doi:10.1017/S1431927621009892
Microsc. Microanal. 27 (Suppl 1), 2021

Image archiving at EMBL-EBI - EMPIAR and the BioImage Archive

Gerard Kleywegt

EMBL-EBI, United States

The past decade has seen spectacular developments in the fields of both light (LM) and electron microscopy (EM), with Nobel Prizes being awarded for super-resolution LM and cryo-EM. These techniques enable scientists to gain insights at scales from individual atoms to entire organisms.

These new microscopy and other imaging methods generate huge amounts of data that need to be processed, visualised, analysed and interpreted. There is a strong case to be made that these data should be publicly archived, including:

- to allow independent quality assessment and validation of the analysis and interpretation reported in the scientific literature;

- the sample and/or data may be unique or difficult to reproduce;

- the data can be re-used to develop, test, benchmark or improve new software and methods;

- the data may be re-used, e.g. in machine-learning (ML) applications.

Some of these benefits are closely tied to the publication (e.g., validation), whereas other may be realised much later. In fact, experience with archives in other areas, such as the Protein Data Bank (PDB; established in 1971) in structural biology, has shown that data may be re-used decades after it was generated for applications that couldn't even be imagined at the time.

A number of initiatives are underway, in Europe, Asia and the USA, to archive bioimaging data. Two of these efforts, both at the EMBL-European Bioinformatics Institute (EMBL-EBI, Cambridge, UK), will be discussed.

EMPIAR (EM Public Image ARchive; https://empiar.org/; ref [1]) was established in 2013, originally with the intention to archive the raw 2D image data underpinning cryo-EM and electron cryo-tomography studies (the Electron Microscopy Data Bank, EMDB, has provided a home for the processed EM maps and tomograms since 2002). The idea for this effort was an outcome of a workshop for cryo-EM community experts about data handling, and it was fortuitously timed to be in place when the "resolution revolution" in this field took off (2014). Since then, the archive has grown in both size (EMPIAR is expected to reach one Petabyte in the Spring or early Summer of 2021) and scope. Nowadays, there are archived data sets from volume EM (vEM) techniques, soft X-ray tomography, electron crystallography, etc. A major focus for the next five years will be on developing extensive support for the vEM community. In addition, EMPIAR's scope is expected to be broadened further, e.g. covering data from X-ray imaging, as well as ML benchmark datasets and results of Citizen Science projects.

A more recent effort is the establishment in 2019 of a BioImage Archive at EMBL-EBI (https://www.ebi.ac.uk/bioimage-archive; ref [2]). This initiative comprises two components. One is the establishment of storage and other IT infrastructure to support archiving of large amounts of bioimaging data. This infrastructure is designed to support an ecosystem of imaging resources, meaning that these resources do not have to acquire and maintain their own infrastructure. For example, EMPIAR stores and serves all its data in the object storage provided by the BioImage Archive. The other part is an actual archive of imaging data for which no dedicated archive exists (for instance, Cryo-EM and vEM data will continue to be stored in EMPIAR, but many kinds of LM data will be in scope for the BioImage Archive).

https://doi.org/10.1017/S1431927621009892 Published online by Cambridge University Press

The history, scope, applications and future plans of EMPIAR and the BioImage Archive will be discussed at the meeting.

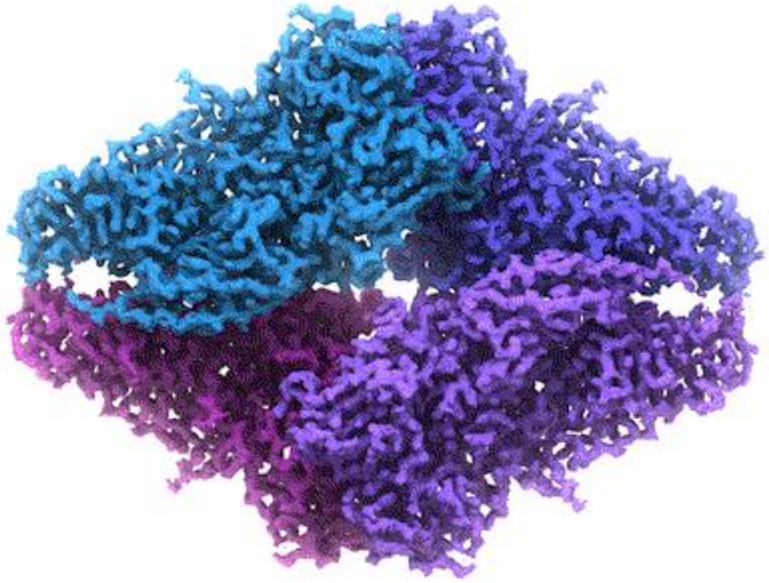

Figure 1. In 2015, a breakthrough cryo-EM structure was published of beta-galactosidase at 2.2Å resolution, a record at that time. The 12.4 TB raw data set (also a record for EMPIAR at that time) was deposited in EMPIAR (https://empiar.org/10061) and has been used many times since then. Several groups (including the original authors) have reprocessed the data to even higher resolution (and in some cases, published and deposited those improved maps). Developers of cryo-EM data-processing software have used it to test and improve their algorithms and several deep-learning methods for automated particle picking benefitted from it as well. For more details, see https://empiar.org/reuse

References

[1] Iudin, A., et al., EMPIAR: a public archive for raw electron microscopy image data. Nat Methods, 2016. 13(5): p. 387-388.

[2] Ellenberg, J., et al., A call for public archives for biological image data. Nat Methods, 2018. 15(11): p. 849-854.

[3] Bartesaghi, A., et al., 2.2 A resolution cryo-EM structure of beta-galactosidase in complex with a cell-permeant inhibitor. Science, 2015. 348(6239): p. 1147-51.

https://doi.org/10.1017/S1431927621009892 Published online by Cambridge University Press