

## Finding an image in a haystack: the case for public image repositories

**Image repositories are growing. Their potential impact is huge; they require support and proper funding.**

Jason R. Swedlow

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 online helps ensure integrity and enables measurement, comparison and interrogation ensuring that data is 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 and the Edinburgh MAGE are perhaps the best examples of the current power of image analysis 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 or allow searching of online journals for similarity between published images, but currently do not accept community submissions. Still others accept data submissions and use defined ontologies for annotation and query. Original image data are also becoming integral parts of scientific publications as two biological journals support submission, publication and download of image data alongside conventional online publications.

The sophistication of all these repositories is significant, but more development is required to exploit the potential of these rich, multidimensional data. Just as repositories for genomic data evolved from efforts initiated around the world and 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 continued investment using the same criteria and mechanisms that built and continue to maintain current genomics and macromolecular structure repositories. The genomics community adopted a commitment to open distribution of data and software in return for public and charity funding. Relatively few image repositories use open source distribution of software, limiting reuse and sharing of expertise. 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.

Another key issue is the need for standardized metadata associated with images. Open, supported, metadata-rich formats for medical and light microscopy images are available. There is some progress towards common descriptions of biological experiments, but real use is limited, maybe because the demands of manual form filling are simply too high to gain substantial compliance from busy bench scientists. Automating the generation and recording of this data in laboratories is a critical requirement for the future and could be promoted by making funding for purchase of digital imaging instruments contingent on output of complete records of experimental metadata. Software tools like Bio-Formats, that read image acquisition metadata in proprietary formats and convert it into a common model can provide a practical way of capturing and using critical metadata and avoid long debates on standards definition and enforcement.

As important as experimental metadata is for validation and reproducibility, the real utility of image repositories will be realized through data analysis and mining tools that generate scientific insight and meaning. In genomics, computational sequence analysis provides enormous value and removes the effects of bias and mistakes introduced by human annotation. Similar tools are available for images: feature-based image-similarity calculations are well-developed and can recognize and classify image sets from a broad range of biological systems and problems. Most results have been achieved with two-dimensional images, and implementations that support space, time and spectral channels need more work. Download of full datasets from some repositories is possible, but will be prohibitive for many larger datasets. A compromise may be the systematic calculation of image feature sets, which are then archived and distributed with other image metadata. In the Open Microscopy Environment (OME), Bio-Formats already writes TIFF-based image files with the metadata in the TIFF header. In the near future, OME's data management system, OMERO, will publish images and metadata via a URL. This approach can enable a completely new kind of image search, one based on content and not just annotations. We will not know the exact details of how to do these calculations until the data is available. As in genomics, the data comes first, then the science.

Existing image repositories are invisible to commercial search engines. There are few links to the individual data pages, the critical metadata deviate from what most search engines actually read, and the Image Search functionality in commercial search engines is not optimized for scientific image data or metadata. With more complete and sophisticated metadata, the 'looks like this' search can extend beyond a single repository and maybe to all scientific image data. Open and accessible data repositories subject to rigorous review, following the well-established templates of genomics and structure repositories, are the foundations for this long-term goal.

#### COMPETING FINANCIAL INTERESTS

Jason Swedlow directs the Open Microscopy Environment and Glencoe Software which has built the JCB DataViewer and ASCB CELL image repositories.

Jason R. Swedlow is in the Wellcome Trust Centre for Gene Regulation and Expression, College of Life Sciences, University of Dundee, Dundee, DD1 5EH, UK.  
e-mail: jason@lifesci.dundee.ac.uk