doi:10.1017/S1431927622006134
Microsc. Microanal. 28 (Suppl 1), 2022

Towards Public Archiving of Large, Multi-Modal Imaging Datasets

Matthew Hartley[1*], Gerard Kleywegt[1]

[1.] European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridgeshire, UK
* Corresponding author: matthewh@ebi.ac.uk

Reproducibility of results is a fundamental pillar of scientific integrity. Biological images are a key data source in biology, often requiring substantial computational analysis and integration with other data to generate useful quantitative information. Providing access to raw image data underlying published results is thus critical to scientific progress. Beyond improving reproducibility, open access to well organised image data in standard formats, following the FAIR principles[1], can drive methods development and allow new discoveries to be made. For these reasons funders, publishers and scientific institutions increasingly expect that researchers will make their data openly available at some point in the experimental lifecycle.

Publishing large, multimodal datasets, however, involves several challenges. Biological imaging covers an extremely diverse set of subdomains, each with its own standards, practises, and communities. This gives rise to huge heterogeneity in data types, formats, and metadata standards. Additionally, the scale of data generation varies over multiple orders of magnitude and dealing with very large datasets (which can be 100s of TB in size) is particularly challenging. Committing to storing and providing access to these kinds of datasets for the long term is difficult, particularly for smaller institutions and laboratories. Multimodal datasets amplify these problems since they involve different types of imaging and require additional consideration to how different modalities must be integrated.

The European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) is the "home for big data in biology" in Europe. EMBL-EBI operates two data resources dedicated to imaging data, which together provide public archival services across imaging modalities. EMPIAR, the Electron Microscopy Public Image Archive, was originally launched in 2013 to archive raw 2D image data supporting cryo-EM and electron cryo-tomography derived 3D structures[2]. EMPIAR's scope has now expanded to include volume EM and soft and hard X-ray imaging data. The BioImage Archive[3] (BIA), launched in 2019, accepts all other modalities of biological imaging data associated with a publication, except for patient-identifiable medical data.

Together, EMPIAR and the BIA support the archival of multimodal imaging data. Currently, the components of a correlative study are deposited separately in the two archives, with linking information, such as transformations mapping the different components into a common physical coordinate space, deposited into the BioImage Archive[4]. Each deposition is annotated with links to the other components of the study. The archives are hosting datasets from the COMULIS community challenge (https://www.comulis.eu/challenges) to provide further demonstrations of this approach.

Further developments of both resources will allow better integration, more possibilities for interactive exploration of multimodal data, and easier data reuse. The recently released Recommended Metadata for Biological Images (REMBI) guidelines[5] provide a common standard for metadata across different imaging modalities. Implementing this standard will provide a more consistent experience for depositors

https://doi.org/10.1017/S1431927622006134 Published online by Cambridge University Press

and users of multimodal data, as well as improving usability of archived data for example through development of standardised representations for coordinate transforms within REMBI's correlative information module. New advances in file formats designed to provide random/streaming access to subcomponents of large datasets will improve the accessibility of these large datasets. This will allow better interactive visualisation and annotation tools, both those provided by the archives and third-party software.

References:

[1] M. D. Wilkinson et al., 'The FAIR Guiding Principles for scientific data management and stewardship', Sci. Data, vol. 3, p. 160018, Mar. 2016, doi: 10.1038/sdata.2016.18.

[2] A. Iudin, P. K. Korir, J. Salavert-Torres, G. J. Kleywegt, and A. Patwardhan, 'EMPIAR: a public archive for raw electron microscopy image data', Nat. Methods, vol. 13, no. 5, pp. 387–388, May 2016, doi: 10.1038/nmeth.3806.

[3] M. Hartley, G. Kleywegt, A. Patwardhan, U. Sarkans, J. R. Swedlow, and A. Brazma, 'The BioImage Archive - home of life-sciences microscopy data'. bioRxiv, p. 2021.12.17.473169, Dec. 21, 2021. doi: 10.1101/2021.12.17.473169.

[4] A. Iudin et al., 'Data-deposition protocols for correlative soft X-ray tomography and super-resolution structured illumination microscopy applications', STAR Protoc., vol. 2, no. 1, p. 100253, Mar. 2021, doi: 10.1016/j.xpro.2020.100253.

[5] U. Sarkans et al., 'REMBI: Recommended Metadata for Biological Images—enabling reuse of microscopy data in biology', Nat. Methods, vol. 18, no. 12, Art. no. 12, Dec. 2021, doi: 10.1038/s41592-021-01166-8.

https://doi.org/10.1017/S1431927622006134 Published online by Cambridge University Press