CAL POLY LIBRARY SERVICES
ROBERT E. KENNEDY LIBRARY

TO: Office of Science and Technology Policy

FROM: Anna Gold, University Librarian, California Polytechnic State University Library Services, San Luis Obispo, California

DATE: January 12, 2012 (revised response)

SUBJECT: Response to RFI: Public access to **digital data** resulting from federally funded scientific research.

About this response

The following information was prepared by the Kennedy Library at California Polytechnic State University in San Luis Obispo, California, in response to the request for information issued November 3, 2011, by the Office of Science and Technology Policy.

Information was primarily provided by Marisa Ramirez, Digital Repository Librarian at Kennedy Library, in consultation with Timothy Strawn, Director of Information Resources and Archives, and University Librarian Anna Gold, with additional input provided by David Beales, Associate Librarian for Engineering at Kennedy Library.

The Robert E. Kennedy Library and Cal Poly Library Services at California Polytechnic State University (San Luis Obispo) provide a comprehensive program of library services within a teaching-led comprehensive public polytechnic university, including access to scholarly and professional information, data services, and a digital institutional repository that recently passed the 1.3 million download mark. The Library's Data Services Team is actively working to define, develop, and sustain a library program of data and GIS services that contributes to Cal Poly's mission and programs of learning and research.

The mission of Cal Poly Library Services is to promote open and informed inquiry, foster collaboration and innovation, support the unique needs of every student and scholar at Cal Poly, and contribute to the cultural life of our community. In common with other libraries in higher education, the Library is in a unique position in our organization to do this through access to technologies that support the creation, reuse, sharing and preservation of new knowledge.

Background

The U.S. government funds tens of billions of dollars in basic and applied research each year, with the goals of speeding the pace of scientific discovery, fueling innovation, and improving the public good. At its core, the most significant research findings are supported with underlying data sets, often collected in digital form on a large or small scale.

Traditionally, libraries have served as steward of the record, and this continues to be true in the digital realm. In the last five years, there have been an increasing number of hires in the Library and Information Science (LIS) fields for data services librarians, data curators, digital curators, and digital repository services librarians. In this era of data-driven science, academic and research libraries are strategically repositioning

themselves to align more closely with their institutions' research dissemination strategies. Academic and research libraries are considering digital curation issues broadly, across all subject disciplines, in an effort to share information as a community and work together to determine best practices and standards.

A host of parties have interests in the curation of digital research data, and academic libraries are well-positioned to represent their interests in access, outreach and advocacy, information management, digitization and digital technologies, support for teaching and learning, and preservation. In collaboration with other campus partners, libraries also have a role to play in developing strategies to capture, collect, manage and preserve data streams created by faculty. Sharing and coordination of these efforts can contribute to the overall data management and preservation efforts.

The following further comments on the questions posed in the RFI are organized into two interrelated themes: preservation, discoverability and access; and standards for interoperability, reuse, and repurposing.

1. Preservation, Discoverability, and Access

Cal Poly Library Services recommends four major approaches to these interconnected challenges:

1.1 Develop a national infrastructure based on existing models.

Below are several models that we recommend be explored:

- **Australia National Data Service (ANDS)** has created the Australian Research Data Commons to support initiatives designed to enhance existing data creation and capture infrastructure commonly used by Australian researchers and research institutions. This will ensure that the data creation and data capture phases of research are fully integrated to enable effective ingestion into a local data and metadata store published through Research Data Australia. This integration enables researchers to contribute descriptions of data to the Australian Research Data Commons directly from the lab, instrument or fieldwork site. It also ensures that higher quality metadata, critical for reuse and discovery, is produced through automated and semi-automated systems.

- **Joint Information Systems Committee (JISC)** is a United Kingdom funding body that supports research by providing leadership in the innovative, shared use of information and communications technologies and infrastructure to support education, research and institutional effectiveness. JISC offers support at local, national and international level by creating and supporting shared resources, knowledge, expertise and services, particularly where it gives rise to immediate cost savings.

- The **Digital Curation Centre (DCC)** is the United Kingdom's leading hub of expertise in curating digital research data. Launched in 2004 by JISC, the DCC provides a national centre for solving challenges in digital curation that could not be tackled by any single institution or discipline. The DCC is responsible for developing resources, training opportunities, and funding projects that promote the development of innovative methods for the preservation, discoverability, and access of research data.

- **DRIVER i**s a pan-European effort whose primary objective is to create a cohesive, robust and flexible infrastructure for digital repositories, offering sophisticated services and functionalities for researchers, administrators and the general public. Aimed to be complimentary to GEANT2, the infrastructure for computing resources, data storage and data transport, DRIVER delivers resources that result from

scientific output, including scientific/technical reports, working papers, pre-prints, articles and original research data. The vision is to establish the successful interoperation of both data network and knowledge repositories as integral parts of the E-infrastructure for research and education in Europe.

1.2 Utilize and encourage integration of current sources of expertise in the library, archives and records management fields.

The library profession has many professional organizations devoted to exploring, adapting and implementing emerging digital curation services, technologies, and infrastructures, whether repository-based or platform-agnostic, born-digital or digitized, for the lifecycle management of research, scholarship and other academic activities.

We recommend that the federal government leverage these professional organizations to assist in developing methods to curate a variety of content in digital form; to use scalable, efficient, and sustainable methods to inform and educate librarians on digital curation trends and new technologies; and finally, to collaborate with other organizations within the library profession and academe on issues concerning digital curation. Such library organizations include, but are not limited to:

- Association of College and Research Libraries (ACRL), specifically the Digital Curation Interest Group and SPARC;
- American Society for Information, Science and Technology (ASIS&T), specifically the Digital Libraries Interest Group and the Research Data Access and Preservation Group;
- Association for Library Collections and Technical Services (ALCTS), specifically the Preservation & Reformatting Sections including the Intellectual Access to Metadata Interest Group, Digital Conversion Interest Group, Digital Preservation Interest Group;
- Association for Information and Image Management (AIIM), specifically the Electronic Records Management section; and
- Society for American Archivists (SAA), specifically the Electronic Records Section.

1.3 Cultivate a workforce capable of addressing the new challenges posed by data curation and cyberinfrastructure development.

Expanding current data curation and cyberinfrastructure activities and embarking on new ones will require investment in professional development for library staff, and in some cases the creation of entirely new positions. Funding should go towards identifying new facets of library graduate education and subsequent professional development to prepare librarians to support data curation and cyberinfrastructure activities.

A challenge in identifying suitable models is to provide sustained, practical professional development opportunities suitable for working professionals, and not limited to campus-based residential programs, though these also have an important role to play in fostering the knowledge, experience, and skills to contribute to data curation and cyberinfrastructure activities.

1.4 Integrate and universally adopt existing mechanisms to educate faculty regarding copyright and intellectual property, and improve compliance with federal data stewardship.

We suggest exploration and possible adoption of the following models and approaches:

- **United Kingdom Intellectual Property Office Audit Model, and the Hargreaves Review.** In the United Kingdom, a key process in managing the intellectual property stemming from research is to conduct an Intellectual Property audit. The purpose of conducting an audit is not simply a stocktaking exercise, but instead it is undertaken to further exploit the intellectual property assets in hand, to implement procedures to minimize the risk of litigation by infringing others' copyrighted material as well as an opportunity for researchers to ask questions they may have about Intellectual Property laws and for the University to identify areas of further education.

  In November 2010, the UK Prime Minister commissioned an independent review by Ian Hargreaves and a team of consultants of the UK's intellectual property framework. The review made ten recommendations designed to ensure that the UK IP system promotes innovation and growth in the 21st century, both nationally and internationally. The United States may consider adopting elements from the UK Audit Model as well as conduct a study to determine how to realize efficiencies within the existing IP system.

- **Develop tools to assert author rights to research data.** High-impact academic publishers such as Nature Publishing Group are now requiring authors to deposit their raw research datasets with them as part of the peer-review process. This raises concerns about intellectual access to the raw data: journals have traditionally required authors to sign over intellectual property rights in exchange for getting published, and may well extend these terms to the raw data. Criteria or tools must be developed to help authors assert their intellectual property rights to their research. Failure to do so could result in stifling academic creativity and intellectual progress.

- **Develop criteria to guide academic publishers' policies on embargo periods.** The purpose of an embargo is to protect the revenue interests of the publisher, but it is generally considered frustrating to academic researchers who rely on current publications to further their work. In essence, the publisher's embargo stifles academic creativity and intellectual progress. As publishers increasingly require raw datasets from authors in order to publish scholarly articles, a further concern is that these embargoes may be extended to limit access to research datasets. A more equitable balance must be reached, based on established and publicly available criteria, to better guide academic publishers' policies on embargo periods. Furthermore, timely deposit of research data in open disciplinary repositories as may be frustrated by a publisher embargo. It is more desirable that data be openly accessible without embargo, with deposit in a disciplinary or institutional repository to be preferred over deposit with publishers. This would not prevent publishers from requesting that authors provide data as part of the peer review process, nor prevent publishers from linking to that data for published articles, using community-based citation standards.

- **Investigate ways to integrate attribution initiatives into reporting systems to ensure compliance with Federal data stewardship policies. Ensure that appropriate attribution is provided to the creators of data by promoting methods such as:**

- **Open Researcher and Contributor ID (ORCID)**. ORCID aims to solve the author/contributor name ambiguity problem in scholarly communications by creating a central registry of unique identifiers for individual researchers and an open and transparent linking mechanism between ORCID and other current author ID schemes. These identifiers, and the relationships among them, can be linked to the researcher's output to enhance the scientific discovery process and to improve the efficiency of research funding and collaboration within the research community.

- **ResearcherID**, a multi-disciplinary scholarly research community developed by Thompson Reuters' Web of Knowledge, which assigns a unique identifier to each author to eliminate author misidentification while simultaneously adding dynamic citation metrics and collaboration networks to an author's profile.

- **Use existing academic channels to educate, verify and improve compliance with Federal data stewardship and access policies for scientific research.**

  Such channels include:

  - **Institutional Review Boards (IRB) and Institutional Animal Care and Use Committees (IACUC).** These are ethical committees formally designated to approve, monitor and review research involving humans and animals. Federal regulations have empowered these committees to approve, require modifications in planned research prior to approval, and to perform critical oversight functions for research conducted on human and animal subjects. Most, if not all, research institutions have one or both of these committees. These committees often require researchers to complete educational modules, such as the CITI Program, which is a service providing research ethics education to all members of the research community. A similar online program could be developed for data stewardship and could be required by local IRB and IACUC committees as a condition of local research approval.

  - **Campus departments such as Research and Grants Development Offices.** These campus offices can verify that grant applicants are including data management costs in grants. Grant applicants are typically required to report back to a central campus office on disbursement of funds and requirement compliance. Utilize this existing framework by leveraging their current activities including verifying grant compliance.

2. Standards for Interoperability, Reuse and Repurposing

It is suggested that metadata standards generally are most usefully considered within the limits of their user communities' standard practices. So long as they are XML-based, there is a useful degree of interoperability; the trend towards interoperable schema will continue while it is still useful. However, librarians are aware that any effort to maintain quality metadata standards is difficult. Metadata schema for data that are not directly related to the needs of the disciplinary community of interest are unlikely to be embraced wholeheartedly.

2.1 Recognizing the important role of disciplinary communities in developing their own descriptive and administrative metadata standards, and recognizing the powerful potential of XML-based and semantic web

approaches to assuring interoperability, it will also be useful to continue to exploit existing national and international structures to develop and promote common standards. These structures include:

- **ISO (International Organization for Standardization)** is a network of the national standards institutes of 162 countries and is the world's largest developer and publisher of International Standards. Because ISO wields influence in both the public and private sectors, this organization has the ability to form consensus on solutions that meet both the requirements of government, education and the broader needs of society. ISO, for example, may be a relevant standard for the registry, management, sharing, and delivery of research data across digital repositories and discovery services.

- **Cross-disciplinary metadata initiatives, such as the Dublin Core Metadata Initiative (DCMI), and web content interoperability standards, such as the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and the Open Archives Initiative Object Reuse and Exchange (OAI-ORE).** The DCMI provides core metadata vocabularies that are widely used in digital repositories to support management, discovery and interoperability of resources, such as Darwin Core. Darwin Core is a stable and versatile standard that facilitates the discovery, retrieval, and integration of information about modern biological specimens and their supporting evidence housed in digital or physical collections. The Open Archives Initiative has its roots in the open access and institutional repository movements, and has developed widely used interoperability standards (such as OAI-PMH and OAI-ORE) that aim to facilitate the efficient dissemination, description and exchange of content.

2.2 *Develop permanently funded tools that enable wide scale registering and verification of data repositories.*

Two examples of these include:

- **DataBib**, a grant-funded project by the Institute of Museum and Library Services, aims to create a community-driven, annotated bibliography of research data repositories. Once funding ends, however, it is unclear how this resource will be maintained. Ideally, this would be a project that could be developed externally, but then adopted and permanently manage by public sector stakeholders.

- **OpenDOAR**, a directory of open access academic repositories, is a current example of how best to develop a single comprehensive, authoritative list which requires registration, verification and harvesting of data repository metadata.