Bioimaging Data

Introduction

Bioimaging specialists are acquiring an ever-growing amount of data: images, associated metadata, etc. However, image data management often does not receive the attention it requires or is avoided altogether since it is considered burdensome. At the same time, storing images on personal computers or USB keys is no longer an option, assuming it ever was!

Data volume is exponentially increasing, and not just the acquired images need storing, but potentially processed images will be generated and must be kept alongside the original images. It is critical to proactively identify where the data will be stored, for how long, who will cover the hardware cost, and who will manage the infrastructure.

All the stakeholders need to be involved in the preliminary discussions: biologists, facility managers, data analysis, IT support, etc., to ensure the requirements are understood and met.

What constitutes Bioimage Data

An image is much more than a collection of zeros and ones. The image will contain the binary representing the pixels on the screen, but it is usually packed with helpful metadata. You will find the apparent keys indicating how to interpret the zeros and ones; you can also find a lot of acquisition metadata, e.g., hardware/instrument used, settings used, etc.

The number of image proprietary formats is very large and keeps increasing. It is challenging to support many proprietary file formats, i.e., read/extract metadata. The Bio-formats library currently supports over 150 different file formats. The Dataset Structure Table shows the extension of the files to read and indicates the structure of the image, e.g., a single file, multiple files, one image file, and a companion file, etc.

Data management challenges

The number of files and their size could be extremely large. Deleting/misplacing a file could invalidate the study, preventing its reuse.

Managing images immediately becomes a more significant problem; not only the binary files need to be handled, but also the associated metadata. Several efforts have been made and are still ongoing to capture that metadata. Understanding and capturing the metadata is critical for many reasons, including analyzing and detecting possible faults in acquisition systems. Deciding how many details will be recorded is essential since this could dramatically increase the metadata volume and, therefore, the effort required to capture the metadata.

The collection of images could be:

- data acquired within a facility;
- data acquired in another facility (commissioned work or external guest user) and "transported" by the users to their facility;
- slides scanned.

After the acquisition, data are usually moved to more permanent storage with different levels of permissions. This depends on the facility policies and could prevent collaborative work. Users will also adopt their own "organization" conventions; this could make it very difficult to find or understand the data when, for example, the data are migrated to a new location or when the researcher who acquired the data leaves the lab.

Standard (meta)data formats

Description

Unlike other domains, the bioimaging community has not yet agreed on a single standard data format generated by all acquisition systems. Instead, the images described above are most frequently collected in proprietary file formats (PFFs) defined by hardware vendors. Currently, there are several hundred such formats that the researchers may encounter. These formats combine critical acquisition metadata with multidimensional binary data but are often optimized for *quickly* writing the data to disk. Tools and strategies are outlined below to ease working with this data.

Considerations

- Consider carefully how the resulting files will be processed when purchasing a microscope. If open-source tools are used, proprietary file formats may require a time-consuming conversion. Discuss with your vendor if an open format is available.

- If data from multiple vendors are combined, a similar conversion may be necessary to make the data comparable.

- Imaging data brings special considerations due to the data's extensive, often continuous nature. Single terabyte-scale files are not uncommon. Sharing these can require a particular infrastructure, like a data management server (described below) or a cloud-native format (described below). One goal of such infrastructure is to enable the selective (i.e., interactive) zooming of your image data *without downloading* the entire volume, thereby reducing your internet bandwidth and costs.

- Notably, most acquisition systems produce proprietary file formats. Understanding how well the imaging community supports them could be vital to a successful study. Will it be possible to analyze or view the image using open-source software? Will it be possible to deposit the images to public repositories when published? The choice of proprietary file formats could prevent from using any other tools that are not related to the acquisition systems.

Solutions

Vendor libraries; Some vendors provide open-source libraries for parsing their proprietary file formats. See libCZI from Zeiss.

Open-source translators: community members have developed multi-format translators that can be used to access your data on-the-fly, i.e., the original format is preserved, and no file is written on disk. This implies that you will need to perform this translation each time you access your data and depending on the size of the image(s), you could run out of memory. Translation libraries include,

- Bio-Formats (Java) - supports over 150 file formats

- OpenSlide (C-r-i-) - primarily for whole-slide imaging (WSI) formats

- aicsimageio (Python) - wraps vendor libraries and Bio-Formats to support a wide range of formats in Python

Permanent conversion; An alternative is to convert your data permanently

- OME-Files - The Open Microscopy Consortium (OME) has developed an open format,"OME- TIFF," to which you can convert your data. The Bio-Formats (above) library comes with a command line to tool bfconvert that can be used to convert files to OME-TIFF

- The bioformatsZraw and rawZometiff toolchain provided by Glencoe Software allows the more performant conversion of your data but requires an extra intermediate copy. If you have available space, the toolchain could also be an option.

Cloud (or "object") storage; If you are storing your data in the cloud, you will likely need a different file format since most current image file formats are unsuitable. OME is currently developing a next-generation file format (NGFF) that you can use.

Metadata; If metadata are stored separately from the image data, the format of the metadata should follow the subject-specific standards regarding the schema, vocabulary or ontologies, and storage format used, such as:

- OME model XML-based representation of microscopy data.
- Quality assessment and Reproducibility in Light Microscopy (QUAREP-LiMi).
- REMBI.

(Meta)Data collection

Description

The acquisition of bioimaging data takes place in various environments. The (usually) light or electron microscope may be in a core facility, research lab, or even remotely in a different institution. Regardless of the instrument's location, the acquired imaging data will likely be stored temporarily in a local, vendor-specific system's PC next to the acquisition system due to its complexity and size. It is often unavoidable to securely store the data as quickly as the acquisition process.

Due to the scale of data, keeping track of the image data and the associated data and metadata is essential, particularly in the life sciences and medical fields. Organizing, storing, sharing, and publishing image data and metadata can be challenging.

Considerations

- Consider using an image management software platform. Image management software platforms offer a way to centralize, organize, view, distribute, and track all of their digital images and photos. It allows you to control how your images are managed, used, and shared within research groups.

- When evaluating an image management software platform, check if it allows you to:

    - Control the access you wish to give to your data and how you wish to work, e.g., PI only can view and annotate my data, or you can choose to work on a project with some collaborators.
    - Access data from anywhere via either Web or Desktop clients and API.

    - Store the metadata with your images. For example, analytical results can be linked to your imaging data and can be easily findable.
    - Add value to your imaging data by, for example, linking them to external resources like ontologies.
    - Make your data publicly available and slowly move towards FAIRness.

- Try to avoid storing bioimaging data in the local system's PC.
- If possible, make a transfer to central storage mandatory. If not possible, enable automation of data backup to the central repository.
- Consider your domain's support for minimal standards (metadata schemas, file formats, etc.).
- Consider reusing existing data.

Solutions

- Agnostic platforms that can be used to bridge between domain data include:
    - iRODS.
    - bZshare.

- Image-specific data management platforms include:
    - OMERO - broad support for a large number of imaging formats.
    - Cytomine-IMS - image specific.
    - XNAT - medical imaging platform, DICOM-based.
    - MyTardis- largely file-system-based platform handling the transfer of data.
    - BisQue - a resource for the management and analysis of 5D biological images.

- Platforms like OMERO and bZshare also allow you to publish the data associated with a given project.

- Metadata standards can be found at the Metadata Standards Directory Working Group.

- Ontologies Resources available at:
    - Zooma - Resource to find ontology mapping for accessible text terms.
    - Ontology Search - Ontology lookup service.
    - BioPortal - Biomedical ontologies.

- Existing data can be found by using the following resources: - LINCS. - Research Data repositories Registry.

Data publication and archiving

Description

Public data archives are an essential component of biological research. However, publishing image data and metadata can be very challenging for multiple reasons, to mention a few: limited infrastructure for some domains, data support, and sparse data.

Bioimaging tools and resources are behind compared to what is available in sequencing, mainly due to limited infrastructures capable of hosting the data. There are a few ongoing efforts to breach that gap.

Two distinct types of resources should be considered:

- Data archives ("storage") are long-lasting storage for data and metadata, making those data easily accessible to the community.
- Added-values archives: store enhanced curated data, typically aiming at a scientific community.

Considerations

- If you only need to make your data available online and have limited metadata associated, consider publishing in a **Data archive.**

- Consider an Added values archive if your data should be considered as a reference dataset**.**

- Select and choose the repositories based on the following characteristics:
  - Storage vs. Added-value resources.
  - Images format support.

- Supported licenses, e.g., CCO or CC-BY license. For example, the Image Data Resource (IDR) uses Creative Commons Licenses for submitted datasets and encourages submitting authors to choose.
  - Which types of access are required for the users, e.g., download only, browse search and view data and metadata, API access?
  - Does an entry have access, e.g., idr-xxx, EM PI AR-#####?
  - Does an entry have a DOI (Digital Object Identifier)?

Solutions

Comparative table of some repositories that can be used to deposit imaging data:

Repository	Type	Data Restrictions	Data Upload Restrictions	DOI	Cost
BioImageArchive	Archive	No PIH data	None	—	Free
Dryad	Archive	No PIH data	300GB	Yes	over 50GB (*)
EMPIAR	Added-value	Electron microscopy imaging data	None	Yes	Free
IDR	Added-value	Cell/Tissue imaging data, no PIH data	None	Yes	Free
SSBD: database	Added-value	Biological dynamics imaging data	None	—	Free
SSBD: repository	Archive	Biological dynamics imaging data	None	—	Free
Zenodo	Archive	None	50GB per dataset	Yes	Free

- PIH: Protected health information.
- (*) unless the submitter is based at a member institution.