Metadata matters

Linkert, Melissa; Rueden, Curtis T.; Allan, Chris; Burel, Jean-Marie; Moore, Will; Patterson, Andrew; Loranger, Brian; Moore, Josh; Neves, Carlos; MacDonald, Donald; Tarkowska, Aleksandra; Sticco, Caitlin; Hill, Emma; Rossner, Mike; Eliceiri, Kevin W.; Swedlow, Jason R.

Published in:
Journal of Cell Biology

DOI:
10.1083/jcb.201004104

Publication date:
2010

Document Version
Publisher's PDF, also known as Version of record

Link to publication in Discovery Research Portal

Citation for published version (APA):
Linkert, M., Rueden, C. T., Allan, C., Burel, J-M., Moore, W., Patterson, A., ... Swedlow, J. R. (2010). Metadata matters: access to image data in the real world. Journal of Cell Biology, 189(5), 777-782. DOI: 10.1083/jcb.201004104
Metadata matters: access to image data in the real world

Melissa Linkert,1,2,4 Curtis T. Rueden,1,2 Chris Allan,3,4 Jean-Marie Burel,3 Will Moore,3 Andrew Patterson,3 Brian Loranger,3 Josh Moore,4 Carlos Neves,4 Donald MacDonald,3 Aleksandra Tarkowska,3 Caitlin Sticco,1,2 Emma Hill,3 Mike Rossner,5 Kevin W. Eliceiri,1,2 and Jason R. Swedlow3,4

1Laboratory for Optical and Computational Instrumentation, Department of Molecular Biology and 2Department of Biomedical Engineering, Graduate School, University of Wisconsin at Madison, Madison, WI 53711
3Welcome Trust Centre for Gene Regulation and Expression, College of Life Sciences, University of Dundee, Dundee DD1 5EH, Scotland, UK
4Glencoe Software, Inc., Seattle, WA 98101
5The Rockefeller University Press, New York, NY 10065

Data sharing is important in the biological sciences to prevent duplication of effort, to promote scientific integrity, and to facilitate and disseminate scientific discovery. Sharing requires centralized repositories, and submission to and utility of these resources require common data formats. This is particularly challenging for multidimensional microscopy image data, which are acquired from a variety of platforms with a myriad of proprietary file formats (PFFs). In this paper, we describe an open standard format that we have developed for microscopy image data. We call on the community to use open image data standards and to insist that all imaging platforms support these file formats. This will build the foundation for an open image data repository.

Recent letters and editorials have highlighted the importance of open access to the large datasets now being collected by biologists in laboratories around the world (COSEPUP, 2009; Field et al., 2009; Schofield et al., 2009). Researchers, universities, and funding bodies all agree that scientific data produced from public- and charity-funded research (not just the results, but complete workflows including raw data) should be shared and accessible. The arguments in favor of open access data are now well established, and protocols and principles for data sharing are emerging (http://sciencecommons.org/projects/publishing/open-access-data-protocol). However, access to and sharing of scientific data require substantial effort and investment to define specifications and build resources to support them. For the successful sharing of DNA sequence data, the genome communities built, maintained, and in some cases fought for the standards and resources that were ultimately accepted by the whole community. This effort laid the foundation for the release of genomic data and the development of online resources, accessible by anyone, for any purpose, that now underpin all modern biomedical research.

We believe the imaging community can achieve the same success for digital image data. In this paper, we review the current status of online biological image repositories and provide a set of recommendations to drive the use of open standardized data formats in biological microscopy as a prerequisite for creating a global image data repository.

Scientific image data repositories for the life sciences

In December 2008, the Journal of Cell Biology (JCB) launched the JCB DataViewer, an online repository for original image data in the life sciences (Fig. 1). To our knowledge, this system is the first open repository that enables routine archiving and sharing of original image datasets supporting published scientific articles. One key attribute of the JCB DataViewer that distinguishes it from past and current data repositories is that the original binary data and metadata, additional information captured by acquisition software about an image, such as the instruments used, acquisition settings, image size, and resolution, are preserved and accessible by the community. As of this writing, the JCB DataViewer contains 6,446 multidimensional (5D; including space, channel, and time) images in support of 186 published articles. The JCB DataViewer is a customized application based on the open source and open development Open Microscopy Environment (OME) Remote Objects (OMERO) and Bio-Formats projects, released by the OME Consortium (http://openmicroscopy.org).

One goal of the JCB DataViewer was to initiate the development of a functional, scientifically valuable online image repository. The first step was to make original data available alongside a publication, available for examination by reviewers and readers of a submitted or published manuscript. Currently, the JCB DataViewer allows access to original data for viewing, simple measurement, and review, but users cannot download...
the original data files, and sophisticated image analysis and querying tools are not included in the application. In the next update, users will be able to download video versions of data stored in the JCB DataViewer, and original image data will be available in an open, standardized data format that preserves the original image metadata (OME tagged image file format [TIF]). Authors will also retain access to their original data, thereby making the JCB DataViewer an archive where authors can store their own published data. These updates represent one more step toward the development of a fully functional data repository.

The data in the JCB DataViewer are freely available to the public immediately upon publication, without a subscription to the JCB. In the future, as image repositories mature, we plan to merge the data held in the JCB DataViewer with whatever resources emerge as the definitive public repository of image data in the life sciences.

The JCB DataViewer is one of a growing number of image data repositories that are now available, each focused on providing access not only to results but also to some combination of sophisticated visualization, analysis, and mining of these complex data (Table I and Fig. 2). Each of these efforts has emphasized specific applications and functionality and reflects the simple fact that the diversity of scientific exploration and images cannot yet be addressed by a single resource. However, there are ongoing efforts to align data models where possible, and perhaps most importantly, simplify submission and subsequent processing through the definition and use of file formats that support standardized metadata. These are examples of real progress toward the goals that many have discussed and that have recently been reiterated (COSEPUP, 2009; Field et al., 2009; Schofield et al., 2009).

In summary, significant effort by peer-reviewed, competitively funded groups in the US and Europe has produced image informatics tools that the research community uses. The tools and resources are by no means finished, and our current status seems analogous to the state of the genomics resources in the mid-1980’s, when individual authors submitted their own sequence data to GenBank, SWISSPROT, and others. The diversity of imaging platforms, experiments, techniques, and data makes this analogy only partially correct and undoubtedly makes the challenge of building and running scientifically useful image repositories harder. Regardless, the sophistication of centralized scientific image resources is growing, and as a result, so will the value they deliver to the scientific community. Those resources that depend on submissions from the community will require the development, adoption, and use of standardized file formats that support as rich a metadata structure as possible. This is why the development and use of standardized image data and meta-data formats are so important.

Microscopy file formats

Many laboratories have at least one sophisticated imaging system, and many large shared-use facilities provide access to an array of imaging systems. After many years of innovation and development, modern digital imaging systems enable temporally and spatially resolved, multichannel measurement and visualization of molecular and ion concentrations in cells and tissues. Emerging imaging techniques such as multispectral, polarization, fluorescence lifetime, and fluorescence correlation are extending the complexity of analysis of biological cells and tissues. This rapid growth and evolution within the field is a double-edged sword. It certainly enables new discovery and insight. However, most digital microscope imaging systems, whether commercial products or laboratory prototypes, are usually run by custom software that saves and processes data using a PFF. In general, every new imaging platform comes with a new PFF, so rapid advances in imaging simultaneously make data exchange and access more difficult. To realize the dream of open data access and sharing, we first must solve the basic problem of accessing the data contained in PFFs. Any solution will not directly lead to new scientific insights, but it is a prerequisite for submission...
Table I. **Scientific image data repositories for cell and developmental biology**

| Resource                                           | Description                                      | Reference                                                                 |
|----------------------------------------------------|--------------------------------------------------|---------------------------------------------------------------------------|
| Allen Brain Atlas                                  | Mouse brain gene expression patterns             | http://www.brain-map.org; Lein et al., 2007                               |
| Edinburgh Mouse Atlas of Gene Expression           | Developmental atlas of mouse gene expression, including image data submitted by the community | http://genex.hgu.mrc.ac.uk/emage/home.php; Christiansen et al., 2006       |
| Fly-FISH                                           | mRNA localization in *Drosophila* embryo          | http://flyfish.ccbri.utoronto.ca; Lécuyer et al., 2007                    |
| BDGP In Situ Database                              | Gene expression patterns during *Drosophila* development | http://www.fruitfly.org/cgi-bin/ex/insitu.pl                              |
| Zebrafish Model Organism Database                  | Gene expression patterns during zebrafish development | http://zfin.org; Sprague et al., 2006                                    |
| 4DXpress                                           | Cross-species gene express pattern comparison     | http://4dx.embl.de/4DXpress; Haudry et al., 2008                         |
| Subcellular Localization Resource                  | Web-based resources for the computational determination and mining of subcellular localization | Qian and Murphy, 2008                                                     |
| SUIf and PSLID                                      | Image datasets mapping subcellular localization in *S. cerevisiae* | http://depts.washington.edu/yeastrc/ Riffle et al., 2005                  |
| MitoCheck                                           | Genome-wide siRNA screen of mitotic phenotypes in HeLa cells | http://www.mitocheck.org Neumann et al., 2010                            |
| PhenoBank Database                                  | Genome-wide C. elegans screen for functional roles in early embryonic mitotic divisions | http://worm.mpi-cbg.de/phenobank2/ Sönntichsen et al., 2005               |
| American Society for Cell Biology                  | Scientific image and video archive               | http://cellimages.ascb.org/                                                |
| Image & Video Library                              | Image data management system; provides powerful web interface and integrates several commonly used image analysis functions | http://www.bioimage.ucsb.edu/bisque Kvalekval et al., 2010                |
| Bisque Database                                    | Annotated images of cells using ontologies that specifically define the anatomy of cells and tissues | http://cclb.ucsd.edu/ Martone et al., 2008                               |
| Cell Centered Database                              | Original image data viewed through browser-based interface linked to publications in JCB | http://jcb-dataviewer.rupress.org Hill, 2008                              |
| JCB DataViewer                                     | Downloadable image data available for user viewing and rendering using downloadable software | http://www.opticsinfobase.org/bsp.cfm                                    |

SUIf, Subcellular Location Image Finder; PSLID, Protein Subcellular Location Image Database.

*In development as of 2010.

Metadata matters: access to image data in the real world • Linkert et al.

...to repositories and the discoveries they enable through reanalysis. For example, if the data from cell-based phenotypic screens were available, they could be reanalyzed for aberrations that were not of interest to the investigators who did the original screen.

Generally speaking, image data are written in formats that include the binary data and the actual image measurement, along with some representation of the metadata: the size of the binary data, its dimensions, acquisition system settings, and any other information that the developer of the acquisition software considered useful. In our experience, storage of binary data in many commercial microscopy formats is based on common formats (TIFF, HDF5, and OLE2, etc.) or other formats that most software tools can read (although there are some notable, extreme exceptions). The much more challenging problem is the metadata. Because standards are not yet agreed upon, microscope and imaging companies define their own metadata formats in their PFFs, and these are often incompatible with those from competing companies.

Since 2000, the OME has been dedicated to building tools for specification, management, and sharing of biological light microscopy data (Swedlow et al., 2003, 2009; Goldberg et al., 2005). OME has developed and released the OME Compliant specification (Fig. 2), which covers most of the metadata in PFFs from many sources and includes most of the fundamental imaging metadata in cell and developmental biology. This specification, used within the context of a TIFF file (OME-TIFF), provides a simple, easy to use format for microscope imaging data that can be used by any software that reads the TIFF file format. Several commercial imaging systems now support OME-TIFF in their software. A popular tool (>13,000 installations worldwide) is Bio-Formats, a software library that interfaces with a large number of software tools (such as ImageJ), enables the reading of >75 PFFs, and supports output to OME-TIFF.

Future directions and recommendations

For many years, the imaging community has expressed a desire to move away from the current ad hoc approach toward more defined standards for metadata representation (Goldberg et al., 2005). However, creating a reasonable standard takes years of community discussion and effort. For the standard to be successful, it must be widely used and functional enough to be worth the effort of conformance, and it takes time for the “snowball effect” to occur. Given the diversity and rapid evolution of imaging applications in biology, we don’t believe that standards can be mandated by any one entity. Instead, we argue that standards for biological imaging must be supported and developed, and once they are valuable for scientific discovery and data...
Figure 2. Recommendations for OME Compliant image metadata. The Image and Instrument Elements from the OME Data Model, with attributes and hierarchies shown in diagrammatic form. The Image Element contains core metadata that can be used for display and processing of the associated binary image data. Currently, an OME Compliant image completes all of the metadata in the Image Element. By the end of 2010, we aim to include the Instrument Element in the OME Compliant specification. The Bio-Formats library provides support for writing OME-XML either as a stand-alone file or within the header of an OME-TIFF file. The full XML Schema version of the OME Data Model is available at http://ome-xml.org/browser/Schemas/OME/2010-04/ome.xsd. Updates to the OME Data Model are announced on the project’s roadmap site (http://ome-xml.org/roadmap).
Box 1. Recommendations for use of PFFs

1. Image metadata must be associated with the binary image data, preferably as a single file.
2. Microscope systems must not store metadata in proprietary databases that are available only on the data acquisition system.
3. Metadata must be readable by third party software using a common, openly accessible software package or library. PFF developers must work with developers of open translation libraries to ensure their format is correctly interpreted.
4. Scientists must use image processing and analysis tools that preserve image metadata.
5. Image data must reflect the original measurement. If compression is supported, the user must be given the option of saving uncompressed or losslessly compressed images (which allows the exact original data to be reconstructed after compression). If compression or encryption is used, the algorithm and parameters must be stated and stored in the metadata.
6. Commercial software programs must provide data export to an open metadata specification. To ensure that commercial software stores these formats correctly, open, freely available libraries and format validators must be available to enable compliance.
7. Public and charity funding for imaging systems must include a requirement that the system writes data in an open, accessible format, wherever possible.
8. All file formats must use versioning to reflect any changes in the data model.
9. When PFFs must be used, new versions must be announced to the scientific community, and users and funding bodies must predicate their purchases on this type of support for the scientific community.
10. Once a standardized repository is available, journals must require deposition of original data supporting scientific manuscripts.

As noted in the Box 1, the use and adoption of these file formats won’t happen by itself, the community must work to drive their adoption. Individual scientists and their funding bodies must require support for these formats when they purchase or fund new imaging systems. The argument for this concerted action is based on a simple, practical goal: scientific data, funded by the public and nonprofit charities, must be publicly available. Over the next few years, the technical capabilities in image repositories will mature. Data to fill these repositories must be open, accessible, and ready for use.

We thank Dr. Alexia Ferrand for preparation of samples for structured illumination data and Angus Lamond for critical reading of the manuscript.

Work on OME in J.R. Swedlow’s laboratory is supported by the Welcome Trust (grant 085982) and the Biotechnology and Biological Sciences Research Council (grant BB/G022585).

Submitted: 21 April 2010
Accepted: 5 May 2010

References

Christiansen, J.H., Y. Yang, S. Venkataraman, L. Richardson, P. Stevenson, N. Burton, R.A. Baldock, and D.R. Davidson. 2006. EMAGE: a spatial database of gene expression patterns during mouse embryo development. Nucleic Acids Res. 34:D637–D641. doi:10.1093/nar/gkj006

Committee on Science, Engineering, and Public Policy (US), and Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age. 2009. Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age. National Academies Press, Washington, DC. 162 pp.

Field, D., S.A. Sansone, A. Collis, T. Booth, P. Dukes, S.K. Gregurick, K. Kennedy, P. Kolar, E. Kolker, M. Maxon, et al. 2009. Megascope: ‘omics data sharing. Science. 326:234–236. doi:10.1126/science.1180598

Goldberg, L.G., C. Allan, J.-M. Burel, D. Creager, A. Falconi, H.S. Hochheiser, J. Johnston, J. Melien, P.K. Sorger, and J.R. Swedlow. 2005. The Open Microscopy Environment (OME) data model and XML file: open tools for informatics and quantitative analysis in biological imaging. Genome Biol. 6:R47. doi:10.1186/gb-2005-6-5-r47

Haudry, Y., H. Berube, I. Letunic, P.D. Weaber, J. Gagneur, C. Girardot, M. Kapushesky, D. Arendt, P. Bork, A. Brazma, et al. 2008. 4DXpress: a database for cross-species expression pattern comparisons. Nucleic Acids Res. 36:D847–D853. doi:10.1093/nar/gkn797

Hill, E. 2008. Announcing the JCB DataViewer, a browser-based application for viewing original image files. J. Cell Biol. 183:969–970. doi:10.1083/jcb.200811132

Kvilekval, K., D. Fedorov, B. Obara, A. Singh, and B.S. Manjunath. 2010. Bisque: a platform for bioimage analysis and management. Bioinformatics. 26:544–552. doi:10.1093/bioinformatics/btp699

Hill, E. 2008. Announcing the JCB DataViewer, a browser-based application for viewing original image files. J. Cell Biol. 183:969–970. doi:10.1083/jcb.200811132

Kvilekval, K., D. Fedorov, B. Obara, A. Singh, and B.S. Manjunath. 2010. Bisque: a platform for bioimage analysis and management. Bioinformatics. 26:544–552. doi:10.1093/bioinformatics/btp699
Lécuyer, E., H. Yoshida, N. Parthasarathy, C. Alm, T. Babak, T. Ceronina, T.R. Hughes, P. Tomancak, and H.M. Krause. 2007. Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell.* 131:174–187. doi:10.1016/j.cell.2007.08.003

Lein, E.S., M.J. Hawrylycz, N. Ao, M. Ayres, A. Bensinger, A. Bernstein, A.F. Boe, M.S. Boguski, K.S. Brockway, E.J. Byrnes, et al. 2007. Genome-wide atlas of gene expression in the adult mouse brain. *Nature.* 445:168–176. doi:10.1038/nature05453

Martone, M.E., J. Tran, W.W. Wong, J. Sargis, L. Fong, S. Larson, S.P. Lamont, A. Gupta, and M.H. Ellisman. 2008. The cell centered database project: an update on building community resources for managing and sharing 3D imaging data. *J. Struct. Biol.* 161:220–231. doi:10.1016/j.jsb.2007.10.003

Neumann, B., T. Walter, J.K. Hériché, J. Bultescher, H. Erfle, C. Conrad, P. Rogers, I. Poser, M. Held, U. Liebel, et al. 2010. Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature.* 464:721–727. doi:10.1038/nature08869

Platani, M., I. Goldberg, J.R. Swedlow, and A.I. Lamond. 2000. *In vivo* analysis of Cajal body movement, separation and joining in live human cells. *J. Cell. Biol.* 151:1561–1574. doi:10.1083/jcb.151.7.1561

Qian, Y., and R.F. Murphy. 2008. Improved recognition of figures containing fluorescence microscope images in online journal articles using graphical models. *Bioinformatics.* 24:569–576. doi:10.1093/bioinformatics/btm561

Riffle, M., L. Malmström, and T.N. Davis. 2005. The yeast resource center public data repository. *Nucleic Acids Res.* 33:D378–D382. doi:10.1093/nar/gki073

Schofield, P.N., T. Bubela, T. Weaver, L. Portilla, S.D. Brown, J.M. Hancock, D. Einhorn, G. Tocchini-Valentini, M. Hrabé de Angelis, and N. Rosenthal; CASIMIR Rome Meeting participants. 2009. Post-publication sharing of data and tools. *Nature.* 461:171–173. doi:10.1038/461171a

Sönnichsen, B., L.B. Koski, A. Walsh, P. Marschall, B. Neumann, M. Brehm, A.M. Alleaume, J. Artelt, P. Bettencourt, E. Cassin, et al. 2005. Full-genome RNAi profiling of early embryogenesis in *Caenorhabditis elegans.* *Nature.* 434:462–469. doi:10.1038/nature03353

Sprague, J., J. Bayraktaroglu, D. Clements, T. Conlin, D. Fashena, K. Frazer, M. Haendel, D.G. Howe, P. Mani, S. Ramachandran, et al. 2006. The Zebrafish Information Network: the zebrafish model organism database. *Nucleic Acids Res.* 34:D581–D585. doi:10.1093/nar/gkj086

Swedlow, J.R., I. Goldberg, E. Brauner, and P.K. Sorger. 2003. Informatics and quantitative analysis in biological imaging. *Science.* 300:100–102. doi:10.1126/science.1082602

Swedlow, J.R., I.G. Goldberg, and K.W. Eliceiri; OME Consortium. 2009. Bioimage informatics for experimental biology, *Annu Rev Biophys.* 38:327–346. doi:10.1146/annurev.biophys.050708.133641