Matthew Cockerill
Managing Director BioMed Central
236, Gray's Inn Road
London
WC1X 8HB
UK

tel +44 20 3192 2000
www.biomedcentral.com
info@biomedcentral.com

Office of Science and Technology Policy
digitaldata@ostp.gov

11th January, 2012

Response to FR DOC 2011-28621

Dear Sir or Madam,

On behalf of BioMed Central Ltd, I am writing to respond to the OSTP's RFI on *Public Access to Digital Data Resulting From Federally Funded Scientific Research*.

BioMed Central is a leading open access publisher. Since its launch in 2000, BioMed Central has demonstrated that commercially viable business models exist which allow scientific publishers to make the peer-reviewed research articles they publish immediately and freely available online in their official form, with costs typically covered via a publication fee. BioMed Central is a founder member of the Open Access Scholarly Publishers Association (http://www.oaspa.org/) and since 2008, has been part of Springer Science+Business Media, the world's second largest publisher of scientific, technical and medical journals (STM).

One of BioMed Central's key objectives as a publisher has been to help researchers to share not only the final results of their work, in the form of a research article, but also the data which underlie that work. To that end, BioMed Central has taken an active role in many data sharing initiatives, and has created a Publishing Open Data Working Group involving funders, researchers and publishers to help identify best practices to encourage data sharing, and to identify steps publishers can take to facilitate such sharing. See http://bit.ly/n4U348 for additional information.

Responses to the specific questions in the RFI are given below:

(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?

There are already examples of scientific communities where policies requiring open data sharing have been widely adopted with great success, and which could serve as useful models for a wider range of data resulting from Federally-funded research, and which demonstrate the economic benefits of such data sharing. Specifically, since the early days of the human genome project, general consensus in the genomics community has led to a policy of immediate open availability of all genetic sequence data from publicly-funded research, with little or no embargo period. Community norms give data creators priority in terms of publications and credit resulting from data analysis, but with a relatively short period of exclusivity (6-12 months) [Contreras 2011: http://www.sciencemag.org/content/329/5990/393.short].

Economics and astronomy are other examples of fields where community norms exist requiring data sharing after some limited time period of exclusivity. The appropriate timeframe for such an embargo period is likely to vary by field, but it is important to realize that some fields can learn from one another, and that in fields in which data-sharing has been slow to take off data-sharing may need to be incentivized, and obstacles to such sharing eliminated.

In the field of clinical trials, there is strong public interest in ensuring that results and data are made freely available as soon as possible following the completion of the trial. [Gøtzsche 2011, http://www.trialsjournal.com/content/12/1/249]. Not only are positive results of vital interest to patients, but negative results are also of vital importance. If they are not reported, then the resulting selective reporting of clinical trial results can lead to a systematic bias in the scientific literature, undermining the validity of the evidence-base for the effectiveness of treatments, and ultimately leading to detrimental effects on the quality of healthcare.

Since 2007, the deposit of summary results for clinical trials has been Federally mandated (http://clinicaltrials.gov/ct2/info/results). However, while such summary result sharing is beneficial, a great deal of additional data is captured as part of the clinical trial which could have great value if shared. For example, sharing such raw data can facilitate more accurate meta-analysis of the results from multiple trials. Great care is needed in designing ethical data sharing policies for clinical trial data, however (Hrynaszkiewicz et al. 2010. http://www.trialsjournal.com/content/11/1/9). Informed consent is vital, as is a suitable mechanism to protect the privacy of individual patients. For complex datasets, watertight anonymization may be difficult to guarantee, and for this reason full public access to all raw data may not be possible. In such cases, public access may be given to a limited subset of

data, while the full dataset might be maintained in a suitably protected repository, with access provided only for specifically approved research uses.

(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?

Intellectual property interests of publishers are likely to be less of an obstacle when developing policies on access to research data than when considering policies on access to published research articles. The ALPSP and STM publishing associations issued a joint statement on access to data in 2006 [http://bit.ly/wrPwph] recommending that: *"raw research data should in general be made freely available. When data sets are submitted along with a paper for consideration in a scholarly journal, the publisher should not claim intellectual property rights in those data sets, and best practice would be to encourage or even require that the underlying research data be publicly posted for free access."*

To encourage sharing of data from the private research sector, it may be beneficial to identify types of dataset which relate to "pre-competitive" research work, and which companies may agree to share to create an "information commons" to facilitate new knowledge discovery. Sage Bionetworks [http://sagebase.org/] is one example of such an information commons, formed as a non-profit spin-off from Merck.

(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?

By developing policies in conjunction with the scientific communities the policies are intended to serve, while also sharing experience and best practices between different domains. For example, in ecology and evolutionary biology, researchers in the community have worked with a consortium of different journals in the field to successfully implement a joint data archiving policy (JDAP), which requires the data supporting peer-reviewed publications to be publicly available [Whitlock et al., 2010: http://dx.doi.org/10.1086/650340].

(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?

It is not economically feasible to indefinitely preserve all research data and so data archiving policies need to take account of the likely long term value of data when determining what should be permanently archived. For example, because of the huge amount of genomic data being collected, much of it may only be kept for the length of project, with only the most relevant parts being preserved long term.

To make data part of the permanent citable scientific record, we need confidence that it will remain available long term, so ensuring that robust and sustainable models are in place to achieve this should be a vital part of any Federal data archiving and dissemination policy.

DataCite (www.datacite.org) has helped with one aspect of managing digital permanence for datasets, by associating digital object identifiers (DOIs) with datasets stored in digital repositories, so that even if the dataset moves to a new location, the DOI can still be used to locate it.

Stability of funding for data repositories is the other major concern, when it comes to delivering digital permanence. Currently many data repositories survive on short term project funding, which creates an increased risk that data will be lost.

It may be that dataset archives will need to adopt some of the same approaches to funding used by research journals to achieve long terms self-sufficiency, perhaps by charging a fee for deposits. One possibility is that such a data deposit fee could form part of the fee paid by the author from their research funds, for publication in an open access journal.

The Dryad repository (http://datadryad.org/) – currently publicly funded – has a long-term sustainability plan, which includes deposit fees for data sets associated with peer-reviewed publications. Funders should consider providing explicit, ring-fenced, funding as part of grants, to cover data archiving and data publication costs, as many already do for open access publication fees.

Commercial services offering data archiving exists, such as LabArchives (http://www.labarchives.com/) and Amazon, with usage-based subscription models, and which may have a role to play. While they may not be able to guarantee long-term preservation, they are well placed to help the scientific community by making archived data conveniently available to researchers 'in the cloud'.

International collaborations between data archives to mirror each others' content provide one approach which seems likely to increase the likelihood that content will be preserved long term, especially if the various archives have independent sources of funding.

It is important to recognize that Federal policy could encourage and even mandate data sharing without needing to provide all infrastructure for such data-sharing centrally. In developing data-sharing policies, Federal agencies should also consider distributed options.

A 2008 report by John Houghton and colleagues, for the Joint Information Services Committee, concluded that data archiving offers an excellent return on investment for funders [http://ie-repository.jisc.ac.uk/279/2/JISC_data_sharing_finalreport.pdf]. However, in order to motivate individual researchers to comply with data sharing policies, specific,

relevant case studies – success stories – for data publishing should be catalogued to demonstrate the benefits of data sharing. Examples include secondary use of clinical trial data for a future systematic review, or identification of adverse effects of a medication.

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

Incentives (including penalties for non-compliance) may be needed to ensure widespread compliance with data sharing policies set by funders, because many researchers may be understandably cautious about sharing their hard won results with others, whatever the wider benefits to society, and all the stakeholders listed can play a role in ensuring such incentives are in place. Making data sharing a condition of publication has proven to be viable in the case of DNA sequences, protein structures and clinical trial registration, and the Dryad example shows how such policies can be successfully introduced in new fields, if a critical number of journals can be persuaded to participate. Journals and publishers can also support open sharing of data by enabling citation of data and providing journals and publication formats for publishing and describing published data sets. New measures of research impact, which go beyond traditional Impact Factor measures, are evolving (http://total-impact.org/). Also, existing tools need to be identified, or new tools developed, to enable efficient sharing and management of data. At Oxford University, the DataFlow project provides an open source data management infrastructure (http://www.dataflow.ox.ac.uk/), which aims to make it easier for research groups to manage data. LabArchives also provides online lab notebooks offering the ability to publish data and assign DOIs.

Many funders do already require grantees to create a data management plan. Unfortunately, in practice the existence of such a plan has done little to encourage the real-world sharing of data. Anecdotal evidence suggests that requests sent to labs requesting access to data under the terms of data management plan are often ignored, or rejected on spurious grounds. Therefore we would recommend that funders do not place too much reliance on such plans, but rather, consider explicit mandates relating to data deposit.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?
See response to question (4).

(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?

Ultimately, economic incentives/sanctions could be applied by agencies, but hopefully this would not prove necessary.

The examples of clinical trial registration, and DNA sequence/protein structure databank deposition demonstrate that compliance with a specific data sharing/publishing policy can potentially be easily verified by publishers, by requiring the relevant accession number or other permanent identifier to be provided by the author, as a condition of publication. With appropriate cross-publisher support, this approach could be applied to many other types of data.

(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?

Agencies should ensure data are available in formats and under licensing terms which facilitate integration and re-use. Offering convenient Application Programming Interfaces (APIs) for access to data will also help to stimulate use (See for example: http://www.guardian.co.uk/open-platform). The use of open formats should be encouraged, though support for widely adopted proprietary file formats is also often a pragmatic necessity. Data should be available under licensing terms which remove any concerns about legal barriers to data integration and reuse (see below). Value-added and enhanced products and services can be built on open content, and this can drive the discovery of new knowledge.

(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?

To eliminate potential legal impediments to integration and re-use of data and to help enable long-term interoperability of data an appropriate license or waiver specific to data should be applied. There are a number of licenses for open data, of which the Creative Commons CC0 license is perhaps the most widely recognised. Under CC0, an author waives all of his or her rights to the work worldwide under copyright law and all related or neighboring legal rights he or she had in the work, to the extent allowable by law.

CC0 overcomes the challenge that the CC-BY Attribution license, widely used by open access publishers including BioMed Central, is not always suitable for data reuse, because a derivative built on data may take work from many thousands of sources, making the attribution requirement extremely burdensome and often simply not feasible.

So, in the case of data reuse, rather than relying on copyright law to ensure credit is given, it seems more appropriate to rely on the established academic cultural norms as to when attribution (citation) of another scientist's work is appropriate. See: http://pantonprinciples.org/faq/ In most cases attribution will be required to ensure reliability and validity of the secondary work. For example, a systematic review of data from several clinical trials would not be publishable if it did not cite and attribute its data sources.

(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.

Data standards are important for ensuring the interoperability of data between different research groups and platforms, and for enabling data to be reused efficiently. There are numerous digital standards for scientific data,which are being catalogued by the BioSharing group at Oxford UK (http://biosharing.org/?q=standards) in partnership with the journal *BMC Research Notes* (http://www.biomedcentral.com/bmcresnotes/).

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

Federal agencies might do this by becoming involved in international working groups on data sharing and standardization. For example, we would welcome NIH participation in the Publishing Open Data Working Group led by BioMed Central which already includes representatives of several major European funders.
Collaboration with agencies in other countries on data archiving/mirroring can also be a productive approach, as the National Council for Biotechnology Information has demonstrated through its international mirroring/data exchange agreements covering the Genbank DNA sequence database, and the PubMed Central open access literature archive.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?
To reliably hyperlink between datasets and the associated peer-reviewed journal publications, the associated dataset must be permanently and persistently available, either alongside the journal article as an associated file, or preferably in an appropriate data repository. The issuing of persistent identifiers, such as DOIs, can help to ensure that links to data can remain functional long term. Citation of datasets, following the standards developed by DataCite, should be encouraged or required by journals and publishers. Space issues are sometimes used by journals to justify stringent restrictions on the number of citations that may be included in an article, which may discourage data citation, but in an

online environment such limitations are unnecessary and should generally be avoided. Explicitly citing datasets is an important mechanism to ensure that visible academic credit is gained for data publication and sharing, removing one commonly-perceived barrier to sharing and publishing of data. Journals and publishers should also provide additional tools for consistent linking between publications and the supporting datasets. For example, a number of BioMed Central journals now require the inclusion of an 'Availability of supporting data' section which clearly points readers to the location from which they can obtain the raw datasets supporting an article.
See: http://www.biomedcentral.com/about/supportingdata and http://bit.ly/zbsPRp

Journals which publish data papers – where the primary purpose of a publication is to publish a description of a dataset, rather than methods and results – are also an important means of earning academic credit for data sharing. *BMC Research Notes* is one such journal: http://www.biomedcentral.com/bmcresnotes/authors/instructions/datanote

Finally, sharing data online using appropriate standard formats and technologies to make it machine readable and semantically meaningful can help to achieve Tim Berners-Lee's vision of Linked Data on the web. See: http://en.wikipedia.org/wiki/Linked_data

Yours sincerely,

Matthew Cockerill

Managing Director,
BioMed Central Ltd