



Document Downloaded: Tuesday May 21, 2013

**COGR Response to Public Access to Digital Data**

Author: Carol Blum

Published Date: 01/12/2012

# COGR

an organization of research universities

## COUNCIL ON GOVERNMENTAL RELATIONS

1200 New York Avenue, N.W., Suite 750, Washington, D.C. 20005  
(202) 289-6655/(202) 289-6698 (FAX)

### BOARD OF DIRECTORS

#### CHAIR

DAVID WYNES  
Emory University

JAMES BARBRET  
Wayne State University

ELAINE BROCK  
University of Michigan

SUSAN CAMBER  
University of Washington

PAMELA CAUDILL  
University of Pennsylvania

MICHELLE CHRISTY  
Massachusetts Institute of Technology

KELVIN DROEGEMEIER  
University of Oklahoma

CHARLES LOUIS  
University of California, Riverside

MICHAEL LUDWIG  
Purdue University

JAMES LUTHER  
Duke University

JAMES R. MAPLES  
University of Tennessee

DENISE MC CARTNEY  
Washington University in St. Louis

ALEXANDRA MCKEOWN  
Johns Hopkins University

KIM MORELAND  
University of Wisconsin

CORDELL OVERBY  
University of Delaware

SUSAN SEDWICK  
University of Texas, Austin

JOHN SHIPLEY  
University of Miami

JAMES TRACY  
University of Kentucky

ERIC VERMILLION  
University of California, San Francisco

DAVID WINWOOD  
University of Alabama, Birmingham

MARIANNE WOODS  
University of Texas,  
San Antonio

ANTHONY DE CRAPPEO  
President

January 12, 2012

Interagency Working Group on Digital Data  
National Science and Technology Council  
Office of Science and Technology Policy

SUBJECT: Request for Information: Public Access to Digital Data  
Resulting from Federally Funded Research

The Council on Governmental Relations (COGR) is an association of 188 research universities and their affiliated academic medical centers and research institutes. COGR concerns itself with the influence of federal regulations, policies and practices on the performance of research conducted by its member institutions. Our goal is to ensure that federal policy goals can be met in an effective and efficient manner without creating administrative structures that may hinder compliance.

COGR offered written and testimonial comment to the National Institutes of Health (NIH) as it developed its policy to enable public access to NIH-funded research. As we noted in our January 2010 response to the Office of Science and Technology Policy's (OSTP) request for information concerning Federal government-wide public access policies, we support the goal of providing timely and less costly access to data that result from federally funded research and observed that the public does have such access through various depositories.

This new request for information raises a series of specific questions but we want to begin by introducing a question that may burden any efforts on the part of OSTP to coordinate Federal agencies' policies concerning the long-term stewardship of research data. The meaning of "data" as used in Federal regulations and policies has become increasingly ambiguous. There does not exist a common definition among Federal agencies, and the definitions used by agencies do not always reflect the meaning applied within the research community, which itself does not have a uniform definition. The rights and responsibilities surrounding ownership, access to and retention of data will be affected by the variety of meanings ascribed to "data." Frequently, the term "research data" is confused with what are, by definition, research materials.

Generally, we would argue that research data consists of information that provides a quantitative and/or qualitative description or characterization. This definition is consistent with the Office of Management and Budget's (OMB) Circular A-110, *Uniform Administrative*

*Requirements for Grants and Agreements With Institutions of Higher Education, Hospitals, and Other Non-Profit Organizations*, definition “as the recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues.” And yet, various agencies provide a broader definition.

For example, NIH defines “data” as “recorded information, regardless of the form or medium on which it may be recorded, and includes writings, films, sound recordings, pictorial reproductions, drawings, designs, or other graphic representations, procedural manuals, forms, diagrams, work flow charts, equipment descriptions, data files, data processing or computer programs (software), statistical records, and other research data.” Some would argue this definition embraces “research materials.”

For much of the research community research materials are those materials from which data can be extracted. Materials are tangible or physical objects, e.g., writings like a database, cells, molecules, designs, plans, forms, flow charts, planets, plants, and/or animals. Thus, in making the distinction between research *data* and research *materials*, it’s important to distinguish between the entities *containing* the data and the data *themselves*. For example, a lab notebook, a recording, or an insect are not data but contain data or represent entities about which data (description or characterization) can be created.

All Federal agency policies and regulations do not employ a similar distinction between data and materials. As noted above, NIH’s definition of research data includes materials such as data files, which are recorded but in most cases will not provide a quantitative or qualitative description or characterization in and of itself. The Federal Acquisition Regulations (FARs) that provide general terms and conditions for Federal contracts includes computer software and software documentation in its definition of “data;” the Defense contract regulations (DFARS) reference “technical data” which includes computer software documentation but not the software programs or source codes. The Environmental Protection Agency (EPA) goes further by carving out “raw data” to include “laboratory worksheets, memoranda, notes or exact copies thereof, that are the result(s) of original observations and activities of a study and are necessary for the reconstruction and evaluation of the report of that study” (40 CFR Part 742).

For the research community under OMB’s Circular A-110, the OMB definition which applies across Federal agencies may provide the most useful general framework for discussing the access to and retention of research data. In this definition, preliminary or “raw” data or research materials without analysis should not be included for the purposes of access by the general public.

However, if the OSTP truly intends to harness digital data for public access, the challenges are even greater. If the intent of the OSTP’s proposal is to provide greater public access to “digital data” then the consequences for research institutions is a significant financial and administrative burden. In its 2009 report, “Harnessing the Power of Digital Data” the Working Group on Digital Data broadly defined digital data to encompass:

. . . born digital and digitized data produced by, in the custody of, or controlled by federal agencies, or as a result of research funded by those agencies . . . [including] the full range of data types and formats relevant to all aspects of science and engineering research and education in local, regional, national, and global contexts with the corresponding breadth of potential scientific applications and uses.

As with the distinction between “data” and “materials,” it is important to distinguish between the different types of “digital data” generated by different disciplines into those that will be of use to other scientists, e.g. genomic data, long-term climate data, versus digital data, or perhaps more accurately “digital materials” that really is of no use except to the originating scientist or investigator because it requires access to notebooks and other information to analyze, laboratory-level digital data. In some disciplines given the nature of the equipment and processes used in experimentation, “digital data” encompasses essentially all the original data because virtually all primary data is collected in a digital format. The latter can be massive and is retained by the investigator as support for publications, etc., versus the former such as genomic or climate data sets that will be needed for many years by a variety of scientists and investigators.

If required by Federal agencies to retain laboratory-level digital data or materials, institutions would need to develop extensive management systems that could store the huge diversity of digital data that our researchers develop, remembering that most laboratory-level “digital data” would include all the original data we collect every day. The growing number of cooperative data repositories for digital data with a larger, more global interest to other scientists helps institutions meet their responsibilities for data access and data sharing with the scientific public that can make good use of the data. Yet, these repositories come with challenges as well including maintaining the integrity and security of the data housed in the repositories.

The RFI addresses “digital data” albeit reading through questions many of them use the more generic term “data” or “research data.” It is not clear whether the OSTP is making a distinction between digital data and other types of data or not. Thus, we return to our original concern over the meaning of “data” as used in Federal regulations and policies and, as noted, in this RFI. The rights and responsibilities and, perhaps more challenging, the costs and burdens associated with providing public access surrounding ownership, access to and retention of data will be affected by the variety of meanings ascribed to “data.”

With these overriding concerns, we offer the following responses to some of the questions.

### **Preservation, Discoverability, and Access**

**Comment 1:** Growing markets related to access and analysis and using those markets to grow the economy and improve productivity of the scientific enterprise.

Data resulting from research are the foundation for the continuing dialogue among scientists that advances our scientific enterprise and productivity. Research data, whatever the format, serve as the source for inventions, publications, and can with sufficient support, serve as the origin for expending existing and creating new businesses.

The individual most capable and most interested in using the data is the person who created it. It would be a mistake to jeopardize the ability of that individual scientist or investigator to exploit the potential of the data.

Data derived from Federally supported basic and applied research are data that can be intriguing to potential investors. But without further investigations and directed proofs of the concepts suggested by the basic research, the ability of the general public to use the data is limited. Like access to research publications, access to research data will not provide a direct, uninterrupted link to a new business or activity without significant investment.

**Comment 2: Protection of Intellectual Property**

Unlike patentable inventions, where the Bayh-Dole Act and implementing regulations provide a uniform Federal framework, there is no single source of authority on ownership and protection of data resulting from Federally funded research. Generally the research agencies that provide Federal financial assistance allow recipients to copyright and own data developed under the award, subject to the right of the agency to use the work for Federal purposes and, as appropriate, subject to specific requirements like the NIH and National Science Foundation's data sharing policies. Under Federal contracts, the government may allow copyright, but normally the government retains the ability to exercise all the rights of the owner, e.g., distributing copies to the public. The FAR provides that universities and colleges may claim copyright in data developed under a contract for basic and applied research that they perform solely. This provision may be a serious constraint in the current funding climate that otherwise encourages teaming arrangements and public-private partnerships. The DFARS makes no special provision for educational institutions.

Institutions have been aggressive in identifying research activities with the potential to stimulate economic development and work to patent and license the intellectual property to the benefit of the business partners and, ultimately, the public. The ability to protect the intellectual property is what attracts businesses to make the investment in time and resources to license the technology and bring products to the market. Public access may not serve these purposes, nor may the approach followed in the various Federal acquisition regulations.

OSTP notes the examples of NIH and the National Science Foundation policies concerning data management and data sharing. These policies work because the focus of access and sharing is the scientific community. The repositories established by various agencies, notably NIH, are built on carefully crafted policies and procedures that protect the intellectual property rights associated with the data. Any agency contemplating policies or regulations must ensure similar protections.

**Comment 3: Differences in Disciplines and Data**

It is precisely these differences in disciplines and data that pose a significant challenge to data management and preservation and, we would add, the establishment of standards for interoperability, reuse and repurposing. Rather than building separate, prescriptive policies similar but different across agencies and disciplines, OSTP should advocate a simple policy requiring federal grantees and contractors to enable the transfer of research data in a manner that meets the specific disciplines and data. In this way, those who understand the data can establish a framework for its transfer to others with appropriate contractual protections including privacy, confidentiality, etc. Institutions can transfer information on how the data were collected, what unique tools are needed for analysis, etc. This type of approach avoids the inevitable pitfalls in attempting to create a single set of standards applicable across disciplines, data and agency.

**Comment 6: Real Costs of Preservation and Access**

For grantee institution, the costs of preserving and sharing are significant. As a general rule, most institutions rely on the investigator to help meet its obligations to preserve and share research data. The creation of national repositories in some disciplines has helped in the

management of access to research data. Charging reasonable fees for the duplication of research data can not cover the full and significant costs associated with long-term preservation.

## **Standards for Interoperability, Reuse and Repurposing**

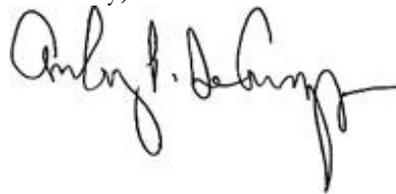
### **Comment 10: Digital Standards**

As noted above, we would caution OSTP and the Federal agencies to avoid establishing standards for data preservation and transfer that penalizes the investigator in terms of the loss of first use and requires a significant investment in re-formatting of data to meet a national standard. Establishing a general principle that requires transfer or sharing of data within a reasonable time frame will ensure that the data is available to other interested scientists and investigators while preserving appropriate protections.

Thank you for this opportunity to comment on OSTP's continuing consideration of the value of public access to peer-reviewed publications. We would note that the America COMPETES Reauthorization Act of 2010 (PL 111-358) Sec. 103 does not assume that a single, government-wide policy is appropriate and charges the Interagency Public Access Committee with coordinating agency activities concerning access to publications and data.

COGR has long supported harmonization and coordination among the Federal agencies in order to streamline the compliance with Federal mandates and regulations. In the case of access to data, we would suggest that setting a principle of long-term preservation and access without prescribing the institutional approach is the wisest course. As OSTP observes, disciplines have begun defining standards that meet the needs of the research community. Relying on those standards and efforts of institutions to transfer new, innovative technologies to the marketplace has led to extraordinary economic developments to date.

Sincerely,

A handwritten signature in black ink, appearing to read "Anthony P. DeCrappeo". The signature is fluid and cursive, with a long horizontal stroke extending to the right.

Anthony P. DeCrappeo  
President