Euro-BioImaging
European Research Infrastructure for Imaging Technologies in Biological and Biomedical Sciences

WP11 Data Storage and Analysis

Task 11.1
The data storage, curation and access challenge: architecture, tools, and construction planning

Deliverable 11.1

State-of-the-art and community requirements in Biomedical Image Analysis, Storage and Remote Access

Task leader
UNIVDUN, Fraunhofer

August 2012

Introduction

Despite the growing diversity of methods of imaging in the life sciences and biomedical research, they each generate large, complex datasets and in most cases, many steps of processing and analysis are required to convert the original data into a result that can be understood, published and ultimately used by others. These steps - often referred to as a *workflow* - usually consist of some combination of commercial, open source, and custom software tools to record, access, process, view, and analyse data.

As described below, these workflows are inevitably defined by scientific goals. As science is by its very nature innovative, these workflows change rapidly, and are modified by scientists driving to achieve a certain result. As data analysis and visualization play a critical role in bioimaging, software tools that are adaptable, modular, or even completely open to modification are often preferred, as they can be tailored to scientific and experimental demands.

Many funding agencies now require the storage and management of data acquired as a result of their funding. However, the resources and repositories to manage and store this data are still being developed, and there are not yet defined, recognised applications or resources that support all the different types of bioimaging data or applications.

In this document, we survey previous analyses of the role of software resources in bioimaging, the status of existing data repositories, and opportunities for future development of these resources.

What are the informatics challenges in biological and medical imaging?

Given the rapid development in image acquisition systems in the last twenty years, it is worth considering why a corresponding rapid development of informatics tools has only occurred recently. Certainly one of the barriers to providing universal tools for bioimage informatics is the diversity of data structures and experimental applications that produce imaging data. In optical microscopy alone there are a substantial number of different types of imaging modalities, and indeed a method like fluorescence microscopy encapsulates a huge and rapidly growing field of image acquisition approaches. Informatics tools that support this range of methods must be capable of capturing not only the raw data (the individual pixels) but also metadata around the acquisition methodology including instrument settings, exposure details etc. This diversity of data structures makes delivering common informatics solutions difficult, and this complexity is multiplied by the large number of commercial imaging systems that use individually specified and oftentimes proprietary file formats for data storage. Current estimates are that there are >150 proprietary file formats for optical microscopy alone (and not including other common imaging techniques), that must be supported by any bioimage informatics tool that aims to provide a general solution.

A deeper challenge resides in each individual laboratory that uses imaging as part of its experimental repertoire. The sheer size of the raw data sets and the rate of production means that individual lab researchers can easily generate tens of Gigabytes up to Terabytes of data per day. This means that large labs or departmental imaging facilities generate hundreds of Gigabytes to tens of Terabytes per week, and are thereby now enterprise-level data production facilities. However, the expertise for developing enterprise software tools or even simply running the hardware necessary for this scale of data management and analysis rarely exists in individual laboratories. In short the sophisticated systems, their storage capacity and development expertise that are sufficient to deliver genomics databases and applications are required here in individual imaging labs and facilities. The delivery of tools that provide access to a broad range of data types, manage and analyze large sets of data, and help run the systems that store and process this data is the challenge that bioimage informatics seeks to address.

Commercial and open resources for image analysis and processing

Most research laboratories have access to commercial imaging platforms, and almost all of these include very sophisticated software made available at time of purchase or through upgrades. Such software may control image data acquisition platforms, provide sophisticated data processing, analysis, and visualization functions, or may provide a library of functions that are accessed through an open programming interface.

Such vendor software is usually provided through a commercial license: it is powerful, and it is responsible for much of the productivity and discoveries in bioimaging.. However, the flexibility described above and the need to rapidly develop prototype tools has led to the emergence of open-source or open application programming interface (API) packages that provide a foundation for adaptation and customization. ImageJ and ITK are two widely used examples, which are available and commonly used in nearly all research laboratories. The open and pluggable nature of many of these applications makes them ideal for use in scientific environments where custom applications are almost always required.

There is no general requirement that all software tools should be completely open. For example, Matlab is a commercial scripting environment that provides an open, standardized interface where users can program their own scripts based on proprietary functionality. Matlab is heavily used in cell and developmental biology simply based on its open programming interface. Among the open source programs, ImageJ is a very successful open image software package. ImageJ's core code is controlled by one developer (Wayne Rasband, NIH), but its architecture easily allows 'plug-ins' to extend its functionality. Recently a new "ImageJ2" development team has emerged, beginning the transition of the ImageJ codebase to well-established open-source frameworks, and the Fiji project now releases complete, supported versions of ImageJ and a large suite of plug-ins. The key point is that as powerful as existing commercial tools are, providing open interfaces to their underlying functionality is an attractive and effective way to enable scientists to use existing tools but also to add on their own. The resulting increased flexibility contributes to the growing use of quantitative tools in biological research and ultimately to scientific discovery.

Well-recognised deficiencies in software for bioimaging

The need for usable, effective tools for processing, analysis and visualization of bioimaging datasets is well-recognised. A recent ERA Instruments/Euro-BioImaging report (http://www.era-instruments.eu/) highlighted the need for improvements in software for bioimaging. This analysis concluded:

> "The availability of adequate software tools and standards, especially for image data management and analysis is currently unsatisfactory. The reasons are manifold:
>
> - The broad range and rapid development of image based research applications makes commercialization difficult and too

> slow to satisfy user needs.
>
> - Funding for sustainable academic software development is difficult to obtain, even more so for implementing user-friendly ready-to-use software, than for developing new algorithms.
> - A reward system for open access and open source software development is missing.
> - Software often depends on companies and their proprietary data formats that are not standardized and/or change rapidly with new software versions.
> - Metadata is very important for using and interpreting the image data correctly. However, metadata of the experiment, i.e. the instrument settings, are often difficult to access independently from the company's proprietary software. Metadata and protocols regarding for example the sample composition and preparation, are often insufficiently recorded by the users, if at all.
> - Public domain software is developed on many different software platforms and is often difficult or impossible to use, because it is not user-friendly.
>
> As a result the current landscape consists of many independently developed software tools that are not standardized or interoperable in terms of the data input and output, which has made community efforts at standardization inefficient.“

This analysis highlights the need for open efforts to define and build software tools that address the needs of scientists using bioimaging for their research. Efforts to date have been sporadic and haphazard and there is a real opportunity to develop a coordinated strategy for delivering software solutions for bioimaging.

Types of Open Source Software

A critical development in the field of bioimage informatics has been the introduction of many open source projects in the last few years. These projects range from being open source distributions where the code is available but new development is not specifically encouraged, to open development projects that are community driven projects that actively encourage the help and participation of projects for the support and addition of new features. Therefore before we proceed further it is worth considering what constitutes open source and open development efforts and why they are valuable or even necessary for bioimage informatics.

Open source software is a well-established movement with strong paradigms in many very successful projects such as Linux (http://www.linuxfoundation.org/), Java (http://java.sun.com/), MySQL (http://www.mysql.com/products/database/), Apache (http://www.apache.org/) and GNU (http://www.gnu.org). A fundamental tenet of open source software projects is that the copyright holder (usually the software developer or his/her employer) determines the software license. This license defines how the software is distributed and what end users may do with the software. For open source software, the original source code is made available under the terms of this license. An open source license usually allows end users to use the software for any purpose, make changes to the software source code or link their own software to it, and if they desire, distribute those “derivative works”. However, the software license also defines under what terms and license derivative works may be distributed. A comprehensive list of open source license is available on Wikipedia (http://en.wikipedia.org/wiki/Comparison_of_free_and_open-source_software_licenses). For any users or developers, these details are important and must be understood given the great implications for development and deployment.

The ability to see and make changes to the work of another developer is a critical component of open source software. The attractive aspect of this approach for science is that users and developers can directly see, evaluate, and use another's work (really, their intellectual property), and if necessary, build upon it. This is a key and often overlooked part of open source software. Successful open source software development projects are dynamic, evolving enterprises allowing input, feedback, and often contributions from their community.

This evolving, adaptable aspect makes open source software particularly useful for scientific discovery, and more specifically, to the rapidly evolving and diverse set of imaging applications used in biological research. Commercial and closed source applications have certainly supported many significant advances in imaging. However, an essential part of bioimaging data analysis is the ability to easily try new methodologies and approaches or even to combine existing ones to generate a derivative result based on the combination of two approaches. Open source approaches make this possible. There is therefore a natural fit between open source software and the process of scientific discovery. In addition, a consequence of the growth of the open source community is a de facto establishment of standardized documentation methods (http://java.sun.com/j2se/javadoc/) and software specifications (http://java.sun.com/products/ejb/docs.html). These specifications ensure that developers can understand and use each other's code, and most importantly, that two independent software packages can use a specified, common interface. This software "interoperability", enforced by the community either formally or informally, is a general hallmark of open source software, and perhaps one of its most underappreciated strengths. Because standardization and specification of software interfaces is so well established in the open source community, open source software has a critical role to play in providing the specifications and tools for common file formats or common interfaces that enable two otherwise incompatible packages to communicate their input and output data to one another. This type of interoperability is critical to support the rapidly evolving needs of bioimage informatics. For all these reasons, much of the recent developments in bioimage informatics are based on an open source foundation.

Recently, a subclass of open source project known as "open development" has been defined (http://www.oss-watch.ac.uk/resources/odm.xml). Open development projects take the open source concepts and add a significant role for the community in the development process. In truth, community interaction and feedback was an initial component of many open source projects, but as open source projects have expanded, not all have included efforts to engage and respond to their user community. Community interaction and support is expensive—it takes precious developer time and often requires the use of forums, mailing lists, and other resources to manage the interactions with the project's community. However, open source, and open development approaches in particular, have proven to be particularly attractive for funding agencies supporting biomedical research. They provide a way to measure the success of the project, by providing measures of uptake and participation. As the community grows around an open development project, it provides a measure of impact of the research investment and sustainability of the software past the duration of the initial award. Many agencies are now requiring that applicants have a software sharing plan in their grant application and if open source is not possible justify this decision. In general, the value for the developers, the community and the funding investment will be maximized if open development models are also followed.

Open BioImage Software Resources

There are a number of open source image data projects (see Table 1). These projects cover a wide range of functionality, are heavily used by the scientific community, and as they are open, there are a number of examples of linkages between these projects. ImageJ, CellProfiler, Endrov, and KNIME use Bio-Formats to open image files, and ImageJ, Cellprofiler and Icy all make use of the same plug-ins. The ITK image processing toolkit is probably the most comprehensive image processing project available and is used by many other image processing tools (e.g., BioImageXD, GIMIAS). The value of these linkages is just beginning to be realised. For example, as of this writing, Bio-Formats is installed and used by >45,000 sites worldwide. It is not possible to directly measure the impact of this usage, but it is likely that the integration with other tools gives scientists the facility to perform the analyses demanded by their experiments. As these linkages grow in the next 1-2 years, there will be substantial benefits to scientists who can use these increasingly stable and sophisticated tools for analysis of their data.

Opportunities for standardization

As noted above, a bioimaging workflow is determined by the specific requirements of the experiment, and thus specific steps comprising imaging, processing and analysis are not good candidates for standardization. Any defined standard acquisition or processing module will be ignored as soon as it does not satisfy a new technique or other innovation. However, mechanisms for accessing, storing or communicating data derive real benefits from being standardised, as long as they are actively supported and maintained. An example is OME's Bio-formats project, where a single software library can import dozens of file formats. The software is developed by reverse engineering >60,000 datasets submitted by community scientists and is used by 1000's of scientists every day, enabling access to a wide variety of datasets by whatever analysis software they use.. Another project is HDF5, for efficient cross-platform management of large data arrays. Thus, any tool that links different software programmes can provide standardisation. By sharing it between scientists, laboratories and institutions, efforts in individual laboratories can focus on solving their own scientific problems, and not on duplication of efforts.

Image Repositories

In the last few years, repositories for imaging data hosted by laboratories, consortia, and even journals have emerged. Image data stored in these repositories include spatial and temporal measurements of gene expression, macromolecule localization, or phenotypes of cells, tissues, or animals and provide actual measurements of the distributions, dynamics, and changes in biological systems, as recorded from digital imaging systems. Their availability on-line helps ensure integrity and enables measurement, comparison, and interrogation ensuring that data are re-used and shared with the whole scientific community. In addition, data availability drives the development of new analysis and mining applications, improving the utility of the repositories themselves, but also providing benefit to all scientists who use imaging. The sophistication and size of these resources is growing and, one day, they may reach the level and importance of the gene sequence, expression and molecular structure resources that are the foundation for much of modern biology.

The Allen Brain Atlas (http://www.brain-map.org/) and the Edinburgh MAGE (http://www.emouseatlas.org/emage) are very good examples of the current power of image analyses applications. They combine comprehensive datasets of gene expression patterns in the mouse brain with sophisticated applications. Users can search for specific expression patterns ('fgfr3 diencephalon') and even graphically define expression patterns and query their datasets for "whatever looks like this".

Other image repositories provide access to the output of genome-wide phenotypic screening projects (http://mitocheck.org; http://phenobank.org; www.cellmorph.org) or allow searching of on-line journals for similarity between published images (http://murphylab.web.cmu.edu/services/SLIF2/), but currently do not accept community submissions. There are others that do accept submissions and use defined ontologies for annotation by human annotators and querying (e.g. http://celllibrary.org; http://ccdb.ucsd.edu). These can provide especially useful educational resources, but their utility as a research resource has yet to be demonstrated.

Original image data are also becoming integral parts of scientific publications. Two biological journals support submission, publication, and download of image data alongside conventional on-line publications. The JCB DataViewer (http://jcb-dataviewer.rupress.org/) publishes original light and electron microscopy image datasets alongside the conventional figures and text published in the Journal of Cell Biology. The JCB DataViewer uses OMERO and Bio-Formats, open source software from the OME Consortium as a foundation, and delivers a custom data upload and web browser-based user interface. As of this writing, authors have submitted original data in conjunction with 261 papers and 923 figures. These include four high content genome-wide screens, which display the original images, genomic and image metadata, and the authors' phenotyping results (http://jcb-dataviewer.rupress.org/jcb/browse/4609/; http://jcb-dataviewer.rupress.org/jcb/browse/4608/). In August 2012, the JCB DataViewer will publish several electron microscopy datasets, including one image composite of a whole fish embryo, where >26,000 tiles have been aligned to display the whole organism at 1.6 nm resolution, viewable in a standard web browser (http://jcb-dataviewer.rupress.org/jcb/browse/5553/17145/). For medical imaging, the Journal of the Optical Society of America has built the OSA ISP (http://www.opticsinfobase.org/isp.cfm), a system that allows authors to upload and publish medical imaging datasets associated with publications in the journal. The system uses a custom desktop client application that downloads derived datasets at reduced resolution for 3D volume viewing and simple measurements and analysis. JOSA has published a number of "special issues" that contain supplemental original data.

The technical sophistication of all these repositories is significant, yet more development is required to fully exploit the potential of these rich, multi-dimensional data. In much the same way that repositories for genomic data evolved from various efforts initiated around the world that finally coalesced into centralized resources, the maturation of image repositories depends on strategic community-led management, consistent public funding, and a commitment to develop them into powerful, essential resources for the biological community. This will require significant continued investment using the same criteria and mechanisms that built and continue to maintain current genomics and macromolecular structure repositories. Genomics adopted a commitment to open distribution of data and software, in return for public and charity funding. In order to promote the reuse and sharing of expertise, image repositories should build upon and be accessed through open source software. Given the well-developed templates for open source software distribution and licensing and the success of open source software in genomics, making image repository software open and available should be a priority.

Table 1. Open Source Biological Microscopy and Medical Imaging Software Projects

Project	Description	URL	Reference
3D Slicer	Medical imaging visualisation and analysis, with support for ITK	http://www.slicer.org/	
ImageJ	Open source image visualization and analysis	http://rsbweb.nih.gov/ij/	(*1, 2*)
ImageJ2	Re-architecting of ImageJ	http://imagejdev.org	---
Fiji	Distribution of ImageJ	http://fiji.sc	(*3*)
Open Microscopy Environment (OME)	Releases Bio-Formats, a file format translator, and OMERO, a data management platform	http://openmicroscopy.org	(*4, 5*)
CellProfiler	Automates feature calculation and analysis, especially for HCS data	http://cellprofiler.org	(*6*)
Bisque	Web-accessible open analysis framework and a flexible annotation structure for microscopy data	http://bisque.ece.ucsb.edu/	(*7*)
BioImageXD	Python-based desktop image processing; incorporates ITK image processing functionaity	http://www.bioimagexd.net/	(*8*)
CellCognition	Development of a fast and cross-platform image analysis framework for fluorescence time-lapse microscopy in the field of bioimage informatics.	http://www.cellcognition.org	
Endrov	Open source multi-dimensional image visualization and analysis	http://www.endrov.net/	---
GIMIAS	Medical image workflow processing and visualisation	http://www.gimias.org/	
Icy	Open source multi-dimensional image visualization and analysis	http://icy.bioimageanalysis.org/	(*9*)
KNIME	Workflow tools for building simple or complex data processing pipelines	http://www.knime.org/	---
ITK, VTK	Advanced tools for image analysis and visualization; very popular in biomedical imaging, but applicable to biological microscopy	http://kitware.com	(*10*)
Micro-Manager	Open source microscope control platform	http://valelab.ucsf.edu/~MM/MMwiki/index.php	(*11*)
Micropilot	Enables fine control of automated microscopes allowing automated acquisition of specific types of events or structures during fixed cell or timelapse imaging	http://www.embl.de/almf/almf_services/hc_screeing/micropilot/index.html	(*12*)
Osirix	DICOM Viewer	http://www.osirix-viewer.com/	

References

1. W. S. Rasband. (U. S. National Institutes of Health, Bethesda, Maryland, USA, 1997-2011).
2. M. D. Abramoff, P. J. Magalhaes, S. J. Ram, Image Processing with ImageJ. *Biophotonics International* 11, 36 (2004).
3. J. Schindelin *et al.*, Fiji: an open-source platform for biological-image analysis. *Nature Methods* 9, 676 (2012).
4. M. Linkert *et al.*, Metadata matters: access to image data in the real world. *Journal of Cell Biology* 189, 777 (May 31, 2010).
5. C. Allan *et al.*, OMERO: flexible, model-driven data management for experimental biology. *Nature methods* 9, 245 (Mar, 2012).
6. A. Carpenter *et al.*, CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biology* 7, R100 (2006).
7. K. Kvilekval, D. Fedorov, B. Obara, A. Singh, B. S. Manjunath, Bisque: a platform for bioimage analysis and management. *Bioinformatics* 26, 544 (Feb 15, 2010).
8. P. Kankaanpaa *et al.*, BioImageXD: an open, general-purpose and high-throughput image-processing platform. *Nature Methods* 9, 683 (2012).
9. F. de Chaumont *et al.*, Icy: an open bioimage informatics platform for extended reproducible research. *Nature Methods* 9, 690 (2012).
10. T. S. Yoo, Ed., *Insight Into Images: Principles and Practice for Segmentation, Registration and Image Analysis*, (CRC Press, 2004), pp. 410.
11. A. Edelstein, N. Amodaj, K. Hoover, R. Vale, N. Stuurman, Computer control of microscopes using microManager. *Current Protocols in Molecular Biology* Chapter 14, Unit14 20 (Oct, 2010).
12. C. Conrad *et al.*, Micropilot: automation of fluorescence microscopy-based imaging for systems biology. *Nature Methods* 8, 246 (Mar, 2011).

13 Held M, Schmitz MHA, Fischer B, Walter T, Neumann B, Olma MH, Peter M, Ellenberg J, and Gerlich DW. *CellCognition: time-resolved phenotype annotation in high-throughput live cell imaging. Nature Methods* 7(9):747-54 (2010).