Euro-BioImaging

European Research Infrastructure for Imaging Technologies in Biological and Biomedical Sciences

WP11

Data Storage and Analysis

Task 11.2

Coordination

Deliverable 11.5

Community needs of current architectures for large-scale image processing and analysis

Task leaders

UNIVDUN

Additional Task Contributors:

October 2013

Contents

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 biomedical image data.

• to support open-source software for biomedical 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. Scientists in these disciplines generate a substantial amount of imaging data, and therefore are constant users of software tools that acquire, process and handle data. There are now a wide variety of tools for all steps in the standard imaging workflow, from acquisition and management through to processing, analysis and publication. Image analysis tools play a key part of the process of generating an understanding of large datasets, and as such they almost always reduce the data volume by converting large sets of pixel data into defined regions (circles, squares, ellipsis etc.) and single value measurements of image objects (intensity sums, features etc.). The need for these tools is universal, thus if delivered correctly and effectively by Euro-BioImaging, their impact will be significant and enabling.

In this deliverable we explore the resources necessary for European life and biomedical research scientists to access image processing and analysis software tools that enable the conversion of large imaging datasets to quantitative measurements and data that underpin regression and modelling. In particular, we focus on the user community's need to access tools and expertise for large-scale image processing and analysis that take advantage of pluggable image analysis frameworks emerging data repositories, web-based social interactivity, and cloud-based computing. The function of the existing infrastructure is to build and deliver software that has well-defined technical capability and is accessible and usable by the end user scientists. We also consider the emerging community of software developers and image analysts representing the emerging field bioimage

informatics, and discuss the requirements for fostering the development of user and developer communities and interactions between them.

2. Software Infrastructures for Image Processing

The function of biomedical imaging software resources is to deliver the capability for handling data all the way from initial acquisition to the final publication of the data in a report or published manuscript. The steps in this workflow require several distinct software tools to store, manage, process, analyse, visualise and ultimately disseminate the data collected. For any imaging-based experiment, a scientist will therefore necessarily select a number of different tools to use in each step of the experiment. The ability to build and execute this image data "workflow" is central to completion of the experiment, and indeed to the generation of usable, publishable data.

In the recent Euro-BioImaging community survey "usability" was clearly identified as the most important feature users look for in software tools. Interestingly, the next eight categories, from standardization, pluggable architecture, support and documentation were all emphasised by over half of the survey respondents (Figure 1).

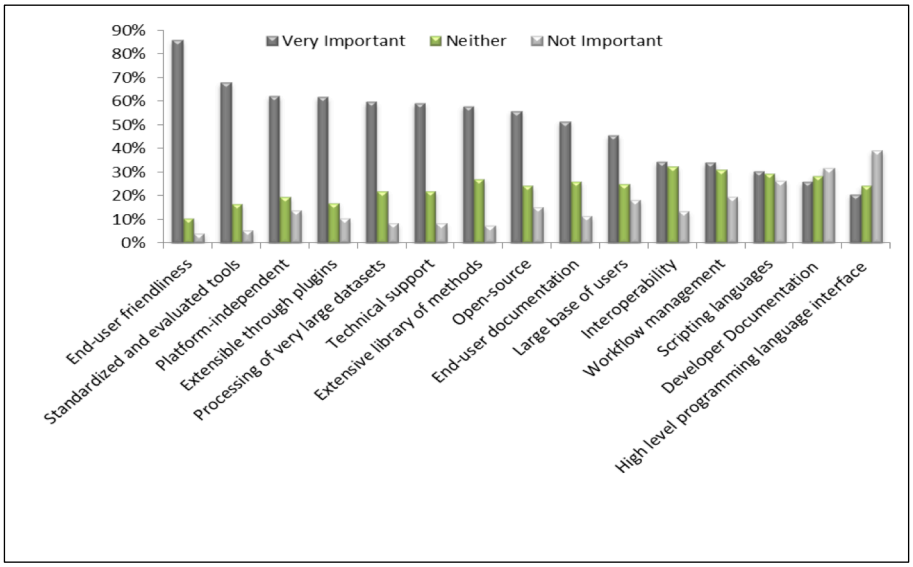

Figure 1. Priorities for choosing software for image analysis. Results from the Euro-BioImaging Survey, 2011. Results are expressed as per cent of total responses. N = 110

The Euro-BioImaging survey also highlighted the serious challenge faced by users when selecting which software tools to use— over 40 image processing and analysis packages were mentioned in the survey responses. This wide array of packages exists because the diversity of imaging technologies and their research applications has spawned a need for an equally diverse set of tools. Many of these tools have been developed commercially by >100 SMEs and LBEs dedicated to imaging. However, an equally diverse and at least as innovative open source bioimage informatics

community has also emerged and now provides many critical tools for the bioimaging user community. Most scientists using imaging don't have the expertise to create their own software tools and as such they depend on these software tools (and the developers and engineers who develop them) to deliver the critical capabilities. From a User's perspective, choosing between literally hundreds of software applications and evaluating them across the range of criteria listed in Figure 1 is simply not possible. From the perspective of software tool developers, the spectrum of user requirements and criteria creates a significant challenge—they must deliver important new methodology and technology and in addition satisfy the broad requirements that users demand. It is unrealistic to expect that the first version of every new software tool will satisfy the whole range of demands posed by its users, so an iterative cycle of development, trial, feedback, and updates are certainly essential to satisfy the needs of tools developers and user scientists.

2.1. Commercial versus Open Source Software

The Euro-BioImaging survey demonstrated that both open and closed software tools are the workhorses of bioimaging (Figure 1) and thus the two must be considered as part of the ecology and of any image processing infrastructure. In our experience, open source tools have the advantage that they can be rapidly customised and adapted by end user scientists. In addition they tend to include the most innovative algorithms and methods for analysing images. However, they are also commonly less user friendly, as they have not been through the process of developing a full commercial product. Moreover support for users is often not at the level of standard commercial service. Thus, recognising the strengths of both approaches and supporting the development of tools in both domains is critical to provide the full range of tools needed for the EU bioimaging community.

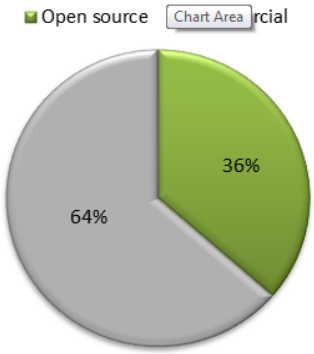

Figure 2: % of survey respondents using open source and/or commercial software ***remove chart area box***

A recent trend in software is to the development and delivery of open Application Programming Interfaces (APIs; for more detail and examples, see D11.2). These allow programmatic access to the software that runs behind the interface, regardless of whether the software itself is open or not. Open APIs have recently appeared in many commercial applications (Table 1) and are becoming a popular way for commercial software to expose controls for image acquisition systems and/or image processing and analysis functions to external 3rd party software. This is an important development,

as it allows more flexible access to commercial tools, and in some cases, interoperability between open source and commercial tools.

Table 1. Commercial BioImaging Software with Open APIs

Carl Zeiss Imaging	http://forums.zeiss.com/microscopy/community/forumdisplay.php?f=14
Nikon	http://www.nisdk.net/
SVI	http://www.svi.nl/HuygensCore
Bitplane	http://www.bitplane.com/imaris/imarisxt

2.2 Pluggable platforms

The standard for most image processing currently performed in European laboratories involves running a software application on a desktop computer, using images collected in a microscope within the same lab and processing files that are stored on that computer or on a lab or department server. Starting in the late 1990s the development of ImageJ and ITK forged a new concept in open source software where a desktop application provided a platform for application specific extensions using a series of “plugins”. These plugins could be constructed relatively easily without significant knowledge of the underlying software code. Over time the continued development of ImageJ by Wayne Rasband (NIH) and ITK by KitWare (USA) and the development and distribution of hundreds of plugins for ImageJ and ITK by the community has made this software platform concept the de facto standard for scientific image processing and thus invaluable for the life science and biomedical imaging communities. In practice this means that users can access a wide variety of functional modules and assemble them into a workflow that solves their specific scientific needs. The power of this concept is demonstrated by the **strong endorsement of ImageJ in the Euro-BioImaging survey with over 70% of respondents indicating that they routinely use ImageJ** for image processing and analysis. ImageJ was the most commonly used platform for image processing in the biological community, and the third most commonly used platform in the medical community, where Matlab, a commercial scripting language, was reported as most commonly used.

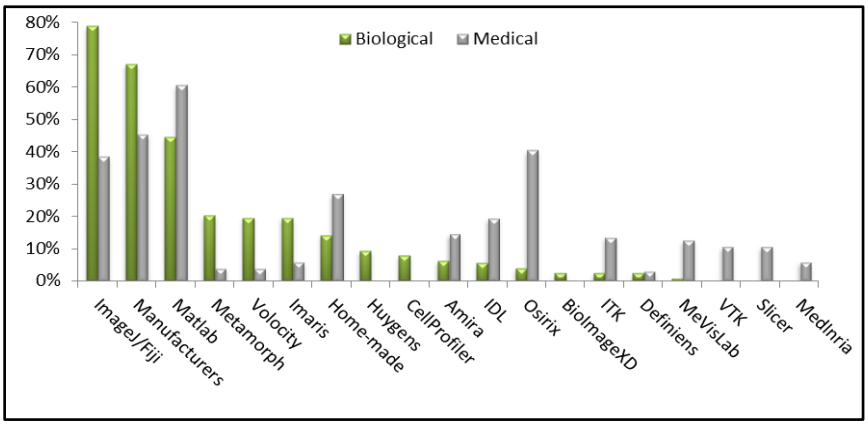

Figure 3. Image processing software use by the biological and medical imaging communities. . Results are expressed as per cent of total responses. N = 128, 104 respectively.

It should be noted that ImageJ's original underlying software architecture was designed specifically for 2D images and is poorly suited for the complex, multidimensional imaging that is now routinely performed in most bioimaging laboratories. Several efforts to address this issue are however now in progress; the ImageJ2 team[1] are rebuilding the software architecture for ImageJ while maintaining the overall appearance of the application and Icy [2] are developing an alternative Java-based application that uses modern software development practices to provide an alternative pluggable software platform for bioimaging processing. Icy has the advantage over ImageJ in that its development is led by Dr Jean-Christophe Olivo-Marin's group (Institut Pasteur) and thus can take advantage of the imaging processing expertise available in that group. Several other image processing platforms have also been developed notably Endrov, BioImageXD and OsiriX, so there are a number platforms for users to choose from.

As demonstrated by the Euro-BioImaging survey each of these software tools has their own communities, but ImageJ and ITK currently appear to be the dominant applications in biological and medical imaging respectively. The success of these platforms has been achieved not by top down decision but rather by a community-based *de facto usage* process based on the accessibility and utility of the applications.

2.3 Plugins for Image Processing and Analysis

The platform architectures described above provide an opportunity for software developers and indeed the whole bioimaging community to develop plugins that extend the core functionality of the image processing and analysis platforms. As described above the development of these plugins is a great demonstration of the power of a community-based approach. The ability to develop a broad

[1] http://imagejdev.org
[2] http://icy.bioimageanalysis.org/

set of plugins for commercial and open source software depends on well documented and engineered software interfaces. There are at present many examples of these interfaces so it is possible to consider this approach and technology mature and well understood by the developer community. The key challenge faced by developers now is how to properly communicate the availability and functionality of plugins as they are developed and receive feedback on their performance and usability from the users. A standardised web-based portal resource to address this need is discussed in D11.6.

3. File and Data Formats

The rapid growth and innovation in biological and medical imaging technologies has delivered new possibilities for scientific discovery, diagnosis and therapy. While constantly delivering new discovery and insights most imaging systems are run by custom software that writes data in some kind of proprietary file format (PFF). The arrival of every new technology or imaging platform generates a new PFF that the community must contend with. This panoply of data formats has resulted in many calls for standardized file or data formats for imaging. As imaging has transformed into quantitative science the scope of these calls has extended beyond the image data itself to include analytic output and any generated annotations. The drive to enable the rapid innovation in the field of biological imaging inevitably challenges standardization.

Several image data formats standards have nonetheless emerged. Perhaps the most well-known is DICOM which is heavily used in the medical imaging field[3]. It is supported by all of the major imaging system and now includes several specifications for file formats, viewer applications and PACS systems. One disadvantage of DICOM is the speed with which the DICOM standardization process can cope with new imaging technologies. For example the arrival of digital pathology and virtual slide imaging has resulted in a new DICOM standard, supplement 145[4] the creation of this supplement required several months of meetings and discussions. As a foundation for future standardization this is useful, but during its development, several new scanning technologies have emerged and it is not clear how well the metadata in these new technologies can be supported in this current standard.

[3] http://medical.nema.org/

[4] Singh.R, Chubb.L, Pantowitz.L and Parwani.A. (2011) **Standardization in digital pathology: Supplement 145 of the DICOM standards.** J Pathol Inform. 2011; 2: 23

In the biological imaging field OME-TIFF is now heavily used in the community and indeed supported by many commercial vendors[5] and the open source image format interoperability tool Bio-Formats (http://www.openmicroscopy.org/site/products/bio-formats). OME-TIFF uses a TIFF-based binary image storage scheme supplemented with a standardized XML based metadata model for biological imaging. In neuroscience, a separate open file format, NifTi uses an ANALYZE-style data format, which is standard in the medical imaging community. OME-TIFF and NifTi are just two examples of open file formats that support current applications in biological and medical imaging. Looking forward the desire to link in analytical results has driven the development of yet more open file formats for example based on HDF5 technology. SDCubes[6] and cellh5[7] are two examples of these types of more extensive data formats. We anticipate that the movement of more types of data into open image file formats will continue in the future, driven by continuing innovation in biological imaging.

5. Cloud-based resources for image processing

The exponential growth in size and complexity of image-based datasets has driven the demand for increased storage and rapid data processing and where possible reduction of data size. Building and maintaining local computing and storage architectures to meet this demand at each imaging facility will not be easily feasible nor economic as the speed of imaging technology development and data volume growth outpaces the reduction in storage and compute cost. Therefore on-demand, rapidly deployable, pre-specified and configured cloud-based solutions are likely to gain appeal in the future, in order to satisfy some of the needs of the community, if their current limitations in data ownership, bandwidth, and cost can be overcome.

Currently, several commercial cloud providers dominate the scene, and are likely to be the first test cases for deploying solutions for imaging. However, an emerging trend involves “science clouds” where institutions or communities build and deploy their own cloud-based resource for specific applications (http://en.wikipedia.org/wiki/Cloud_computing; Helix Nebula: http://www.helix-nebula.eu). Regardless, two barriers to common use of cloud-based solutions must be overcome before they can be considered to be a part of Euro-BioImaging’s strategic plan for data management and processing. First, transfer of image data requires high bandwidth to and from the cloud provider. Our data from proof-of-concept imaging facilities suggests that Euro-BioImaging Users will record, during each visit to a Node datasets between 200 GB to 20 TB in size (light sheet microscopy

[5] http://www.openmicroscopy.org/site/support/ome-model/ome-tiff
[6] http://www.nature.com/nmeth/journal/v8/n6/full/nmeth.1600.html
[7] http://www.cellh5.org/

is an example of this latter category). Moving 200 GB of data between physical locations can be quite slow, and moving 20 TB is prohibitive (we assume the use of standard research grade networks for these data transfer; specialised applications like Aspera (http://aspera-eu.org) can be used for this purpose, but retail costs for the server software-- ~$100,000—for routine installations at all Nodes is currently impractical). Even if bandwidth increases, charges for data I/O during computation and data egress on commercial clouds are prohibitive (currently, holding ~0.5 TB of data accessed by 3 users is ~€1000/month on Amazon Web Services). Nonetheless, over the next 3-5 years, it is likely that national and trans-national research cloud infrastructures will become available, and at least some (based on current trends, we estimate up to 30%) of Euro-BioImaging's data might be processed on these resources (Note: Australia has deployed its NeCTAR platform for this purpose; http://nectar.org.au/). For this reason, Euro-BioImaging's community will be well-served by building and deploying standardised sets of operating system and image processing packages that can easily deployed on research cloud infrastructures, as they become available. These can include tools for processing and analysing biological and/or medical imaging data and leverage emerging standards for cloud computing (e.g., the EU Digital Agenda).

6. Conclusion

The large portfolio of image processing and analysis tools that are currently available is very powerful but can be prohibitively complex for the typical imaging user as selecting the best softwares from the multitude of available packages to build the required image data storage and analysis workflow is not currently a simple activity. The aim of Euro-BioImaging should therefore be to build a processing and analysis infrastructure that can deliver software which has both the required technical capabilities and is accessible and usable by the end user scientists. Development of this model will critically depend on close collaboration between bioimaging software developers, image analysts and end users to develop systems that are user friendly, interoperable and openly accessible.