From: Ginny Steel
Date: January 9, 2012
To: digitaldata@ostp.gov
Subject: Response to RFI on digital data

To Whom It May Concern,

I am writing in my capacity as the Chair of the University of California Council of University Librarians to submit the attached comments in response to the RFI on digital data issued by the Office of Science and Technology Policy in late 2011. Collectively, the UC libraries make up the largest research/academic library in the world, with over 35 million volumes in our holdings as well as significant digital collections. The 10 campuses and the California Digital Library work together to expand the scope of our collections, improve access to information, and develop alternative modes of scholarly communication so that all faculty, students, and staff have access to the resources they need to support their teaching, learning, research, and service. One of the primary goals for UC's Council of University Librarians in 2012-15 is to support efforts to change the current, unsustainable models of scholarly communication that are having a calamitous and wide-reaching effect on academic library budgets. In light of our serious concern that there be effective and sustainable national and international stewardship solutions and the importance of providing public access to data resulting from federally funded research, we offer the attached recommendations.

Sincerely,

Ginny Steel
University Librarian and
Chair, UC Council of University Librarians
University of California, Santa Cruz
1156 High Street
Santa Cruz, CA 95064
vsteel@ucsc.edu
831 459 2076
831 459 8206 fax

To: **Office of Science and Technology Policy (digitaldata@ostp.gov)**
From: **University of California Libraries**
Subject: **Response to the OSTP RFI on Public Access to Digital Data**

Dec. 23, 2011

Introduction

The University of California applauds the recent Request for Information issued by the Office of Science and Technology Policy (OSTP) [1] to solicit comment and recommendations on approaches for ensuring long-term stewardship of, and broad public access to, digital data resulting from federally funded research. These research data play a fundamentally important role in the ongoing practice of scientific inquiry, discourse, and advancement, with many broader societal dependencies, relationships, and consequences, both direct and indirect, in commerce, education, and culture. However, data represented in digital form is inherently fragile with respect to the ever increasing pace of disruptive technological and institutional change. Thus, research data must be placed under careful, comprehensive, and proactive management and stewardship – in short, it must be properly *curated* – in order to ensure that it remains available for use, sharing, and re-purposing by current and future generations of scientists and scholars. Twenty-first century research is driven by the continual, exponential advances of computation, storage and bandwidth. It is imperative that data produced in this environment neither be lost nor remain unused. Unused data has no value.

The University of California believes that truly effective and sustainable stewardship solutions depend upon adherence to five broad strategic imperatives:

- *Know what you have* ("You can't manage what you don't measure")

  The objects of stewardship must be clearly documented in terms of significant form or structure, scientific meaning, and desired behavior in order to facilitate successful preservation, discovery, and use.

- *Express and share that knowledge widely* ("It takes a village")

  Comprehensive description of research data must itself be the object of affirmative custodial care. This information should be formally documented as an aid to widespread discoverability and potential transfer of stewardship responsibility. Note that the proper emphasis of all stewardship activity is on the persistence of the research data (and its concomitant description) itself, not the systems in which that data is managed, which are inherently ephemeral.

- *Make lots of copies* ("Be redundant, be redundant")

  Both coarse and fine grained replication and redundancy in all aspects of the stewardship infrastructure – technical, curatorial, and procedural – are one of the most powerful means to minimize the potential for debilitating single points of, systemic, or correlated failures.

- *Protect the copies* ("Trust, but verify")

An effective stewardship infrastructure incorporates affirmative safeguards into all technical systems and workflows, such as storage-level use of error correcting codes, media refresh, and fixity audit; data center high availability hosting and operational practices; and established procedures for obsolescence detection and mitigation.

- *Plan and watch* ("Proactive when you can, reactive when you must")

  Successful stewardship outcomes require a comprehensive program of services, policies, and practices. It is important to be prepared for anticipated eventualities through action plans, ongoing technology watch, and stakeholder engagement activities, while also being alert for, and responsive to, unexpected conditions requiring attention.

Preservation, Discoverability, and Access

The OSTP RFI properly casts preservation and access as complementary, rather than disparate, stewardship activities: access being dependent upon preservation up to a *point* in time, while preservation ensures access *over* time. Furthermore, data access enables new forms of scholarly collaboration and communication that will lead to revolutionary means of exploring knowledge. This can only be achieved with effective preservation and discovery. The full range of necessary stewardship policies and practices encompasses both technical and organizational facets. While effective technical systems are a necessary foundation, truly effective and sustainable stewardship solutions for long-term preservation and access must rely significantly on human competencies, analysis, and decision making.

(1) Encouragement of Access and Preservation

One of the most direct positive impacts that federal granting agencies can have in encouraging desirable behaviors for the long-term management, preservation, and sharing of research data is through instituting appropriate requirements as a pre-condition, and an auditable post-condition, for funding. (Current NIH and NSF data management requirements serve as exemplars.) Although compliance to new behaviors for preservation and sharing will initially be driven by external requirements, over time, as they are increasingly integrated and internalized into researchers' work practices, they will eventually come to be accepted merely as normative patterns of scientific activity. It would be useful for granting agencies to couch the intent underlying their requirements in terms of the many tangible benefits that may personally or organizationally accrue from following these practices, such as increased opportunities for collaboration, directed or serendipitous discovery, publication, and citation [2].

(2) Protection of Intellectual Property Interests

US law does not provide copyright or intellectual property protection to facts or factual data. US Copyright Office Circular Number 1 states, "[c]opyright does not protect facts, ideas, systems, or methods of operation, although it may protect the way these things are expressed." [1]

[1] See [3]. US copyright law contains additional exemptions and restrictions, and "several categories of material are generally *not* eligible for federal copyright protection": for example, "works that have not been fixed in a tangible form," "titles, names, short phrases and slogans," work created by the federal government, "ideas, procedures, methods, systems processes, concepts, principles," "works consisting

There are three questions that must be asked to determine whether an intended use of data triggers copyright or any intellectual property protection. The first question in respect to data and copyright and intellectual property protection is whether the intended use triggers copyright. Does the intended use of data fall within one of the copyright owner's exclusive rights? Title 17 USC Section 106 defines the exclusive rights of copyright owners [4], and if the intended use does not fall within those exclusive rights, then the use is permitted unless there is a prior contract that prohibits it. The mere extraction or copying of "ideas, facts, processes, or methods" or federal government information that is excluded from copyright protection does not trigger copyright (17 USC 102(b), 17 USC 105, and 17 USC 106). The reproduction and transmission rights of copyright owners cover only the making and or distribution of "copies," and not every "copy" qualifies. Specifically, "copies" that are not capable of being "perceived, reproduced, or communicated" and that are not sufficiently "fixed" do not qualify. The US Copyright Office defines the word "copy" for purposes of US copyright law as follows: a copy is "[t]he material object, other than a phonorecord, in which the copyrighted work is first fixed, and from which the work can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device" [5].

The second question regarding data and copyright and intellectual property protection is, even if the intended use does trigger copyright, does it fall within copyright's limitations or exceptions such as Fair Use (17 USC 107), the Exemption for Libraries and Archives (17 USC 108), or other exemptions? If the intended use falls within an established exemption, then it does not require any further copyright or intellectual property protection.

The third question is, even if copyright does cover the intended use and the intended use is not covered by an exemption, is the work available under license such as Creative Commons or other public licenses? If so, then the use may fall under the scope of the license. Again, this use of data is pursuant to the scope of a license or contract, so no further copyright protection is necessary or warranted.

One growing use of digitized data is that of non-consumptive research. Non-consumptive research relies on computation and, typically, computational queries to a database or collection of digitized data in which the content and corpus of the database is not accessed for display or reading. This new form of research and discovery is a user's right, not a copyright owner's right, unless the user agrees to a contract or license that restricts computational and non-consumptive research or unless the database owner employs technology to prevent, prohibit, or restrict a user's right to engage in non-copyrighted computational research and discovery.

Clifford Lynch, director of the Coalition for Networked Information, notes that opportunities for new knowledge creation, production, and innovation require "new ways to think about the scholarly literature (and the underlying evidence that supports scholarship) as an active, computationally enabled representation of knowledge that lives, grows" and further "suggests ways in which information technology can accelerate the rate of scientific discovery and growth of scholarship" [6]. Given the enormous potential for public access and use of preserved and accessible publicly funded data, it would be a serious educational, scientific, and economic

---

entirely of information that is common property and containing no original authorship (for example: standard calendars, weight charts, tape measures, and rulers, and lists or tables taken from public documents or other common sources)" [3].

setback if opportunities for the creation of new knowledge and discoveries were limited due to incorrect interpretations and applications of existing copyright law and overly broad expansion of intellectual property protection. Current copyright, intellectual property, and licensing regimes provide more than adequate protection for creators, scientists, federal agencies, and publishers with respect to digital data.

(3) Disciplinary Differences

The broad strategic imperatives underlying long-term stewardship certainly apply, at least in the abstract, regardless of scientific discipline. Even at the more tactical level, many specific activities can be performed without particular regard to scientific discipline, for example, persistent identification, citation, storage, replication, fixity, and technical characterization. Most meaningful disciplinary differences come into play in terms of formats and tools, descriptive practices, methods and methodology, and modes of discovery and use, which can be highly specific to both broad and niche communities of practice. Funding agencies should recommend the widest possible use of the most common data formats and analytical tools. Comprehensive information documenting availability, deployment, and use of all such formats and tools, whether common or not, should be publicly available in well-known technical registries, such as PRONOM [7] or the Unified Digital Format Registry (UDFR) [8], being developed by University of California Curation Center (UC3) at the California Digital Library (CDL) as part of the Library of Congress's National Digital Information Infrastructure Preservation Program (NDIIPP).

Many of the discipline-specific descriptive practices have come into being as the result of long collaborative experience and represent an optimization of effort and productivity. As such, the development and codification of such practices should be accepted by federal agency policies, albeit with encouragement of their establishment with the widest possible scope. The primary challenge raised by narrow descriptive standardization comes from the rise of cross-disciplinary research, particularly when diverse communities rely on substantially inconsistent practices. Linked data and other semantic web technologies hold out great promise for facilitating automated cross-disciplinary discovery, and federal granting agencies should encourage the development and use of appropriate ontologies for this purpose.

Another potential disciplinary distinction is the "big-data/small-data" divide. Extremely large datasets, especially those arising in the hard sciences through large-scale simulation or fine-grained observational instruments, pose significant administrative and technical challenges. Funding agencies should continue their efforts to support, publicize, and move onto a sustainable footing high performance computing facilities. However, a recent survey of over 1700 *Science* peer reviewers reported that the largest dataset generated or used locally by over 48% of the respondents was less than 1 GB in size [9]. Thus, while the problems of big data receive much public scrutiny, it is important that small data, particularly in the life and social sciences, whose usage is undoubtedly much more diffuse, continues to receive adequate attention and support. It should also be noted that the Humanities is rapidly changing as well. Modern scholars working on interpreting the cultural fabric of our world are doing so in a largely data driven environment, so much so that recent trends in humanities' scholarship are taking the shape of data curation as a publication [10]. It is imperative for all forms of scholarship that data be prepared and shared with a networked mindset.

(4) Costs and Benefits

The allocation of scarce curatorial resources, whether financial or otherwise, is always based on evaluations of the current and future value proposition for the curated data. Evaluation criteria should include the scientific value, scope of applicability, and degree of uniqueness and reproducibility. Since any assessment of future value can be problematic with respect to fundamental underlying assumptions and ever-changing conditions, it is important that all plausibly useful research outputs are subject to minimally sufficient baseline practices, and that there is some level of ongoing curatorial assessment to select data deserving of added value attention in light of evolving circumstances.

(5) Stakeholder Contribution

Many well-established memory institutions – libraries, archives, and museums – have developed deep expertise, experience, and resources for dealing with the long-term preservation of and access to cultural heritage material, which in many cases can be directly applied to the stewardship of research data. These institutions should be encouraged to make available their preservation and curation systems and services to the research community. Many research universities, such as the University of California, already have mature local, centralized, and consortial solutions in place to address the long term stewardship needs of the University's digital assets.

As an example, an international group of academic libraries, research projects, and government and non-profit organizations have collaborated under the leadership of UC3, UCLA, and UCSD to create the DMPTool [11], a publicly available online system that aids researchers in creating data management plans meeting funder requirements. This system is configured to provide campus-specific guidance and advice regarding the availability of services appropriate for long-term stewardship.

(6) Funding Mechanisms

Based on UC3's experience working with data owners internal and external to the University of California, a "pay as you go" model for stewardship services may not be appropriate in all contexts. The majority of research data derives from grant funded project activities, which leaves little provision for sustainable funding beyond project completion. Most researchers are therefore eager to embrace an alternative "pay once" model whose charges can be built into grant proposals. In order to be sustainably viable, however, it is important that stewardship service providers understand the full gamut of lifecycle costs [12][13]. Note that this may require service providers to take on an unfamiliar fiduciary role in the long-term management of endowment funds. These funds should be dedicated for the purpose of sustaining the research data in perpetuity, and not be available for reallocation towards other service provider priorities.

(7) Policy Compliance

Voluntary compliance to Federal stewardship policies can be enhanced by proactively bringing all of the affected stakeholders – researchers, service providers, funders – into the process of developing those policies, so that all parties can feel that their particular needs and concerns have been considered and incorporated. Reporting requirements should be kept to a minimum and based on commonly accepted objective measures. If possible, these measures should be

independently verifiable, as suggested by the "neighborhood watch" concept proposed by the UC Curation Center [14], so that compliance can be determined with minimal intrusion and cost. Since funder requirements have only recently started to be promulgated, many in the research community are not fully aware of their intention or significance. Affirmative efforts are needed, both by funding agencies and local institutions, to raise the awareness of all of the implications of the new policy requirements.

(8) Innovative Use

The use and re-use of research data is largely dependent upon four factors: knowing that the data exist; knowing where to get it; having it delivered in a form that is easily integrated, either directly or through minimal transformation, into local work practices; and ensuring the long-term public access and use of publicly funded data. The first and second are questions of widespread dissemination of descriptive metadata in appropriate intra- or cross-domain discovery services integrated with datacenters and access repositories. The third factor is more difficult as it is facilitated by growing conformance to common data practices regarding the acquisition and representation of data. While some scientific communities have broad internal agreement regarding these practices, in many cases idiosyncratic usage is the norm. Funding bodies should consider supporting efforts at standardizing and codifying disciplinary practice on the broadest terms as part of a more general encouragement of translational research and centers of excellence. Finally, funding bodies and government agencies should require long-term public access to and use of publicly funded data.

(9) Attribution

Providing scientists with assurance of appropriate attribution and credit for making available their research output can be facilitated through support for formal data publication. While the historical practice has been to provide public visibility to only one of the many outputs of a research program – the summarizing paper or conference presentation – there is no reason why the other data products could not be similarly treated, wrapping those products in the familiar façade of academic publication [15]. Providing datasets with persistent identifiers and descriptive citations enables the entire scholarly publication infrastructure to come into play to provide sophisticated aggregation, indexing and abstracting, enhanced discovery, and attribution, all of which should combine to encourage more widespread use and repurposing.

Standards for Interoperability, Re-Use, and Re-Purposing

Effective solutions for the long-term preservation of, and access to, digital research data will almost certainly involve a community of committed stakeholders. Increased access actively promotes preservation outcomes: data that are used widely or frequently are much more likely to receive the appropriate stewardship attention. The global distribution of stewardship expertise and experience will always be uneven and it is natural to assume the orderly or ad hoc development of specialized centers of excellence offering tools, best practice recommendations, and services to the broader community. Furthermore, preservation is a serial, rather than a one time, activity. Given the inevitable evolution of organizational mission, resources, and priorities, over any sufficiently extended period of time it is likely that the responsibility for the physical and curatorial custody of research data will be transferred from institution to institution. Thus, broad

community conformance to accepted standards – both de facto and de jure – is a necessary concomitant to sustainable long-term success in preservation and sharing. We are in a research environment where technology is of secondary importance; information – and the widest possible distribution and sharing of that information – is what matters.

(10) Interoperability Standards

As discussed in the context of question (8), the use of research data is predicated on three factors: knowing that the data exist, knowing from where the data are available, and having it made available in a form that is easily integrated into local workflows. These suggest the need for common standards for data description, publication, discovery, and representation formats. Data description must be supported at sufficiently fine grain to enable direct and, ideally, automated determinations of the suitability of a given dataset for a particular local purpose. In other words, descriptive practice should extend down to the level of individual variable fields, units of measure, spatial and temporal coverage, normalization procedures, etc. It is also imperative that research data move from inside the academy to the outside. This suggests that descriptive standards should be developed in a manner that is usable to those with deep disciplinary expertise as well as broad synoptic understanding.

(11) Standards Process

The ecological science community has been successful in fostering a number of open source informatics standards and projects, including the EML metadata standard and its attendant tools. The ecoinformatics.org organization [17] provides a central platform and lightweight process for harnessing the voluntary collaborations of domain scientists in areas of broad concern and applicability. In accordance with open source principles, this work is self-directed and self-governing, leading towards community empowerment and commitment.

(12) Standards Coordination

Coordination and standardization of scientific data practices can be best performed on a disciplinary, rather than governmental, basis, taking advantage of long-standing disciplinary channels for intra- and cross-domain discourse and collaboration. Conformance to standards and best practices will be greatest when those practices are perceived as arising from within the community of concern and practice, rather than being imposed externally. That being said, governmental agencies and funding bodies can play an important role in encouraging and funding disciplinary working groups constituted on the broadest possible basis. Today, and certainly in the future, many innovative avenues of scientific advance result from research that transcends traditional disciplinary boundaries. It is therefore important that cross-disciplinary efforts at common standardization or standardized cross-walks be established and encouraged.

(13) Data and Publication

The DataCite consortium develops and promotes DOI-based standards and services for data publication, and is working with the scholarly publishing community to provide greater visibility to research data in the familiar context of discovery portals and indexing and abstracting services [18]. Federal funding decisions should be planned to encourage the support by publishers and data repositories of bi-direction linking between traditional academic publications and the data that underlies their analysis and conclusions, with all of the attendant mechanisms and incentives

for citation, attribution, and impact analysis.

References

[1] Office of Science and Technology Policy (2011), "Request for Information: Public Access to Digital Data Resulting from Federally Funded Scientific Research," 71 *Federal Register* 214 (4 November 2011), pp. 68517-68518.

[2] Heather A. Piwowar, Roger S. Day, and Douglas B. Fridsma (2007), "Sharing detailed research data is associated with increased citation rate," *Public Library of Science* 2:3 <http://dx.doi.org/10.1371/journal.pone.0000308 >.

[3] US Copyright Office, Circular 1, *Copyright Basics* <http://www.copyright.gov/circs/circ01.pdf>.

[4] 17 USC Section 106 <http://www.copyright.gov/title17/>.

[5] US Copyright Office, *Definitions* <http://www.copyright.gov/help/faq/definitions.html>.

[6] Clifford A. Lynch, "Open computation: Beyond human-reader-centric views of scholarly literatures," *Open Access: Key Strategic, Technical and Economic Aspects*, ed. Neil Jacobs (Oxford: Chandos Publishing, 2006), pp. 185-193.

[7] National Archives [UK] (2001), *PRONOM* <http://www.nationalarchives.gov.uk/PRONOM>.

[8] UC Curation Center (2011), *Unified Digital Format Registry (UDFR)* <https://bitbucket.org/udfr/main/wiki/Home>.

[9] "Challenges and opportunities" (2011), *Science* 331:6018 (11 February 2011): 692-693 <http://www.sciencemag.org/content/331/6018/692.full.pdf>.

[10] UCLA Institute for Pure & Applied Mathematics, *Networks and Network Analysis for the Humanities: An NEH Institute for Advanced Topics in Digital Humanities*, August 15-27, 2010 <https://www.ipam.ucla.edu/programs/hum2010/>.

[11] Andrew Sallans (2011), "DMPTool: Supporting the data lifecycle," *NSF Workshop on Research Data Lifecycle Management*, Princeton University, July 18-20, 2011 <http://www.columbia.edu/~rb2568/rdlm/Sallans_UV_RDLM2011.pdf>.

[12] Serge J. Goldstein and Mark Ratliff (2010), *DataSpace:: A Funding and Operational Model for Long-Term Preservation and Sharing of Resource Data* <http://dspace.princeton.edu/jspui/handle/88435/dsp01w6634361k>.

[13] University College London/British Library (2011), LIFE: Life Cycle Information for E-Literature <http://www.life.ac.uk/>.

[14] Stephen Abrams, Patricia Cruse, John Kunze, David Minor, and Mike Smorul (2011), "'Neighborhood watch' for repository quality assurance," *Designing Storage Architectures for Preservation Collections*, Library of Congress, September 26-27, 2011 <http://www.digitalpreservation.gov/news/events/other_meetings/storage11/docs/cdl_neighborhood_watch_paper.pdf>.

[15] John Kunze, Rachel Hu, Patricia Cruse, Catherine Mitchell, Stephen Abrams, Kirk Hastings, and Lisa Schiff (2011), "Baby steps to data publication," *Beyond the PDF*, University fo California, San Diego, January 19-21, 2011

<http://sites.google.com/site/beyondthepdf/workshop-papers/baby-steps-to-data-publication>.

[16] John Kunze, Rachel Hu, Patricia Cruse, Catherine Mitchell, Stephen Abrams, Kirk Hastings, and Lisa Schiff (2010), *Practices, Trends, and Recommendations in Technical Appendix Usage for Selected Data-Intensive Disciplines*, Report for the Gordon and Betty Moore Foundation, <http://escholarship.org/uc/item/9jw4964t>.

[17] Ecoinformatics (2011), *Ecoinformatics Online Resource for Managing Ecological Data and Information* <http://www.ecoinformatics.org/index.html>.

[18] DataCite (2011), *DataCite: Helping You Find, Access and Re-use Research Data* <http://datacite.org/>.