Euro-BioImaging

European Research Infrastructure for Imaging Technologies in Biological and Biomedical Sciences

WP11

Data Storage and Analysis

Task 11.1

Coordination

Deliverable 11.2

Community Needs of Current Solutions for Biomedical Image Storage and Remote Access Services

Task leaders

UNIVDUN

Additional Task Contributors:

October 2013

1. Report Summary

WP11 Objectives

To define a roadmap towards the construction of a "European Biomedical Imaging Data Storage and Analysis Infrastructure". The key objectives of this infrastructure will be:

• to support efficient and standardized storage for and access to curated biological and medical image data.

• to support open-source software for biological and medical image analysis through coordination of community efforts, provision of an actively maintained repository of state-of-the-art validated algorithms for quantitative image analysis and thorough training.

• to interface with high performance computing facilities for high-throughput and/or computation-intensive image analysis.

• to provide seamless collaboration and access to other relevant computing and data resources in ESFRI and in European and national infrastructures.

Digital imaging is now routinely used across both the life and biomedical sciences and has become an essential tool for all aspects of research, training and clinical practice. As image data volumes and complexity grow, it becomes increasingly difficult to access, view, share and analyse datasets using standard desktop-based solutions. In addition, the need for interdisciplinary collaboration, where datasets are shared with consortia of scientists with expertise in experimental biology, data analysis, image processing, modelling, physiology and/or medicine has also grown. In this new age, processing, analysis and sharing of image data based on conventional desktop-based solutions is simply no longer possible. Multi-dimensional images from unique clinical cohorts must be securely shared among defined collaborations to enable full analysis and query, while ensuring that identifiable data are never made publicly available. In biological imaging, technologies such as multi-dimensional fluorescence or high-content screening are becoming standard approaches to reveal fundamental biological mechanisms that explain human physiology and disease. Datasets produced by these technologies are routinely 10's to 100's of GBs, and in some cases, many TBs. To deliver the potential of these data, and the ambition and potential of European science, the technology that enables access to image data regardless of where it is produced has become a critical scientific need, and one which the Euro-BioImaging infrastructure must address.

In this deliverable we explore the technological requirements for remote access image systems, and in particular examine image databases and central repositories that are the foundation for the next generation of biological and biomedical science.

2. Image databases and repositories

2.1 Common use cases of image databases

The volume of data now routinely collected in laboratories and the complexity of the experimental metadata and analytic results generated by imaging experiments has driven the development of databases for imaging data. There are currently a number of open source and commercial products that provide remote access systems for biological and medical images. Almost all mix some ability to access and view an image remotely, e.g. through a web browser, with capabilities for storing, analysing, processing, annotating and searching for images. Thus these systems are referred to as "image databases" or "image management systems". They can be used for many different applications and at many different scales. For example, a laboratory with a single microscope may require an application for managing the data it produces on a microscope with a few users, while a department with an imaging facility will require a resource to manage and distribute the data collected in its multi-system facility to > 100 users. An institution may also require resources for managing the data associated with the manuscripts published by its scientists to achieve compliance with data publication requirements from national funding organisations and/or journals. Software applications are currently available to manage each of these cases and as these resources have matured they have increasingly become an important part of the software ecology for image processing.

2.2 Technical capabilities of existing research image databases

To enable remote image access and processing, image databases must deliver access to the images they hold. Most commonly this is through an "Application Programme Interface" (API). This interface provides standardised access to the data held in the system and is a standard concept in software. Typically these interfaces allow image processing applications to read data from and write data into the image database. If constructed correctly, the APIs and databases can serve very large numbers of users and many different scientific applications simultaneously and even serve data from many different types of imaging modalities.

Currently there are several open source projects providing image database and repository functionality. Two examples are The Open Microscopy Environment[1] (OME) and BISQUE. OME releases OMERO, an open source image data management platform. OMERO is used in thousands of sites worldwide and is the basis for several large public repositories[2]. BISQUE[3] is an open source

[1] http://openmicroscopy.org
[2] http://jcb-dataviewer.rupress.org/; http://www.cellimagelibrary.org/; http://www.emdatabank.org/
[3] http://www.bioimage.ucsb.edu/bisque

data management system built at UCSB. This is used by a major plant bioimaging project (http://bovary.iplantcollaborative.org) and several image processing groups in the USA. Both tools allow web-based remote access to large image datasets, subject to defined permissions specifications that control who can and can't access any particular dataset. For example a laboratory might want to privately hold certain datasets within a specific group of lab members yet publish other datasets on line in support of published manuscripts. Both OMERO and BISQUE provide this capability in the form of layered permissions control for their users. They also allow integration of analysis functions. OMERO'S API supports integration of Matlab functions, Python scripts and plugins from the ImageJ2 project (see D11.6). Going forward submitting data directly from these applications to larger public repositories (see below) will be a key piece of functionality.

2.3 Comparing Image Databases and Rich File Formats

Image databases are necessarily complex applications that require substantial expertise to develop, deploy, and maintain. Facilities that collect large datasets need access to staff with this expertise, which further increases the costs of building and running their data infrastructure. In some cases, User data can be stored using file formats that include the provision for storing rich metadata describing the experiment (e.g., OME-TIFF), or even analytic results (e.g., HDF5-based formats). Several examples of this latter approach have now been proposed and in some cases built e.g., cellh5, for timelapse siRNA screens[4]. These formats work well for single sites, but do not incorporate the complex permissions structures or transaction safety present in bona fide databases. In addition, provisions for remote access still have to be included to make these data useful for remote Users. While rich file formats may provide some utility e.g., for transferring datasets to Users for their own processing and analysis, they may not provide all the functionality necessary for management and access to data needed at Euro-BioImaging Nodes.

2.4 Comparing Image Databases and PACS

Picture archive and communication systems[5] (PACS) are well-established in medical imaging. These systems provide archive and storage of multi-dimensional medical images, usually in DICOM format, and provide remote access, often through a web browser, with provision for multi-dimensional image viewing. PACS systems can be considered image databases, as they store image data and metadata, and enable simple search functionality. Open source PACS (e.g., OSIRIX) are currently available but in general, these systems do not have the flexibility or scalability that OMERO or Bisque provide. In particular current PACS don't use the federated image data standard Bio-

[4] http://www.cellh5.org/
[5] http://en.wikipedia.org/wiki/Picture_archiving_and_communication_system

Formats and thus can't support the range of image data types allowed by OMERO and Bisque (see D11.6 for more details) or the range of heterogeneous data types supported by OMERO (e.g., biobanking data[6]).

2.5 Public Image data Repositories

As noted in Deliverable 11.1, there are now several on-line image data repositories. Together, these provide many tens of TBs of image data, annotations, and links to other resources to the biological community. The JCB DataViewer[7], ASCB CELL Image Library[8], the web microscope[9] and the EMDataBank[10] are examples of such resources that each provide on-line public access to large (e.g. >200,000 images at the JCB DataViewer) annotated multi-dimensional image datasets. They are heavily used, and accessed by many 1000's of unique users per month. Thus the technology to build remotely accessed image data repositories exists and can be deployed. The existing resources can be considered strong proofs of principle that serve as a foundation for the next steps in building the resources needed for the Euro-BioImaging community.

Looking forward, a critical next step for delivering value to the community from these resources will be the integration of various resources to allow searching across different data sets used by larger user groups and communities, so called reference data sets. Searching for "tubulin" should return results from as many different resources that contain standardized image data as possible. In addition, submission and deposition of new datasets that conform to the established standards should be enabled. An example of such an important next step beyond state of the art is the linking of current and future phenotypic genome-wide screens. These imaging-based high throughput datasets (often referred to as high content screens; HCS) systematically sample genome-wide perturbations using siRNA gene product knockdown in cultured cells or tissues and then record effects on cell phenotypes using imaging. Currently, at least 12 HCS datasets are published on-line, with more coming, and linking, cross-indexing and systematical and standardized annotation of these and future datasets is currently being developed in the framework of the SystemsMicroscopy NoE (www.systemsmicroscopy.eu).

Addressing the critical need of resource integration and data deposition by delivering a community resource that delivers value similar to the well-established genomic and microarray databases is a key priority of the Euro-BioImaging data infrastructure. This resource, a Euro-BioImaging Image Data Repository (EB-IDR), represents the next step in the development of on-line

[6] http://openmicroscopy.org/site/products/partner/omero.biobank/
[7] http://jcb-dataviewer.rupress.org
[8] http://www.cellimagelibrary.org
[9] http://www.webmicroscope.net/
[10] http://www.emdatabank.org

image repositories, demonstrating the importance and value of data submission, cross-referencing and standardised annotation for indexing and search. In addition it would crosslink to the available biomolecular databases hosted by the EBI and ELIXIR, wherever images represent a molecular measurement or provide spatial or temporal context for biomolecular information, which is more and more often the case. This effort is supported by a close collaboration and common image data policy developed jointly by Euro-BioImaging and ELIXIR.

Once established, this EB-IDR could serve as the public access point for these and other datasets, as the results generated from them are published. It can also serve as a data hosting service, before public release of data, to foster collaboration between consortia of data producers and image processing and data analysis experts. Eventually, the EB-IDR may emerge as the de facto point of deposition of image data associated with published biological and biomedical research in the EU, much as the Protein Data Bank[11] (PDB) has emerged as the acknowledged community repository for 3D structural data of biological macromolecules.

3. Cloud-based resources for image databases

As discussed in D11.5, cloud-based resources can be used for bioimage data storage and analysis, but the current cost structures of commercial cloud providers make these especially expensive for handling processing and analysis of large datasets. However, over the next 3-5 years, it is likely that national and trans-national research cloud infrastructures will become available as are currently being prepared in the Helix-Nebula project, in which Euro-BioImaging is represented through its coordinating institution EMBL, and at least some (based on current trends, we estimate up to 35%) of Euro-BioImaging's data may be stored and processed on these resources (Note: Australia has deployed its NeCTAR platform for this purpose[12]).

Tools that allow on-demand deployment of image database resources in cloud environments would be very useful for imaging facilities to serve data to remote users for analysis and computation. Image data sets used by individual or small user groups could go online in Euro-BioImaging's remote access resources when needed and then potentially go offline once analysis or mining is completed and data is archived locally, reducing the need for Euro-BioImaging to invest in permanent, large-scale storage and computational resources. Reference image data sets used by larger user groups over long times could be made available through the EB-IDR, which would offer remote compute services on this data at the repository. If larger compute resources were required,

[11] http://pdb.org
[12] http://nectar.org.au/

these reference datasets could be deployed on permanent storage linked to cloud-based HPC resources.

An example of this kind of technology is the deployment of Embassy at the European Bioinformatics Institute (EBI-EMBL). Based on ESX technology from VMware, users are able to request a web-accessible portal into a desktop that includes a defined set of data analysis applications, linked to a large data repository. This technology is currently used by EBI to support defined, limited collaborations with external groups and is not yet scalable to an open public resource. However, Embassy can be considered as a current state-of-the-art concept that might be expanded to a community resource, with the appropriate resourcing and technical development. Perhaps the most challenging problem in this type of architecture will be access to the very large datasets stored in the EB-IDR or elsewhere. In general, moving TB-scale datasets between arbitrary locations on demand is not practical. Euro-BioImaging will need to leverage investments and developments in several other transnational projects, and in particular those under development by ELIXIR, to enable general access to these datasets, and in particular to enable users to link cloud-based high performance computing resources to these data resources.

4. Conclusion

Open source and commercial products that provide remote access systems for biological and medical images are an important part of the software ecology for image access and processing. Application Programme Interfaces (APIs) and databases can serve very large numbers of users and many different scientific applications simultaneously and even serve data from many different types of imaging applications. However, image databases are necessarily complex applications that require substantial expertise to develop, deploy, and maintain and as such facilities that collect large datasets need to access staff with this expertise as an integral part of running their data infrastructure. Image database technology will be a critical resource for Euro-BioImaging Nodes needing to share data and analysis tools with Users in remote locations. In the future Euro-BioImaging infrastructure tools that allow on-demand deployment of image data sets in database resources of its cloud environment may reduce the need for Euro-BioImaging Nodes to invest in permanent, large-scale storage and computational resources. Development of a Euro-BioImaging Image Data Repository (EB-IDR) that integrates standardized image data sets for cross referencing, searching and new data deposition is a critical data infrastructure task. Ultimately, the EB-IDR should become the *de facto* point of deposition of image data associated with published biological and biomedical research in the EU. EB-IDR will serve reference image data sets to large user communities over long times. EB-IDR would offer remote compute services for user analysis needs

of the reference data and may deploy some of its data sets on demand at collaborating high performance compute centers, when specialized HPC analysis is needed by its users.