January 12, 2012

Catherine Casserly; cathy@creativecommons.org
Timothy Vollmer; tvol@creativecommons.org
Creative Commons
Mountain View, CA

Re: OSTP Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research [http://federalregister.gov/a/2011-28621]

Creative Commons (CC) is pleased to submit comments to the Office of Science and Technology Policy's Request for Information (RFI) on the topic of Public Access to Digital Data Resulting from Federally Funded Scientific Research. Creative Commons (http://creativecommons.org) is a 501(c)(3) U.S.-based nonprofit corporation dedicated to making it easier for people to share and build upon the work of others, consistent with the rules of copyright. CC develops legal and technical tools used by individuals, cultural, educational, and research institutions, governments, and companies worldwide to overcome barriers to sharing and innovation. Creative Commons operates globally. The international CC affiliate network consists of 100+ affiliates working in over 70 jurisdictions, and there are over 500 million CC-licensed works available on the web.

Thousands of academic researchers release print and digital datasets, journal articles, and educational materials under Creative Commons copyright licenses and waivers, allowing those materials to be easily found, accessed, reused and re-purposed around the world. CC licenses offer a flexible set of permissions so that authors and publishers can release their data while ensuring—if desired—that they receive attribution for their work, or explicitly place their research into the public domain in cases where copyright does not apply (e.g. for collections of factual data).

We answer specific questions laid out in the RFI below. Note that we have not answered all the questions, as we believe there are other organizations and stakeholders with greater expertise in those areas.

Question 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
The federal government should establish policies that insure the public has cost-free, unimpeded access to the digital data resulting from federally funded scientific research. Data should be made accessible in a manner that explicitly communicates the rights available to data re-users. The public should have the right to reuse publicly funded data free of any legal restrictions, or with the only restriction being that the source of the data is credited.

Access to this data should be made available as soon as possible, with due consideration to confidentiality and privacy issues, as well as the researchers' need to receive credit and benefit from the work (for instance, a short embargo period for data is not antithetical to the recommendation of public access, as original researchers invest significant effort into the creation and analysis of data and would like to be able to capitalize on their efforts).

If the federal government wants to maximize the impact of digital data resulting from federally funded scientific research, it should provide explicit, easy-to-understand information about the rights available to the public. The simple process of posting federally funded scientific data in a publicly accessible repository on the web so that they can be viewed and downloaded will not realize the full reuse potential of the data. Even where it is the intention that digital data created with public dollars be widely shared, as long as those materials are not clearly marked with information describing the rights and permissions under copyright law, the use of those data will be diminished and impact of public investment lessened.

With a renewed attention to clarifying the rights available to data re-users in advance, the federal government can increase the speed of scientific discoveries, promote innovation, and support new economic opportunities. And as scientific researchers, creative startups, and traditional commercial businesses are granted the reuse rights they deserve, these and other groups can best leverage federally funded digital data to advance the scientific enterprise. Unimpeded access to the digital data resulting from federally funded scientific research can increase the speed and variety of scientific discoveries and boost U.S. competitiveness by encouraging the development of new products and services.

The federal government can grant these permissions to the public by 1) dedicating the data to the public domain or 2) adopting a liberal licensing policy where at most downstream data users must give credit to the source of the data.[1] Creative Commons offers public domain tools and licenses to help accomplish these goals. CC0 (read "CC Zero") is a tool that allows data publishers to dedicate data to the public domain by waiving all rights to the work worldwide under copyright law.[2] Waivers such as CC0 are the gold standard with regard to global interoperability and innovation potential because it removes all barriers to reuse from data. In certain domains, such as science and public sector data, there are important reasons to consider using waivers like CC0. Waiving copyright and related rights in a domain like science eliminates all uncertainty for potential re-users, encouraging maximal reuse. Echoing the Panton Principles,

> "[I]n science it is STRONGLY recommended that data, especially where publicly funded, be explicitly placed in the public domain via the use of the Public Domain Dedication and Licence or Creative Commons Zero Waiver. This is in keeping with the public funding of much scientific research and the general ethos of sharing and re-use within the scientific community."[3]

[1] By public domain, we mean the legal concept whereby content is not protected by copyright.

[2] http://creativecommons.org/publicdomain/zero/1.0/

[3] http://pantonprinciples.org/

Use of explicit waivers of rights is necessary to avoid uncertainly around what uses may be made of data, e.g., to integrate with other data to create new datasets or reformat it to support its long-term preservation. Even in the frequent case of data that is automatically in the public domain, e.g., because it is strictly factual in nature and so not subject to copyrights under U.S. law, explicit declarations of the rights-free status of the data is necessary since researchers are not expert in the nuances of copyright law and often make incorrect assumptions about their rights to reuse data. In other cases datasets of mixed factual and creative content, making their legal status murky even to legal experts. This uncertainly can result in researchers avoiding reuse of existing data for fear of inadvertently violating a non-existent copyright or to avoid involving legal counsel to determine possible reusability.

The situation is even more complex for international research collaborations where different countries and legal jurisdictions have different or conflicting laws and policies regarding protection of data, such as the sui generis data rights that apply in the European Union but not in the U.S. For data produced by such international collaborations (an increasingly common phenomenon) a rights waiver such as CC0 is the only existing method of making the data unambiguously legally available to U.S. researchers for reuse.

Where attribution is desired, the federal government might consider requiring the Creative Commons Attribution 3.0 (CC BY) license for digital data resulting from federally funded scientific research. CC BY is a copyright license that grants permission to the public to reproduce, distribute, perform, display or adapt the licensed materials for any purpose so long as the user gives attribution to the author or as otherwise directed by the copyright holder.[4]

Question 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
The current climate around intellectual property (IP) is increasingly one of ratcheting up enforcement and maximization of copyright and other IP rights. However, it's important to understand the underlying priorities of scientific researchers to determine whether the default copyright regime reflects their needs and supports their academic endeavors. The primary motivation for scientific researchers is "to inform others about their work," thus contributing to the scholarly canon and promoting the advancement of science.[5] Scholarly authors wish to have their research read, consumed, and cited. To do their work, scientific researchers must acquire and use data from a wide variety of sources, and need to know how they may reuse data—either by itself—or more likely, in combination with other datasets. So, it makes sense to be able to provide the widest access with the fewest encumbrances to digital data resulting from federally

[4] http://creativecommons.org/licenses/by/3.0/

[5] Hansen, et al., *Intellectual Property Experiences in the United States Scientific Community*, 2007, p. 8. Available at http://sippi.aaas.org/Pubs/SIPPI_US_IP_Survey.pdf.

funded scientific research. For this access to be meaningful, scientists need clear and unambiguous information (metadata) about the rights they are granted in using the data for their research.

Rather than focusing on Intellectual Property as the means for capturing the value of research performed, mechanisms are needed to provide academic and public credit to researchers for their work, to the agencies that funded the work, to publishers who promoted the results, and to other stakeholders such as data archives that made the data available. Such mechanisms will require broadly available and standard infrastructure such as persistent and globally unique identifiers for the data, researchers, institutions, funding agencies, and other stakeholders that allow data use and reuse to be tracked over time and made part of the scholarly record.

Some argue that placing digital data resulting from federally funded scientific research directly into the public domain is suboptimal because it will be impossible to track the attribution, provenance, and credibility of the published data. But these are separate issues: citation has always been a normative practice that scholars enforce through social mechanisms. Whether the cited research is in a peer-reviewed publication or posted on the public Web does not affect its citability. Citation and credibility are made possible via mechanisms like associated metadata and disciplinary norms for determining quality and provenance. Even licenses that require "attribution" (common for copyrighted creative works) does not guarantee appropriate citation, since licenses and attribution are legal tools that are usually enforced via lawsuits. A public domain policy for the sharing of digital data resulting from federally funded scientific research is not misaligned with the need for citation and quality control.

Question 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
With regard to the management of copyright, the federal government should adopt common, global standards for the communication of rights for digital data resulting from federally funded scientific research. Scientists and researchers around the world currently use CC tools such as CC0 to release data into the public domain and CC BY (where copyright adheres to the data) to allow its reuse with simple attribution. Standardized, public licenses and waivers should be strongly recommended instead of customized solutions for each discipline because it increases compatibility and clarity to the end user in how they may use the data and combine it with other datasets.

More generally, federal agencies should adopt a standard data format for the communication of rights and permissions for digital data resulting from federally funded scientific research. The Creative Commons Rights Expression Language (ccREL) is a specification for how license information is described using RDF and how licensing information is attached to works.[6] ccREL

[6] RDF stands for Resource Description Framework. RDF is a "family of W3C specifications originally designed as a metadata data model. It has come to be used as a general method for conceptual description or modeling of information that is implemented in web resources, using a variety of syntax formats." For more information, see http://en.wikipedia.org/wiki/Resource_Description_Framework

is the standard recommended by Creative Commons for machine-readable expression of copyright licensing terms and related information, so that content and data can be exposed via search engines like Google.[7] ccREL is embedded within CC license metadata and the CC0 public domain dedication tool metadata.

Question 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
Especially for the scientific community, digital data resulting from federally funded scientific research should be made immediately available in the public domain in order to speed the discovery of cures for diseases and promote the advancement of science.[8] Immediate access speeds up the research and development cycle, thus leading to faster development of value-added research, products and services.

Clear, unambiguous policies around the management and sharing of digital data resulting from federally funded scientific research should be provided to federal grantees, and the federal government should ensure compliance with these policies. As discussed above, federal grantees that create digital data should be required to place that data in the worldwide public domain using CC0 or made available with the minimal attribution requirement via CC BY. Adopting a clear data management policy should help rectify the widespread over-reach of many data provider Terms of Use statements. Many of these Terms of Use statements curtail re-use through overly restrictive licensing agreements. In addition, the federal government can help educate grantees about copyright with regard to data. For instance, strictly factual information does not rise to the level to warrant copyright protection, and should not be claimed as copyrightable material.

Question 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
Digital data resulting from federally funded scientific research should be shared in the worldwide public domain using the CC0 public domain dedication tool or made available via the liberal CC BY license. Both CC0 and CC BY metadata include ccREL specification for machine-readable expression of copyright licensing terms and related information. Ensuring that licensing metadata

[7] http://wiki.creativecommons.org/images/d/d6/Ccrel-1.0.pdf

[8] The Tuberculosis Commons (TB Commons) Initiative Open Innovation Team "believes that open models can accelerate knowledge turns resulting in a faster drug development process." Content provided by end-users to the site is released into the public domain using the under the CC0 Public Domain Dedication tool. See http://www.tbcommons.org/initiative/.

and related information is clearly communicated to humans and machines promotes interoperability, reuse, and repurposing of digital scientific data.

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

Comment 13
For the OSTP RFI on Public Access to Peer-Reviewed Scholarly Publications Resulting From Federally Funded Research, we recommended that scholarly articles created from federally funded research be released under full open access. Full open access policies will provide to the public immediate, free-of-cost online availability to federally funded research without restriction except that attribution be given to the source.[9] The standard means for granting permission to the public is through a CC BY license. This license is aligned with the principle of full open access because it allows rights (including commercial rights) to be communicated with the only requirement that users give credit to the rightsholder. Full open access policies for publications are necessary to maximize the impact and reach of federal research investments, and promote scientific discovery and economic activity. By adopting a common legal licensing framework for publications and data, the federal government can ensure maximum interoperability and reuse of federally funded research outputs.

We thank OSTP for the opportunity to provide comments to this RFI, and we're happy to answer any other questions you may have.

Sincerely,

Catherine Casserly, CEO, Creative Commons
Timothy Vollmer, Policy Coordinator, Creative Commons

[9] Carroll, Michael. *Why Full Open Access Matters.* PLoS Biology, November 2011. Volume 9, Issue 11. Available at http://www.plosbiology.org/article/info:doi%2F10.1371%2Fjournal.pbio.1001210.