Digital data RFI **January 12, 2012**

Submitters: Prof Jason R. Swedlow and Dr. Emma Hill, Open Microscopy Environment 'http://openmicroscopy.org'

Background

Since 2000, the Open Microscopy Environment (OME) has developed data specifications and software tools to handle complex multi-dimensional scientific image data for the life sciences community. OME is an open-source, international consortium of academic scientists building data management tools for life sciences imaging. All resources built and maintained by OME are available at http://openmicroscopy.org.

OME releases OME-TIFF, a specification for an open, multi-dimensional image file format, Bio-Formats a software plug-in library that accesses >120 scientific image file formats, and OMERO an open-source, enterprise-level application. OME works with a large number of commercial imaging companies to provide support for open file formats and their software and for supporting their own proprietary file formats. In 2005, OME founded a commercial arm, Glencoe Software, Inc., to provide opportunities for customising OME software for specific uses. This led to the development of the JCB DataViewer (http://jcb-dataviewer.rupress.org) the world's first online scientific image publication system. The JCB DataViewer is built upon OME's open source foundation and is an example of the delivery and power of open tools for data publication and archive.

Our responses to the RFI reflect over 10 years of work in the field of scientific image data access and management and our expertise in building tools that are useful for scientists. As is clear from our responses, we do not believe that there are single individual standards that can be used for solving the data problems in modern science. We have worked hard to develop standard interfaces that allow access to complex data types, and seen significant success with this strategy. Our experience suggests that this strategy may be deployed more broadly, in domains beyond life sciences imaging.

Different domains require different solutions, however all domains use publication of scientific results as a medium for communication and dissemination. An effective delimiter of what data should be published can be related to the content of a published paper—if the data is directly related to the results presented in a paper, the data should be published. If the data is supplementary or accessory to a publication, it may be published, but this should be optional and decided jointly between the scientist author(s), reviewer(s), and the publisher.

We are happy to follow up questions or discussions on these issues. We are excited that the problem of publishing scientific data is taken so seriously in the United States and will support these activities in any way we can.

Jason Swedlow j.r.swedlow@dundee.ac.uk
Emma Hill e.e.hill@dundee.ac.uk

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Comment 1: Fundamentally we believe that the publication of scientific data is one of the most important parts of the scientific enterprise. We do not believe that all data generated by scientists in most fields should be published. How much data should be published will vary from field to field depending on individual experiments and the existing scientific culture. One common denominator across all scientific domains is the process of publication in peer reviewed journals or online facilities. Publication represents a determination by the scientist authors, reviewers, and editors that a body of work should be delivered to the community for dissemination and consideration. Following this well-established principle, data associated with experiments reported in a publication should be publicly available.

This policy achieves two things. First, it ensures that data associated with a publication is available to the community. Second it provides a convenient definition of what data should be publicly available and what data is probably supplemental or not necessary for publication. Certainly most experiments generate substantial amounts of data that for whatever reason are supplementary or maybe not even sufficient for future analysis. In our view, this less useful data should not be included in the public record, at least in the first instance, simply because we don't yet have the tools to define its status and utility in a convenient and commonly understood way. This overarching policy can be used by individual domains to define what data should be published in each field.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

Comment 2: All scientific fields have wrestled with the tensions between publication and preservation of intellectual property. There is no reason to develop any new processes here but simply to use those that already exist to protect intellectual property and to support the publication of science for consideration by the community and the public. Specific licenses can be associated with the data such that scientists and/or publications are always cited if required. This issue has been resolved already for many data types (genes, structures, images at the JCB DataViewer). Data can be accessed, to some extent analysed and also downloaded, and is available for re-use and distribution under the Creative Commons Attribution-Non-commercial-Share Alike 3.0 Unported license.

Funding agencies and scientific research institutions already have policies associated with any intellectual property that results from research they have supported. There are already mechanisms to protect patentable findings if necessary. Data publication can use these same mechanisms.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Comment 3: Different fields have different community standards defining what amounts to an individual publication, and Federal agencies can take their cues from these communities to define what is necessary in each field. If and when a scientist publishes something as a result of their analysis of some digital data, ideally those data should be made available.

Federal agencies can support this diversity (and perhaps slowly drive consensus) by funding development of software tools that access and use this data. This investment will help energise the use of data, and very likely help define what can actually be done with the data.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

Comment 4: Consider the costs of a scientist having to run an experiment or create a clone/antibody/etc. from scratch versus the cost of their being able to obtain this from someone who has published some results of that experiment or details of such a clone/antibody. The cost reduction of simply being able to acquire these data is easily apparent. The reduction in up-front costs releases funds for other avenues of research, and the benefits are obvious. In the long-term if there were better locations for all data to be housed and subsequently accessed this would reduce the burden on scientists time and costs associated even further.

An additional consideration is that currently in many research locations the steward of data is the person who produced it rather than any central location within that lab or university. As students and RAs leave laboratories, data is often lost or no longer accessible even to the person who led the research. Submission of data to a central repository serves an important archiving process, and reduces the risk of lost data due to hardware failure and simple inability to properly manage data.

Availability of digital data opens the data up for analysis by people in unrelated fields. For example, computer scientists who work on feature recognition in images might produce advanced new algorithms only if they have good exemplar data sets for development and testing. This applies to both academic and commercial settings, and could serve as a major boost to the US economy.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

Comment 5: Stakeholders can best contribute to the implementation of data management plans via the provision of the relevant and necessary infrastructure. Laboratories, Universities and research organisations could resource core facilities to provide researchers with an easy way to access and archive their data. Most scientific publishers already contribute to the implementation of data management by acting as the enforcers of most current policies by making sure data are deposited before publication can proceed. One journal, The Journal of Cell Biology, has also worked to create their own database to house original image data relating to the manuscripts that they publish. While individual journals can do this, it does not seem optimal and as is done for other data types like protein structures etc. this might be better housed in one centralised repository.

Other critical resources that must be developed are tools for accessing public datasets. These are not simply databases, but full-fledged applications that provide access and analysis of data. OME is an example of such a project funded by charity and government research organisations that works with many different entities and develops tools for the community-- thus the investment from funding bodies has been returned to the community for tools for data publication. We note that OME is developed by an active team of scientists and expert software developers. We believe this is critical to the development of tools that are on the one hand well-designed software, and on the other, useful for scientists.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

Comment 6: In our experience no single entity can take on all of the aspects of managing scientific data. Laboratories, departments, universities, publishers and funders all have a role to play in this process. To date, significant discussion has gone forward but relatively little action has occurred. However, the most important changes have occurred when funding bodies, e.g., The Wellcome Trust and the NIH, have unequivocally demanded that the outputs of their research be deposited in publicly available repositories. These actions created more subsequent action than any other previous discussion or policy process. They also forced the funding agencies to contend with some of the consequences, e.g., the establishment of PubMed Central. Thus the most important funding mechanism is a strong, definitive and unequivocal policy statement from funding bodies requiring the public release of data. This statement, backed up with resources to develop the necessary tools, is the action that will change the way scientific data will be handled.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

Comment 7: As funding agencies review their repertoire, it would be relatively easy to check whether relevant data detailed in any publications resulting from the research funded have been made available. For example, each publication now has a unique DOI. It seems relatively easy to develop DOIs for individual datasets that can be reported by investigators in their

funding reports, follow up studies etc. Generation of DOIs is automated and can be extended to individual datasets.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

Comment 8: We believe a 'bottom up' approach is the most effective response for this question. Funding bodies should be ready to invest in the outputs of publicly available research data. The proper investments will appear once the data is available. As an example Google didn't develop their search expertise and then wait for the World Wide Web to appear. First the data was available, second a number of people tried to solve the search problem and failed (rather spectacularly, because the problem was hard), and third a great solution and a great US company appeared with a new method—based on publicly available data—and soon thereafter, a new, transformative commercial model.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

Comment 9: As mentioned above, use the same existing publication and DOI principles that are already well established across all scientific domains.

Standards for Interoperability, Re-Use and Re-Purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

Comment 10: These standards must be defined by the community of scientists who generate a certain kind of data perhaps in conjunction with the manufacturers and vendors of the instruments used.

In OMEs experience spending significant effort on defining a single data standard is rarely successful and ultimately self-defeating. The development of new technologies happens so quickly that any data standard is rapidly rendered obsolete. Moreover, many technologies are made available by commercial companies whose compliance with specifications-- even those like MIAME (or DICOM, etc.)--is relatively inconsistent. In OME we have taken the strategy to provide an open specification, OME-TIFF, which many commercial providers now use (however, when examined in detail, their compliance is rarely complete). However, our most successful tools are not common file formats but software that provide a common interface to the wealth and breadth of different data types. These strategies are embodied in our image access tool Bio-Formats and our data management application OMERO. This is a pragmatic and flexible approach that recognises that the change of pace of scientific data applications cannot be limited by a data standard—new technology must be able to record the data it produces in whatever form. Software tools can be adapted to read this new data, and if designed correctly,

used by any processing algorithm to access that data. Thus, a relatively small investment in tools that can access the data and keep up with the rapidly developing data generation systems is probably the best strategy for driving digital data standards. Our mantra: don't standardise the data (it's impossible anyway), standardise the interface to the data.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

Comment 11: In OME we have taken an alternative approach to providing standardised access to data. On the one hand we have provided a common data specification, OME-TIFF recognising that the specification will always be behind the cutting edge of data generation. In fact some of the commercial companies who market the microscopes have now also adopted the OME-TIFF format as one of the options for data to be stored in directly from the microscopes. On the other we have built open-source data access libraries that anyone can use. Critically Bio-Formats is built through the contributions of the community: users submit data that should be supported, we reverse-engineer the file formats no matter their source and then-- usually within a day or two-- release software that reads the data. Currently OME holds about 47,000 submitted datasets from around the world. Bio-Formats is installed and running at >37,000 sites worldwide and is started many 1000's times each day.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

Comment 12: When driven by the scientific community, these standards already often involve scientific representatives from multiple countries. In our experience the most important contribution of federal bodies would be the commitment to developing tools like Bio-Formats and OMERO that provide standardised access to data as opposed to standardised file formats. This is a relatively small investment of a few FTEs per year who build and maintain these tools that has huge impact that crosses national boundaries.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

Comment 13: This already works very well for several kinds of data for which there are already well-developed and supported publicly available databases for data to be deposited into (such as genomic DNA and protein sequences, solved protein structures, etc.). Thus the most important practice is the generation of unique identifiers that define individual datasets using the well-established DOI system. While not ideal it certainly is a system that is established, accepted and can be rapidly deployed. These can then be used for monitoring and verifications.

In our experience with OME and the JCB DataViewer probably the most important practice and policy is to accept that the data publication problem is in fact quite challenging. The community has been discussing data publication and data release now for many years. No individual

solution as built today will satisfy all necessary requirements, but developing and deploying these solutions in steps has multiple benefits. It engages with the community and begins the process of training the community to publish their data. It helps develop new tools and identify the true problems and bottlenecks. It provides the technical solutions to problems as they come up. Most importantly it begins the process of making data available and delivering data to the community.