Global BioImaging Project

D4.3 Publication of international recommendation: image data standards and open access repositories

Project N.	653493
Project Title	Global BioImaging
Project Acronym	GBI
Associated Work Package	WP4
Associated Task	Task 4.3
Lead Beneficiary (short name)	UNIVDUN
Nature	Report
Dissemination Level	Public
Estimated Delivery Date (Grant Agreement, Annex I)	30/11/2018
Actual Delivery Date	30/11/2018
Task leader	Jason Swedlow
Contributors	Claire M. Brown; Shuichi Onami; Graham Galloway; Wojtek Goscinski; Pasi Kankaanpää; Ryan Sullivan; Chris Wood.

Funded by the Horizon 2020 Framework Program of the European Union

Abstract

Biological and biomedical imaging datasets record the constitution, architecture and dynamics of living organisms across several orders of magnitude of space and time. Several technologies have now matured so that routine publication of these datasets is now possible. Participants in Global BioImaging from 15 countries and all populated continents have agreed on the need for recommendations and guidelines for the establishment of image data repositories and the formats they use for delivering data to the global scientific community. This deliverable summarizes work by GBI members in defining these guidelines and our future work in this area.

The following international GBI partners contributed to this recommendation:

Country	Name	Affiliation
UK	Jason Swedlow	Euro-BioImaging; Dundee University
Finland	Pasi Kankaanpää	Euro-BioImaging; Turku University
UK/EMBL	Ugis Sarkans	EMBL
Australia	Wojtek Goscinski	Monash University, MASSIVE and CVL
Australia	Graham Galloway	University of Queensland, National Imaging Facility
Australia	Ryan Sullivan	University of Sydney, Microscopy Australia
Canada	Claire M. Brown	McGill University
Japan	Shuichi Onami	RIKEN, ABiS
Mexico	Chris Wood	UNAM

Table of Contents

1. Introduction

Imaging is now used globally as a method of recording quantitative measurements of biological and biomedical structure, constitution and dynamics in the life and biomedical sciences. A consistent theme of discussion in Global BioImaging's Exchange of Experience meetings has been the development of standards for image data formats and public data resources. To properly address the concerns of the global imaging community and also to leverage the cooperation and commitment that has emerged in Global Bioimaging, the project's participants intend to develop recommendations for image data formats, standards and recommendations for public data repositories. This document defines our understanding of the requirement for such resources and the target audience for these standards. These should inform future work by Euro-BioImaging, OME, BIDS, SSBD, the Characterisation Data-enhanced Virtual Laboratory, Trusted Data Repositories and other projects that have contributed to Global BioImaging as they develop tools, interfaces, standards and data resources for the bioimaging community worldwide. The present deliverable will also be the starting point for the future work of the Global BioImaging network, which has established a Working Group to continue working on data management (with the final goal to move towards the identification of interoperability of systems).[1]

2. Guidelines for Standardized Formats

The huge range of modalities and applications for imaging reflects the incredible spread and dominance of imaging as a critical scientific technology in the physical, biological and biomedical sciences. This diversity demonstrates the power of imaging but also creates several technical problems. In particular, the huge number of data formats that are used across many different modalities inhibits access to and exchange of datasets between scientists in collaborative projects, across different imaging applications and research domains.

It is impractical to suppose or recommend that a single data format can satisfy the wide range of imaging applications covered by the Global Bioimaging community. Thus, we have developed a series of specifications and recommendations for potential standards that Global Bioimaging and imaging scientists in general may adopt and use in the future. These recommendations are built upon the successful use of standards in various imaging communities for example DICOM, OME-TIFF, imzML, Nifti, NRRD and many others. These various community standards have had varying success depending on the quality of the implementation and the ongoing maintenance of the format. These various examples present an opportunity to learn from past successes and failures. They also provide a strong set of recommendations for defining and adopting standards within the Global Bioimaging community. In the sections below we detail our current level of experience and recommendations for implementing and adopting standards for imaging data.

3. Data Repositories

Commonly shared open datasets have repeatedly proven to be essential for the development of analytic and processing tools for data across the sciences. Open science initiatives are becoming more widely accepted by the scientific community and open access to research data is now often required by funding agencies. In the life and biomedical sciences, the commitment of the

[1] See D2.4 "Sustainable plan for funding future activities of Global BioImaging including reciprocal use, training, virtual platforms for data handling"

genomics community to rapidly publish genomic sequence data (DOI: 10.1007/s10739-018-9538-7) was the basis of the development and growth of the modern field of bioinformatics. Global Bioimaging aims to catalyze a similar development of bioimage informatics and data analytics by encouraging and supporting the construction, sustainability and continuous availability of repositories for imaging data.

The Global Bioimaging community aims to adopt the principles and methods for the construction and operation of open data archives and resources that are well established in other fields of the life and biomedical sciences. In particular, the construction of open data archives that store and publish imaging datasets and added value databases that provide data curation, mining and re-analysis must be a priority. This separate but complementary pairing of functionalities in data resources has proven effective in other domains and should be used to guide the construction of open data resources for bioimaging. Global Bioimaging strongly endorses the proposals recently published by Ellenberg et al (DOI: 10.1038/s41592-018-0195-8), that outline a vision for such a bioimage data ecosystem.

4. Target Audiences for Global Bioimaging Recommendations

In considering the construction of recommendations by the Global Bioimaging community, we have agreed that the target audiences for any of our recommendations will include imaging scientists - central facility staff and managers who deliver technical know-how and best practice to the bench science colleagues. However, we have also concluded that it is essential to also focus on journal editors and funders. We have concluded, following from several Exchange of Experience meeting discussions and presentations, that focusing on data specifications and scientific staff misses the opportunity to influence the future direction of the entities that help define policy, practice and implementation. Journals, and in particular journal editors, have contributed to the use of open data standards by requiring papers submitted for publication adopt specific standards. Funders contribute by conditioning awards on the use and adoption of data standards and where appropriate the deposition of datasets in open repositories. National funding departments that may not have fully developed expertise in these technologies can draw confidence in policy and decision-making anchored on such recommendations. For these reasons, Global Bioimaging has concluded that the recommendations we have developed should be written in a way that can be easily appreciated and incorporated by a wide cross-section of the scientific community.

5. Recommendations for Data Format Standards

In the following we outline the characteristics of useful, usable data standards. These guidelines can be used by scientists, facility personnel, funders and journal editors to assess the utility of data standards proposed by scientific groups, national programmes or transnational collaborations.

1. *Openness*

Any proposed data format must be openly available, supported by openly accessible, versioned, and editable specification(s) and documentation. Specifications and other related documents must be easily accessible from a URL or other publicly available on-line resource, following the FAIR specification—Findable, Accessible, Interoperable and Reusable—formulated by the

Force11 group (https://www.force11.org/group/fairgroup/fairprinciples). It is insufficient for documents and specifications to only be supplied on demand.

2. *Implementation*

Any proposed format should be supported by openly available software libraries that provide read and write functions for the format, preferably in multiple, community-adopted programming environments (e.g., Java, Python, C++, etc). These implementations should be open source, and include an application programming interface (API) so they can be easily adopted and included in 3rd party software. It is quite useful for the read functions to be incorporated into a validator, an application that can be used to read a file and assess how well it adheres to the standard.

3. *Examples*

Usage and adoption of a proposed data format standard will be catalysed by openly available examples—real data stored in the format. These are useful references for anyone wishing to adopt and use the format, and also can serve as tools for testing and validating software that reads and/or writes the format. For each version of the format specification, up-to-date examples should be provided.

4. *Licensing*

All data standard resources should be published under an appropriate license. Documentation, specifications, implementations, and example data sets should be licensed using an appropriate Creative Commons license, e.f., CC0 or CC-BY. Licenses that forbid commercial use often inhibit adoption by industrial research labs and commercial technology providers and should be avoided. Software for reading/writing data formats should be licensed under a permissive software license, e.g., BSD, MIT, or similar in order to promote adoption by users from across the bioimaging community.

5. Data Types

There are many different data types covering a multitude of different applications, domains and spatial and temporal scales. Any proposed standard will likely only cover one or at most a few applications or domains. The expected types of data the standard supports should be stated clearly in any documentation. In addition, the types of data supported, for example metadata related to experimental or case manipulations, image data acquisition, data processing, and analytic outputs should be clear, easy to understand for any user, and usable for search and data management applications.

6. Governance or change management

For a scientific standard to stay relevant whilst ensuring transparency, it needs a mechanism or structure for decision-making and change management. Due to the varying types of standards, their reach, and differences across their adoptive community, a governance or change management policy and process could take many forms. The most critical attributes are transparency and strong community engagement.

7. Adoption

For a standard to be considered suitable it should be adopted beyond an individual research laboratory, or institution.

6. Resources for Open Access Image Data Repositories

Imaging datasets are rich, heterogeneous and often quite large. Until recently, most image data repositories published datasets from single projects, making large strategic datasets available for query and download. However, in the last 10 years, several repositories have appeared that integrate datasets from independent peer-reviewed studies enabling datasets from electron microscopy, high content screening, multi-dimensional fluorescence microscopy, histology, and several different modalities for whole tissue or organism imaging to be published and accessed online, usually through a web browser-based interface, and sometimes through appropriate APIs. A partial list of online imaging data resources is presented in Table 1. This table shows the large, diverse and increasing number of these resources.

Data Type	Utility & Impact	Types of Users/Applications	Examples of Public Resources
Correlative light and electron microscopy	Link functional information across spatial and temporal scales with ultrastructural detail	Cell biologists, structural biologists and modellers: structural models that span spatial and temporal scales	EMPIAR (https://www.ebi.ac.uk/pdbe/emdb/empiar); BioImage Archive; IDR (https://idr.openmicroscopy.org)[2]
Cell and tissue atlases	Construction, composition and orientation of biological systems in normal and pathological states.	Educational resources; Reference for construction of tissues, organisms, health scientists	Allen BrainAtlas (https://www.brain-map.org); Allen Cell Explorer (https://www.allencell.org/); Human Protein Atlas (https://www.proteinatlas.org); Human Protein Cell Atlas (https://www.proteinatlas.org); Mitotic Cell Atlas (https://omictools.com/mitotic-cell-atlas-tool); Model organism gene expression atlases

[2] resources for CLEM will be constructed and made public in 2019-2020

Benchmark datasets	Standardised test datasets for new algorithm development	Algorithm developers; Testing systems	EMDataBank (http://www.emdatabank.org); BBBC (https://data.broadinstitute.org/bbbc); IDR (https://idr.openmicroscopy.org); CELL Image Library (http://www.cellimagelibrary.org);
Systematic Phenotyping	Comprehensive studies of cell structure, systems and response	Cell biologists, physiologists, Queries for genes or inhibitor effects	MitoCheck (http://www.mitocheck.org); SSBD (http://ssbd.qbic.riken.jp); IMPC (www.mousephenotype.org); PhenoImageShare (http://www.phenoimageshare.org/)
Whole organ and Systems	Studies of		Human Connectome Project (http://www.humanconnectomeproject.org/) http://www.med.harvard.edu/AANLIB/home.html

Table 1. Examples of Potential High Value Datasets. This table is exemplary and is not a comprehensive survey of all imaging datasets. (Adapted from Ellenberg et al, (2018) DOI: 10.1038/s41592-018-0195-8).

7. Recommendations for Open Access Image Data Repositories

Image data repositories are continuing to grow, with some becoming acknowledged resources for publishing imaging data. The Image Data Resource (IDR, Table 1) has published 10s of datasets alongside published papers, includes curated annotations of targeted genes, drug treatments and phenotypes and has been named a recommended data repository by Springer Nature journals. The Systems Science of Biological Dynamics Database (SSBD, Table 1) is collecting datasets from laboratories across Japan and annotating them with trajectories and other dynamic data cast in a formal model. The appearance and growth of these and other resources demonstrates that many of the barriers for managing and publishing large collections of images have been solved. This allows us to look ahead at the possibilities for the construction of a coherent, connected ecosystem for publishing and integrating bioimaging data. We have therefore defined key recommendations that should be implemented to ensure this momentum continues and preferably grows.

1. *Metadata Specifications for Submission*

The value of published imaging datasets can only be realised if they are accompanied by metadata that describe type and state of sample, experimental manipulations, imaging conditions and probes, and any analytic results derived from the data. The value of capturing these metadata as completely as possible has to be weighed against the reality of capturing experimental and analytic outputs from biological laboratories. Collection of biomolecular metadata during the construction of gene expression, proteomic and other datasets has

demonstrated that lightweight metadata requirements are critical for community adoption and use and that overly stringent or laborious submission requirements result in incomplete, failed, or lack of submission. Moreover, the increasing number of image data repositories may result in an equivalent number of metadata submission templates, causing confusion for data submitters and future data users. The developing image data resource should engage with the bioimaging community to define a common metadata specification that is shared across repositories, updated on a regular, predictable basis and relatively easy for data submitters to use, fill out and submit. As far as possible metadata should be harvested from the instrument, and at the time of acquisition. This will minimise any additional workload on the part of the researcher.

2. *Components of the BioImaging Ecosystem*

As noted above, the collection, annotation, storage, integration and publication of biological datasets is well-established with many resources having reached maturity and stability. These existing resources serve as models that the imaging community can use to learn useful and successful design and construction patterns.

An approach that has proven successful in several other fields is to construct two separate data resources. The first, an *archive*, serves as a repository for all data associated with publications, and stores data files and a limited amount of metadata. Data can be browsed, found using search indices and downloaded, but higher level annotation, integration and processing is not attempted, so that the archive can primarily serve a role for keeping pace with the rate of data submissions. A second type of resource, an *added value database* (AVDB), incorporates dataset from the archive, performs curation and integration and seeks to enrich data and enable discovery with the datasets it holds. The separation between the construction and operation of archives and AVDBs is critical to have an efficient data intake workflow and also to allow curation as a sufficient level to enable data re-use and discovery.

The principle of a bioimaging archive and associated AVDBs has recently been published (Ellenberg et al (2018), DOI: 10.1038/s41592-018-0195-8). While the argument appears strong, such archives do not yet exist. The goal is in 2019-2020, AVDBs like IDR and SSDB will use other existing data archives (e.g., BioStudies, https://www.ebi.ac.uk/biostudies/) and work with their communities to build support for the construction and operation of bioimaging archives.

An *archive* is particularly effective (and considered best practice) if it is capturing data from the point of experiment, whereby all data and associated metadata are captured from the point of generation, and associated with pre-experiment preparatory steps. This requires close collaboration and integration between laboratories, instrument facilities and informatics capability to connect microscopes or imaging equipment. In return, it provides the ability to provide significant added value to the data generated by an instrument facility and increases the trustworthiness of generated data. The Characterisation Virtual Laboratory (https://www.cvl.org.au/) and NIF Trusted Data Repositories projects have demonstrated this effect.

3. *Requirements for AVDBs for Deep Learning*

As AVDBs grow and mature, the well-annotated datasets they hold may be valuable training datasets for advanced artificial intelligence (AI) applications, including tools that use deep learning. However, in discussions with members of GBI who run AVDBs, there is a shared sense that there aren't clear, definitive requirements for how training datasets should be constructed, how annotations ("labels") should be formatted, or which datasets should be prioritised for formatting for AI uses. We recommend that AVDBs represented in GBI work with AI experts to define these and other requirements in order to rapidly expand the usage of bioimaging datasets for AI applications. This should include standards for linking the imaging data to other relevant data from the same subject/sample, such as genetic data and biochemical/clinical/behavioral results.

Moreover, there are clearly strong opportunities for applying AI techniques to microscopy and imaging problems. For this to be realised, it is important for communities to establish transparent community standards across data curation, data publication and technique publication. Without community consensus across these attributes, AI techniques risk becoming an irreproducible black box.

4. *Authentication for Submissions and Data Access*

As archives and AVDBs grow, the number of submissions they receive will increase, and the number of authors submitting datasets will also increase. This will inevitably raise an issue where authentication of author identity, affiliation and other critical information becomes an essential part of the data submission workflow. This is especially important for controlling access to personal identifiable information, data submitted while under embargo, and other protected datasets. Several public, diverse identifier and authentication projects, including ORCiD (https://orcid.org/), Elixir Authentication and Authorization Infrastructure (https://www.elixir-europe.org/services/compute/aai), Identifiers.org (http://identifiers.org/), Life Science Authentication and Authorization Infrastructure (LS AAI) (https://tnc18.geant.org/getfile/4229), and Australian Access Federation (AAF https://aaf.edu.au/) are building identification policies and resolution systems to ensure all members of the scientific community are associated with a unique identifier and to provide services to resources like the imaging archives and AVDBs for user identification and authorization.

LS AAI is of interest, as it is an extensive collaborative project where several research infrastructures have together defined requirements for a common AAI and work together with e-infrastructures to develop it. The AAF provides a federated web-login service that allows researchers to access a broad variety of Australian research-focused web services through their University credentials. It is used for authentication to access gateways (CVL, Genomics Virtual Laboratory) repositories (Store.* and ImageTrove) and other resources. These resources are poised to become widely used services that provide users, facilities and infrastructures easy, streamlined and compatible authentication services. While originally developed at the national or regional level, they can be extended to a global scale and/or serve as an example for general science AAI development, such as eduGAIN (https://edugain.org). We recommend that those involved in data services develop a task force to research current and ongoing work and develop proof of concept projects to assess the usage and usability of the various authentication systems that are coming on-line. In the long-term, a truly global identification and authentication system

will not only be used for individual scientists, but could also be used to identify instruments and the datasets they collect.

5. *Trustworthy Research Data Repositories*

The complexity of acquisition techniques, experiments and the resulting research data is increasing - and the ability to recreate experiments, or reuse data is proving more challenging. There is a movement to ensure that data published in repositories is trusted. The CoreTrustSeal's Core Trustworthy Data Repositories Requirements (CTDRR; https://www.coretrustseal.org/) provides a list of requirements that are deemed mandatory for a trustworthy data repository. A trusted data repository service is essential for sharing data. It ensures that data created and used by researchers is "managed, curated, and archived in such a way to preserve the initial investment in collecting them" and that the data "remain useful and meaningful into the future".

A number of Australian projects have undertaken the task to make research data repositories more trustworthy, including efforts in human and preclinical imaging (NIF Trusted Data Repositories) and lattice light sheet microscopy (under the Characterisation Virtual Laboratory). In both cases, the effort has been to create repositories where processing pipelines used more trustworthy and understandable by the researcher community. Another example is FAIRsharing.org, which provides a catalogue and characteristics of databases, data standards and other public resources (https://fairsharing.org/). These reference resources increase reproducibility and repeatability of experiments; increase researcher understanding the data; and make processing pipelines humanly transparent and increase data provenance.

8. Conclusion

Standardised data formats and public data resources are a critical "next step" for the fields of biological and biomedical imaging. The appearance of several open data formats and data repositories has demonstrated that the technology and know-how exists to build these resources. The members of GBI agree that the next step is to drive adoption by all members of the scientific community, but in particular funders and journals who can require use of open formats and data deposition as a condition of funding or acceptance of scientific publications. We have outlined the characteristics of standards that can be used by these critical stakeholders to assess the quality of proposed open formats and data repositories. We aim to use these guidelines to deliver a white paper targeted at these critical members of the global community of bioimaging scientists.