# Elsevier submission to Office of Science and Technology Policy public consultation on Public Access to Digital Data Resulting from Federally Funded Scientific Research

January, 2012

# Introduction

Elsevier’s primary mission is to advance science by providing high-quality academic publications and services. We envision a future world in which data are much more broadly managed, preserved, and reused for the advancement of science. We want to work in partnership with other stakeholders to achieve this vision.

Professional curation and preservation of data is, like professional publishing, neither easy nor inexpensive. The grand challenge is to develop approaches that maximize access to data in ways that are sustained over time, ensure the quality of the scientific record, and stimulate innovation.

- We believe rich interconnections between publications and scientific data are important to support our customers to advance science and health.
- We recognize that scientists invest substantially in creating and interpreting data, and their intellectual and financial contributions need to be recognized and valued.
- Funders too invest substantially in these data and their contributions need to be recognized and valued.
- Where publishers add value and/or incur significant cost then our contributions also need to be recognized and valued.
- There are potential new roles, and we want to embrace an active test and learn approach.
- We will be sensitive to different practices and preferences between subject areas as we test and learn.
- Any role for Elsevier would not be exclusive, and we want to work in collaboration with other stakeholders to establish a sustainable framework for the discovery and use of scientific data.

# Access to data

### Preservation, discoverability, and access

### (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?

One of the biggest challenges for scientists today is the invisibility of data. While 97% of researchers in the US report good access to research articles, only 42% of researchers in the US express satisfaction with their access levels to data[1] and datasets, and this access gap clearly hinders scientific progress. It also leads to duplication of funding as different research funders pay repeatedly to have the same experiments run and the same data captured. Leadership and coordinated action by all stakeholders is needed to improve access to data by scientists.

Elsevier recommends the following:

- Federal agencies should work with stakeholders, including research institutions, funding bodies and publishers to develop and deploy standard approaches for linking publications and data.

[1] *Access vs. Importance, A global study assessing the importance of and ease of access to professional and academic information. Phase I Results. 2010. Publishing Research Consortium* http://www.publishingresearch.net/documents/PRCAccessvsImportanceGlobalNov2010_000.pdf

- Federal agencies should encourage authors to document their data and to deposit their data with an appropriate data center or service and to make their data available for reuse by others.
- All data should be assigned persistent, unique Digital Object Identifiers (DOIs) to aid their discovery, use, and citation. DOIs permanently identify and track scholarly items on the web, and are already used to link millions of items from hundreds of publishers and societies. DOIs integrate with the OpenURL and are completely access-model neutral.
- Appropriate metadata should be generated with the data to enable understanding and reuse.
- Stakeholders, including publishers, should encourage academics to cite datasets that have been used in their research and that are available for reuse via a data curation center or service and enable linking of data to the published journal article.
- Federal agencies should work with other stakeholders on policies for long term preservation of data, and accreditation systems/standards for digital curation services.

Federal agencies should also adopt policies that encourage publishers to continue to invest in their journals and in the development of discovery tools for data. For example our article linking tools facilitate entity text-mining (e.g. Arabidopsis Viewer), pull data associated with a published research article from a data store (e.g. Genome Viewer), support visualizations of data from a data store (e.g. Protein Viewer), link from published research articles to further detail in a data store (e.g. all 3 of the previous examples), and link articles to associated data in data stores (e.g. Pangaea or DRYAD). We are able to make these sorts of investments to make data more easily discoverable and reusable because we have sustainable business models for our journals. Unsustainable public access policies for journals could undermine these efforts.

Members of the public may also wish to access scientific datasets collected/created during federally-funded research projects. We recommend that careful work is done to understand actual needs, so that effective and sustainable approaches to filling these needs are developed.

### (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?

Copyright does not protect ideas or facts, and so does not apply to raw datasets. Respect for copyright should be reflected in any policies, but should not become a smokescreen to prevent the sharing of raw datasets.

Creative investment in the presentation of data can attract copyright, for example in published journal articles. The entirety of a publication is covered by copyright, including any data presented within it, and so permission is required to use their contents unless the intended use is covered by a copyright exception.

Elsevier requests a non-exclusive license from its authors to use supplemental data if they are to be published. This policy means that supplemental data will continue to be owned / controlled by the original researcher. The researcher may, of course, have contractual obligations to his/her employer or funder that will guide whether and how these data can be reused.

A variety of data licenses have emerged which academics may wish to consider for their supplemental data, processed data, or raw data.  Elsevier believes these re-use terms should be the choice of the researcher.

Incentives, rather than mandates, are needed to overcome data access challenges.  There is currently a disparity  between a researcher's willingness to use shared data and to supply it.  Many researchers agree it is necessary, but decline to share their own data.  An example is documented in a survey prepared for the launch of the EconomistsOnline repository in 2010 that indicated a majority of economists were in favor of accessing datasets, but when asked if they would post their own data only 15% indicated that they would be prepared to do so[2].

Publishers can play a role to incentivize the deposit and reuse of data - just as we help to incentivize academics to publish by enabling them to register their scientific discoveries in widely accessed, cited, and respected journals.  We have made data available alongside publications and support initiatives to help researchers to share data (e.g. Pangaea, CCDC and DRYAD).

The publishing industry has also developed standards for inter-linking datasets and publications through the International DOI Foundation.  DOIs permanently identify and track scholarly items on the web, and are already used to link millions of items from hundreds of publishers and societies.  DOIs integrate with the OpenURL and are completely access-model neutral.

### (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?

There is great variation across disciplines with respect to how data are collected or created, analyzed, documented, stored, and shared.  For example, we see variance in the supplementary datasets submitted to underpin published journals articles (e.g. methods/protocols, raw datasets, executable code, videos of experiments, and more) across disciplines.

Terminology, taxonomy, methodology and communication vary significantly across disciplines (and often sub-disciplines).  Controlled, shared vocabulary is essential for facilitating more automated approaches to processing information to underpin scientific discoveries.  It is therefore very important for federal agencies to work with researchers, institutions, publishers and other stakeholders to develop (if necessary) and deploy appropriate subject area classifications and controlled vocabularies.

New professional data curation professionals are emerging with skills sets that combine academic expertise in a subject (or sub-discipline) with information science skills.  Support for enabling more professional data curators and data curation facilities is essential.  Initiatives such as the Archaeology Data Service, CASPAR, the Digital Curation Centre, Planets, OAIS, SHAMAN, and nestor are likely to prove useful examples of effective digital stewardship[3] including the development/deployment of shared vocabularies.

### (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?

[2] EconomistsOnline – http://www.economistsonline.org/home and http://itswww.uvt.nl/its/voorlichting/PDF/NEEO/D1.7-NEEO-Final-Report-2010.pdf  section 7.3.5

[3] PARSE.insight Science Data Infrastructure Roadmap (see http://www.parse-insight.eu/downloads/PARSE-Insight_D2-2_Roadmap.pdf)

US agencies can play a very important leadership role by convening stakeholders to work constructively together to identify and overcome barriers to the consistent curation and reuse of important scientific data.

OSTP itself could helpful ask the General Accounting Office to undertake a study of existing federal data archives and data curation centers to determine the full costs required for start-up, management, and ongoing access, preservation, and migration activity across different subject areas.

Federal agencies should provide funding to:

- Support researchers to document and deposit their datasets in data curation centers.
- Support desk-based research that involves reuse of datasets deposited by other researchers and accessible via data curation centers.
- Establish discipline-specific data curation facilities where these do not yet exist. Both the Open Archive Information System (OAIS) Reference Model and the report of a Blue Ribbon Task Force on Sustainable Digital Preservation and Access (available at http://brtf.sdsc.edu/) are both helpful in identifying best practices for sustainable and high-quality data curation services.
- Incentivize stakeholders who develop/deploy technical standards to facilitate the transparent description, identification, management, retrieval, and reuse of datasets and the integration of distributed data, text, and tools.

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

Incentives, rather than mandates, are needed to overcome data access challenges.

Publishers can play a role to incentivize the deposit and reuse of data - just as we help to incentivize academics to publish by enabling them to register their scientific discoveries in widely accessed, cited, and respected journals. We have made data available alongside publications and support initiatives to help researchers to share data (e.g. Pangaea, CCDC and DRYAD). The publishing industry has also developed standards for inter-linking datasets and publications through the International DOI Foundation.

We encourage our authors to:

- deposit their data with appropriate data centers or services at the earliest possible opportunity, and certainly by the time of publication, recognizing that academics may need an exclusive period of time to analyze their data and publish results based on these analyses.
- seek support from an experienced data curator with expertise in their subject area (e.g. for privacy issues associated with patient images used in medical research);
- register Digital Object Identifiers (DOIs) for their datasets, implement data management plans, and use open standards to facilitate interoperability and successful data curation and
- cite datasets via their DOIs to encourage the fullest possible understanding of the research objectives, design, and methods prior to access/reuse of the data that underpin the publication

In cases where authors submit data alongside their articles, publishers have developed mechanisms to facilitate their upload and to make these data available for peer review. We have taken steps to ensure that reviewers can see all the material they need to complete their review, including data and supplementary materials.

Elsevier and other publishers can also:

- Champion the importance of long term preservation of data, and accreditation systems/standards for digital curation services.
- Communicate the benefits of data curation and reuse for different stakeholders in the scholarly communication landscape including authors, funders, publishers, researchers, and university administrators.
- Deploy our expertise in certification, indexing, and linking to add value to data (e.g. for search and mining).
- Use standard vocabularies, taxonomies, ontologies, and entity resources where possible rather than inventing our own.
- Support the creation and capture of linked data during the authoring and editorial process and maintain linked data through production processes.
- Facilitate the rich linking to and from publications.

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

We agree this is an important issue and evidence-based policy making is crucial. It is essential for there to be clear business cases for, and sustainable business models, to underpin data curation. There are a number of active digital data centers and data repositories that commit to the documentation and digital preservation of their contents. More systematic study of the costs associated with these, and more rigorous attention to their long term sustainability would be extremely helpful to all stakeholders.

Federal agencies should provide funding to:

- Support researchers to document and deposit their datasets in data curation centers.
- Support desk-based research that involves reuse of datasets deposited by other researchers and accessible via data curation centers.
- Establish discipline-specific data curation facilities where these do not yet exist. Both the Open Archive Information System (OAIS) Reference Model and the report of a Blue Ribbon Task Force on Sustainable Digital Preservation and Access (available at http://brtf.sdsc.edu/) are both helpful in identifying best practice for sustainable and high-quality data curation services.
- Incentivize stakeholders who develop/deploy technical standards to facilitate the transparent description, identification, management, retrieval, and reuse of datasets and the integration of distributed data, text, and tools.

### (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?

Some work is already underway in this area through groups such as the ISO Repository Audit and

Certification Working Group. We believe federal agencies should work with other stakeholders to develop (if necessary) and deploy standards and policies in this area.

Publishers can play a role to incentivize the deposit and reuse of data - just as we help to incentivize academics to publish by enabling them to register their scientific discoveries in widely accessed, cited, and respected journals. We have made data available alongside publications and support initiatives to help researchers to share data (e.g. Pangaea, CCDC and DRYAD). The publishing industry has also developed standards for inter-linking datasets and publications through the International DOI Foundation.

### (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?

OSTP itself could helpful ask the General Accounting Office to undertake a study of existing federal data archives and data curation centers to determine the full costs required for start-up, management, and ongoing access, preservation, and migration activity.

Federal agencies should provide funding to:

- Support researchers to document and deposit their datasets in data curation centers.
- Support desk-based research that involves reuse of datasets deposited by other researchers and accessible via data curation centers.
- Establish discipline-specific data curation facilities where these do not yet exist. Both the Open Archive Information System (OAIS) Reference Model and the report of a Blue Ribbon Task Force on Sustainable Digital Preservation and Access (available at http://brtf.sdsc.edu/) are both helpful in identifying best practice for sustainable and high-quality data curation services.

Federal agencies should also:

- Encourage the re-use of publicly funded and accessible research datasets by making them available under unambiguous non-exclusive licenses
- Commit to develop a culture of ethical re-use of data, for example by banning those who willfully misrepresent or distort data created by others from receiving grant funds
- Incentivize (e.g. by respecting copyright and other intellectual property rights) stakeholders who develop/deploy technical standards to facilitate the transparent description, identification, management, retrieval, and reuse of datasets and the integration of distributed data, text, and tools. Create a level-playing field for different sustainable business models that emerge for these products and services

### (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?

Authors should be encouraged to registered DOIs for their data. With a DOI in place datasets become citable and can be linked to other data and to publications. The DOI contains the standard metadata which will supply the required metadata, including author, affiliation, related articles, etc. In addition, as the DOI has successfully worked with millions of published journal articles, the use of the DOI facilitates easy linking across objects, including data set and published articles.

**Standards for interoperability, re-use and re-purposing**

### (10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data?

Digital curation standards are mostly in formative stages, and the following are good examples of active projects in this area:

- Opportunities for Data Exchange (www.ode-project.eu)
- DataCite (http://datacite.org/)
- APARSEN (http://www.alliancepermanentaccess.org/index.php/currentprojects/aparsen/)

### (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?

Recently NISO-NFAIS released Recommended Practices for Supplemental Journal Article Materials. Elsevier has been an integral partner in the development of these guidelines.

More broadly CrossRef and the International Digital Object Identifier Foundation have been transformational in developing standards to link publications and data in a persistent way. The history of these successful standards organizations is perhaps helpful to relate.

The International DOI Foundation was created in 1998 to support electronic publishing through the development and promotion of the DOI (Digital Object Identifier) System as a common infrastructure for content management. The Foundation is a registered not-for-profit organization, controlled by an Executive Board elected by the members of the Foundation.

In 1999 a technical demonstration was made of the DOI at the Frankfurt Book Fair. Representatives of the leading scientific, technical, and medical publishers recognized in this prototype that a lookup system based on the Digital Object Identifier (DOI) held the key to a broad-based and efficient journal reference linking system. They took the unusual step of joining together as the non-profit, independent Publishers International Linking Association Inc. (PILA), which was incorporated in January 2000 and CrossRef went live as the first collaborative reference linking service in June 2000.

CrossRef’s mission is "to be a trusted collaborative organization with broad community connections; authoritative and innovative in support of a persistent, sustainable infrastructure for scholarly communication." CrossRef's general purpose is to promote the development and cooperative use of new and innovative technologies to speed and facilitate scholarly research. CrossRef's specific mandate is to be the citation linking backbone for all scholarly information in electronic form. CrossRef is a collaborative reference linking service that functions as a sort of digital switchboard. It holds no full text content, but rather effects linkages through CrossRef Digital Object Identifiers (CrossRef DOI), which are tagged to article metadata supplied by the participating publishers. The end result is an efficient, scalable linking system through which a researcher can click on a reference citation in a journal and access the cited article.

In parallel the International DOI Foundation has continued to evolve. Millions of DOIs have been assigned, and millions are accessed each month. DOI registration agencies have been appointed across the globe. CrossRef was the first registration agency and has been followed by Office des publications EU (OPOCE) for government documents, *m*EDRA for multilingual resources, EIDR for movie and television assets, and DataCite for scientific data.

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

Federal agencies could helpfully endorse existing initiatives and standards such as those mentioned in our response. Inter-working with the international academic publishing community on data issues is best done through the International Association of Scientific, Technical, and Medical Publishers (STM) and through the involvement of Elsevier and other publishing companies in working groups.

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

As stated the publishing industry has developed standards for inter-linking datasets and publications through two not-for-profits: the International DOI Foundation and CrossRef. Elsevier is working with a number of organizations to link to appropriate data. For example, in earth sciences, we provide interlinking to earth sciences data through Pangaea and our published journal articles. Support by federal agencies for the Digital Object Identifier as the persistent identifier of choice for both data and publications would significantly improve the linking of publications and associated data.

Collaboration with DataCite could also be extremely helpful. This is an organization that aims to increase acceptance of research data as legitimate citable contributions to the scholarly record, and support data archiving to permit results to be verified and re-purposed for future study. DataCite is currently engaged in the process of helping researchers find, identify, and cite research datasets; providing persistent identifiers for datasets, workflows and standards for data publication; and enabling research articles to be linked to the underlying data. To achieve these goals, they are currently working primarily with organizations that host data, such as data centers and libraries.

Y.S.Chi

Youngsuk Chi
Chairman
Elsevier
360 Park Avenue South
New York, NY 10010