage Archive provides a web-based interface to browse, search, view metadata and retrieve datasets. This portal presents the dataset-level metadata associated with an accession together with the images and supporting files that accompany that accession.

A flexible search system allows both simple keyword search and construction of complex queries that enable search over a range of dataset metadata. When viewing a dataset, information associated with each individual file can be viewed in tabular form. Views can be filtered by this image-file level metadata, allowing relevant subsets of the dataset to be viewed. For example, in a high-content screen, those images associated with a specific compound in the screen can be selected and displayed.

Whole datasets, as well as individual images, can be downloaded either directly through the web interface or via the FTP or Aspera protocols. The latter two approaches are provided primarily for the bulk download of large datasets.



Thumbnail	Section Search Section	Plate Search Plate	Well Search Well	Compound Id Search Compound Id	CompoundName Search CompoundName	SMILES Search SMILES
	Screen B	ESP0025717	A1	SPE_A68281735	REV-5901 / L-655,238	CCCCCC(O)c1cccc(OCc2ccc3ccccc3n2
	Screen B	ESP0025731	A1	SPE_K17555800	talmapimod	O=C(C(N(C)C)=O)C2=CN(C)C1=CC(Cl)= (C)CN(CC4=CC=C(F)C=C4)[C@@H](C)(
	Screen B	ESP0025732	A1	SPE_K17555800	talmapimod	O=C(C(N(C)C)=O)C2=CN(C)C1=CC(Cl)= (C)CN(CC4=CC=C(F)C=C4)[C@@H](C)(
	Screen B	ESP0025726	A1	SPE_K39983086	Loteprednol etabonate	CCOC(=O)O[C@@]1(CC[C@H]2[C@@H [C@H]3[C@@H](O)C[C@]12C)C(=O)OC
	Screen B	ESP0025736	A1	SPE_K64881305	Ispinesib (SB-715992)	CC(C)[C@@H] (N(CCCN)C(=O)c1ccc(C)cc1)c1nc2cc(C

Figure 2. Images can be searched, filtered and downloaded based on their image-specific metadata content (sample from S-BIAD29, https://www.ebi.ac.uk/biostudies/BioImages/studies/S-BIAD29).

At time of writing, the BioImage Archive website is accessed by approximately 2000 different users (measured by unique IP addresses) per month, with visitors coming from a wide range of geographical locations.

BioImage Archive ecosystem

Beyond its immediate role in the provision of a service for biological image data deposition, indexing and retrieval of biological imaging datasets, the BioImage Archive is developed to support added-value imaging data resources. The two resources that work most closely with the BioImage Archive are EMPIAR and IDR. We expect that in future other imaging data resources can use this service.

Initially founded to provide a home for raw images underpinning 3D cryo-EM maps and tomograms, EMPIAR has expanded to cover volume EM and 3D X-ray imaging data. In addition to the indexing of EMPIAR datasets described above, the BioImage Archive provides capacious object storage to EMPIAR. The two resources also work together to support the deposition of correlative imaging data[12] where different modalities are used to image the same biological specimen. In an example of this process, soft X-ray tomography images are deposited into EMPIAR, the corresponding light microscopy data are deposited into the BioImage Archive, and the datasets are linked together. Information about the physical transformations required to map the datasets into a common coordinate space are included as part of the BioImage Archive deposition.

As mentioned above, IDR is a platform for bioimage data integration and reanalysis. It runs on EMBL-EBI's Embassy science cloud system (ww.embassycloud.org)[13]. The BioImage

Archive provides the IDR with guarantees of long-term sustainability for data in its collections through import of data from the IDR. Work is also underway to develop a submission mechanism allowing data to be submitted to the IDR through the BioImage Archive, such that submitters can receive an accession identifier quickly for immediate publication of their results, while suitable reference datasets can benefit from the curation and data enrichment provided by IDR.

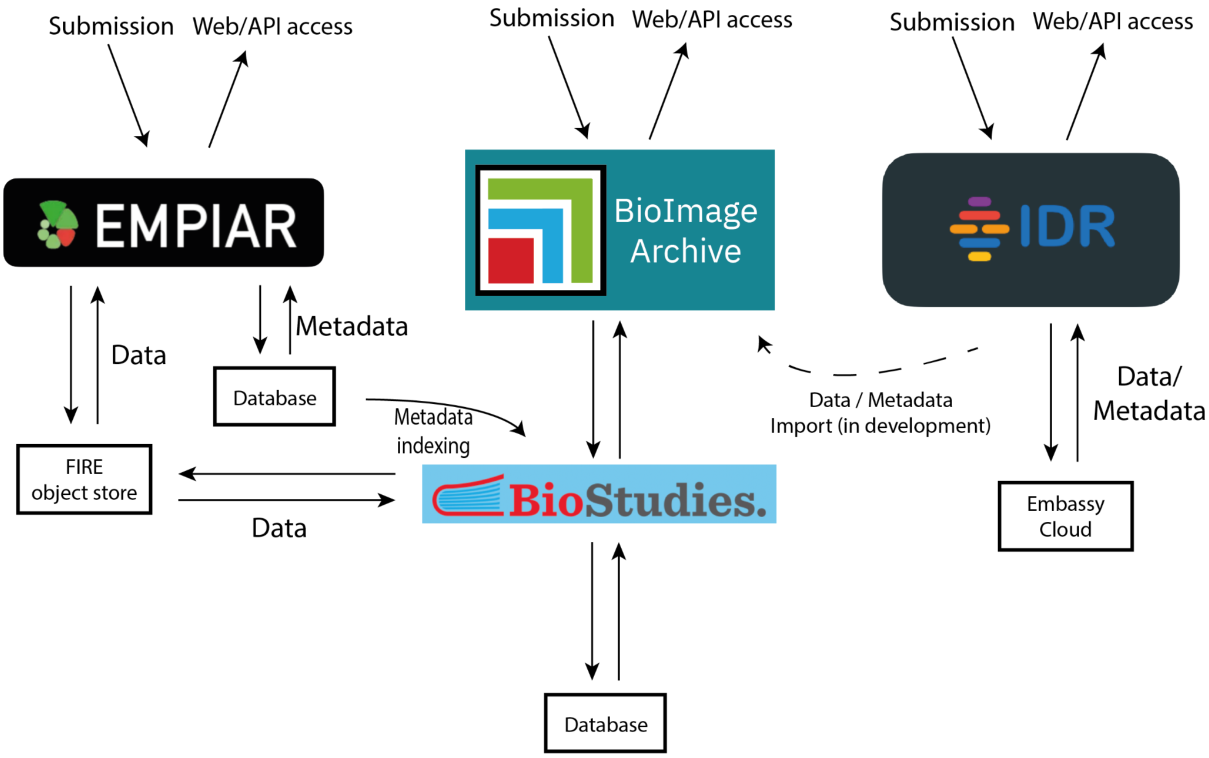


Figure 3 - The BioImage Archive interacts with other image data resources through shared data stores, indexing and cross-linking.

Sustainability

The infrastructure underlying the BioImage Archive is funded by the UK Research and Innovation Strategic Priorities Fund, while staff is funded from EMBL core resources. EMBL-EBI is committed to maintain the BioImage Archive as an essential part of the collection of the EMBL-EBI core resources. This is critical for the continued growth and operation of added value resources (in particular, IDR, EMPIAR) that use the Archive as an underlying foundation to support their separate funding efforts. In this way, the Archive functions as the foundation for a growing bioimaging data ecosystem[4].

Future plans

One of the BioImage Archive's core aims is to maximise the reuse of imaging data. To use the data, it must first be located, and as the archive's collections expand, rich metadata accompanying the data will be increasingly necessary to support data search and comparison[14]. The recently released REMBI metadata model[16] provides draft metadata guidelines to cover different modalities in biological imaging.

The BioImage Archive will implement REMBI, while ensuring that submitters can still meet their need for rapid deposition of data. Several options exist to support this process, including automated metadata extraction ("harvesting") and integration with local data-management solutions such as OMERO[17]. We will also integrate emerging community standards, such as the recently proposed 4DN-BINA extensions to the existing OME metadata model[18]**.** Where possible we will support importing metadata from external applications, such as tools designed to capture image-acquisition metadata at the time of imaging[19].

Supporting the interactive exploration, visualisation and reuse of large imaging datasets is a key long-term goal for the BioImage Archive. These aims are difficult to meet when data are represented across many different file formats and file sizes, and this heterogeneity also presents a significant barrier to reuse. To overcome this barrier requires providing access to data in file formats that are standardised, scalable and optimised for large-scale distribution[20], in particular by supporting random access to subsets of data and parallel access. At least initially, forcing data depositors to engage in complex data conversion will discourage submissions; therefore, format conversion will need to be carried out within the Archive. If community adoption of new file formats grows, the Archive can work with commercial imaging system vendors to add support for these formats in their acquisition control and analysis software systems.

Modern deep-learning-based AI techniques have brought rapid progress to the analysis of biological images[21]. These technologies often rely on large corpora of well-annotated reference data to train or retrain models. The BioImage Archive has great potential to accelerate developments in this field by providing these reference datasets to the community. This requires work to define suitable common formats for annotations, develop deposition pipelines and encourage and support the community in their use.

Finally, the BioImage Archive will work to grow the community of added-value databases, by identifying technical or biological domains where specific communities can have their needs served, as well as supporting easy integration for existing or emerging resources. In the long term, the BioImage Archive will need to work globally with the international bioimaging community to establish common standards and distribute data across multiple locations[22]. Similar models involving international consortia support large-scale archives like the ENA and PDB[23].

Discussion

Although a young resource, the BioImage Archive is already experiencing fast growth. Early uptake has been very rapid, with the demand for fast deposition of data to receive an identifier for publication (driven by journal requirements) a notable factor driving submission.

Key to this growth is the BioImage Archive's ability to fill a major gap in the provision of broad modality image data deposition and open access to image data. As open access deposition of images supporting scientific results increasingly becomes a requirement for publication, the role of the BioImage Archive in providing this service while encouraging data reuse will grow. As the Archive's collections expand, ensuring that the quality of submitted data and metadata increases while maintaining straightforward data deposition will be a key priority. Experience from other data resources shows that when this growth is managed appropriately, the results are of great value to the scientific community.

The challenges of scaling a biological image archive with a very wide scope are considerable. The wide range of different technologies and approaches that together comprise biological imaging gives rise to significant heterogeneity in data types and file formats. Flows of data ingest, linking patterns and methods consumption for high-content screening experiments, developmental biology assays or protein interaction studies, for example, are all very different. A further challenge arises when different modalities of image data must be integrated, such as in correlative imaging[12], or when other forms of data (such as genomic) are required to give context to images. Through careful selection of the right level of metadata, integration with local data management solutions and new file formats, we hope to meet these challenges.

A broad repository of open-access FAIR[1] image data has huge potential to accelerate research in the life sciences. We envision a future in which the BioImage Archive anchors a wide range of community resources representing multiple biological and technical domains, supporting curation and reuse. Together, these would unlock the potential in existing image data, making the investment in biological imaging come truly alive.

Materials and Methods

Submission process

Submission to the BioImage Archive involves four stages:

1. Preparation for submission, organising data and registering an account if needed.
2. Upload of data and preparation of file-level metadata.
3. Completion of dataset-level metadata.
4. Finalisation of submission.

Submitters organise their data locally before submission, by selection of a suitable file and directory structure. The BioImage Archive allows considerable flexibility in the structure of submitted datasets. This flexibility is necessary to allow the archive to support the wide range of imaging modalities and experimental setups, as well as to allow submission of intermediate and downstream analysed data for which no general formal structure exists.

Submitters then upload their image and supporting data files. For smaller datasets, this can be done through the submission tool, which provides a graphical interface for file uploads. For larger datasets, FTP (File Transfer Protocol) and Aspera (specialised software for accelerated file transfer) uploads are supported. These large datasets may take several days to upload, so that planning depositions ahead of time is recommended particularly if publication is dependent on data release.

They then complete a web form which allows metadata about the images and associated data to be provided. This metadata includes information about the submitting team, the study-associated publication that the image data support, experimental and imaging protocols and summary information about the images themselves. In future, the REMBI metadata schema will extend this.

After submission, an identifier is assigned. Submission can take place in as little as one day, though the limiting factor is usually data-transfer time.

Data access

Access to images and other files is provided both directly through the web portal, and via an API. Because the BioImage Archive does not currently convert between formats and allows submission in any data format, it cannot provide visualisation of images. Images can be downloaded directly from the web portal via the HTTP/HTTPS protocol, as well as via FTP or Aspera. Globus support is planned for future development.

The API supports programmatic access to both data and metadata. This allows enumeration of datasets, keyword and advanced search, retrieval of individual image-level metadata and download of individual files.

Relation to BioStudies

The BioStudies database enables authors to package all data supporting a publication, through both direct data hosting and integration of links to other data sources[10]. At EMBL-EBI, BioStudies fills a key role of providing a home for data that do not fit into existing structured archives. This role enabled BioStudies to provide an initial service for storage and indexing of image data at EMBL-EBI (giving rise to the pre-2019 submissions in Figure 1).

The BioImage Archive builds on this foundation provided by BioStudies, using it as a platform. Data submission, search and access processes are customised for the BioImage Archive. Future work will extend this customisation to include extraction of metadata from images, specific submission workflows and templates for different data types, and other features enabling easier data submissions and access.

System architecture

The three main components of the system are its backend that provides data and metadata management and lifecycle services, the submission tool that enables dataset deposition, and the data access interface for both web and programmatic access.

During the submission process, data initially resides on fast but transient filesystem storage. This allows submitters to assemble their data, and for extra actions (compression for example) to be applied to the data prior to completion of submission. Data files can be transferred via HTTP, FTP or FASP protocols. Metadata can be provided via web forms of the submission tool, and advanced users can also format metadata as JSON, XML or tabdelimited files. The submission tool uses the Angular web application framework and is written in TypeScript.

When submission is completed, files are uploaded to FIRE (FIle REplication), EMBL-EBI's very large-scale object data storage system. This provides long-term sustainable storage, operational redundancy, and backup to tape. Dataset level metadata are stored in a MongoDB database. The system backend is coded in Kotlin.

The database is indexed nightly to power the search engine that is a part of the data access application and exported to allow distributed data hosting. The search engine is built using Apache Lucene, and the data access system is a Java web application.

Acknowledgements

We thank all of the community representatives that participated in the workshops that established the scientific rationale for the BioImage Archive and the REMBI metadata framework. We also thank the IDR, EMPIAR, BioStudies and BioImage Archive teams, EMBL-EBI's Bioimaging project team and the storage team within EMBL-EBI's Technical Services Cluster. Finally, we would like to thank all the scientists, past, present and future, who deposit their data in the BioImage Archive.

Funding statement

The development of BioImage Archive has been supported by European Molecular Biology Laboratory member states, Wellcome Trust grant 212962/Z/18/Z and BBSRC grant BB/R015384/1. Work on EMPIAR was funded from 2014 to 2021 by two project grants awarded to EMBL-EBI by MRC and BBSRC (MR/L007835/1 and MR/P019544/1). From 2021, EMPIAR benefits from funding from the Wellcome Trust (221371/Z/20/Z). The EMBL-EBI IT infrastructure supporting the BioImage Archive is funded by the UK Research and Innovation Strategic Priorities Fund.

References

1. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
2. Harrison, P. W. *et al.* The European Nucleotide Archive in 2020. *Nucleic Acids Res.* **49**, D82–D85 (2021).
3. wwwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **47,** D520-D528 (2019).
4. Ellenberg, J. *et al.* A call for public archives for biological image data. *Nat. Methods* **15**, 849–854 (2018).

5. Rung, J. & Brazma, A. Reuse of public genome-wide gene expression data. *Nat. Rev. Genet.* **14**, 89–99 (2013).

6. Iudin, A., Korir, P. K., Salavert-Torres, J., Kleywegt, G. J. & Patwardhan, A. EMPIAR: a public archive for raw electron microscopy image data. *Nat. Methods* **13**, 387–388 (2016).

7. Williams, E. *et al.* Image Data Resource: a bioimage data integration and publication platform. *Nat. Methods* **14**, 775–781 (2017).

8. Tohsato, Y., Ho, K. H. L., Kyoda, K. & Onami, S. SSBD: a database of quantitative data of spatiotemporal dynamics of biological phenomena. *Bioinformatics* **32**, 3471–3479 (2016).

9. Scheffer, L. K. *et al.* A connectome and analysis of the adult Drosophila central brain. *eLife* **9**, e57443 (2020).

10. Sarkans, U. *et al.* The BioStudies database—one stop shop for all data supporting a life sciences study. *Nucleic Acids Res.* **46**, D1266–D1270 (2018).

11. Hill, E. Announcing the JCB DataViewer, a browser-based application for viewing original image files. *J. Cell Biol.* **183**, 969–970 (2008).

12. Iudin, A. *et al.* Data-deposition protocols for correlative soft X-ray tomography and super-resolution structured illumination microscopy applications. *STAR Protoc.* **2**, 100253 (2021).

13. Cook, C. E. *et al.* The European Bioinformatics Institute in 2016: Data growth and integration. *Nucleic Acids Res.* **44**, D20–D26 (2016).

14. Linkert, M. *et al.* Metadata matters: access to image data in the real world. *J. Cell Biol.* **189**, 777–782 (2010).

15. Eng, E. T. *et al.* Reducing cryoEM file storage using lossy image formats. *J. Struct. Biol.* **207**, 49–55 (2019).

16. Sarkans, U. *et al.* REMBI: Recommended Metadata for Biological Images—enabling reuse of microscopy data in biology. *Nat. Methods* (2021) doi:10.1038/s41592-021-01166-8.

17. Allan, C. *et al.* OMERO: flexible, model-driven data management for experimental biology. *Nat. Methods* **9**, 245–253 (2012).
18. Hammer, M. *et al. Towards community-driven metadata standards for light microscopy: tiered specifications extending the OME model*. 2021.04.25.441198 https://www.biorxiv.org/content/10.1101/2021.04.25.441198v3 (2021) doi:10.1101/2021.04.25.441198.
19. Huisman, M. *et al.* A perspective on Microscopy Metadata: data provenance and quality control. *ArXiv191011370 Cs Q-Bio https://arxiv.org/abs/1910.11370* (2021).
20. Moore, J. *et al. OME-NGFF: scalable format strategies for interoperable bioimaging data*. *Nat Methods* (2021) https://doi.org/10.1038/s41592-021-01326-w.
21. Berg, S. *et al.* ilastik: interactive machine learning for (bio)image analysis. *Nat. Methods* **16**, 1226–1232 (2019).
22. Swedlow, J. R. *et al.* A global view of standards for open image data formats and repositories. *Nat. Methods* (2021) doi:10.1038/s41592-021-01113-7.
23. Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.* **10**, 980 (2003).