Data Management

Terrell Vanderah
NIST, Ceramics Division
Editor-in-Chief, Phase Equilibria Diagrams
(NIST Standard Reference Database 31)
Gaithersburg, MD

terrell.vanderah@nist.gov
Tel (301)975-5785

Drivers

The most recent legislation (Jan 2010):

H.R. 5116 [111th]: America COMPETES Reauthorization Act of 2010

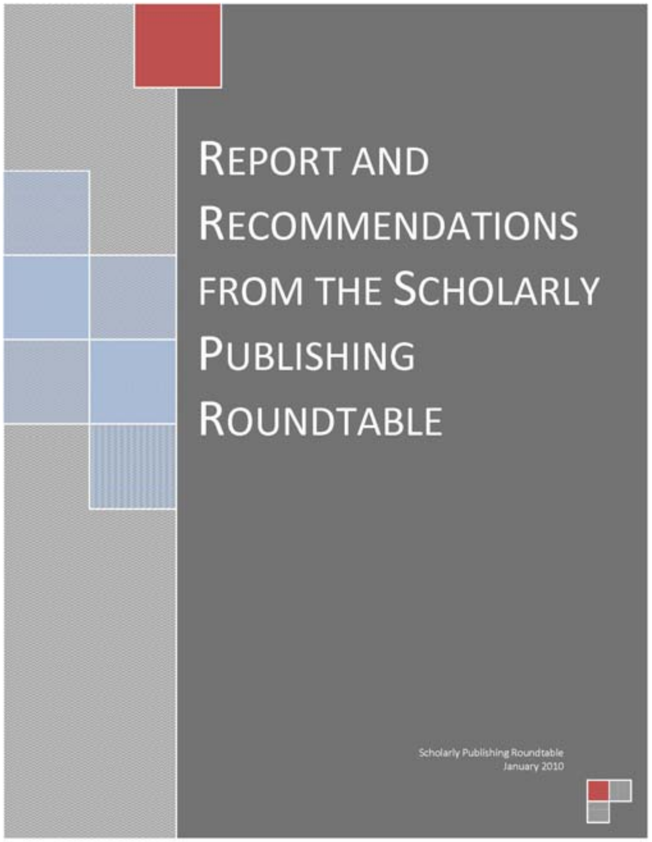
http://www.govtrack.us/congress/billtext.xpd?bill=h111-5116

Section 103 Establishment - The Director (OSTP) shall establish a working group under the National Science and Technology Council with the responsibility to coordinate Federal science agency research and policies related to the dissemination and long-term stewardship of the results of unclassified research, including digital data and peer-reviewed scholarly publications, supported wholly, or in part, by funding from the Federal science agencies.

Section 104 Management of Scientific Collections - The Office of Science and Technology Policy shall develop policies for the management and use of Federal scientific collections to improve the quality, organization, access, including online access, and long-term preservation of such collections for the benefit of the scientific enterprise.

IT IS THE LAW.

The legislation is based on part of the recommendations of a report commissioned by the House. http://www.aau.edu/WorkArea/DownloadAsset.aspx?id=10044



REPORT AND RECOMMENDATIONS FROM THE SCHOLARLY PUBLISHING ROUNDTABLE

Scholarly Publishing Roundtable
January 2010

January, 2010

SHARED PRINCIPLES

1) Peer review must continue its critical role in maintaining high quality and editorial integrity.

2) Adaptable business models will be necessary to sustain the enterprise in an evolving landscape.

3) Scholarly and scientific publications can and should be more broadly accessible with improved functionality to a wider public and the research community.

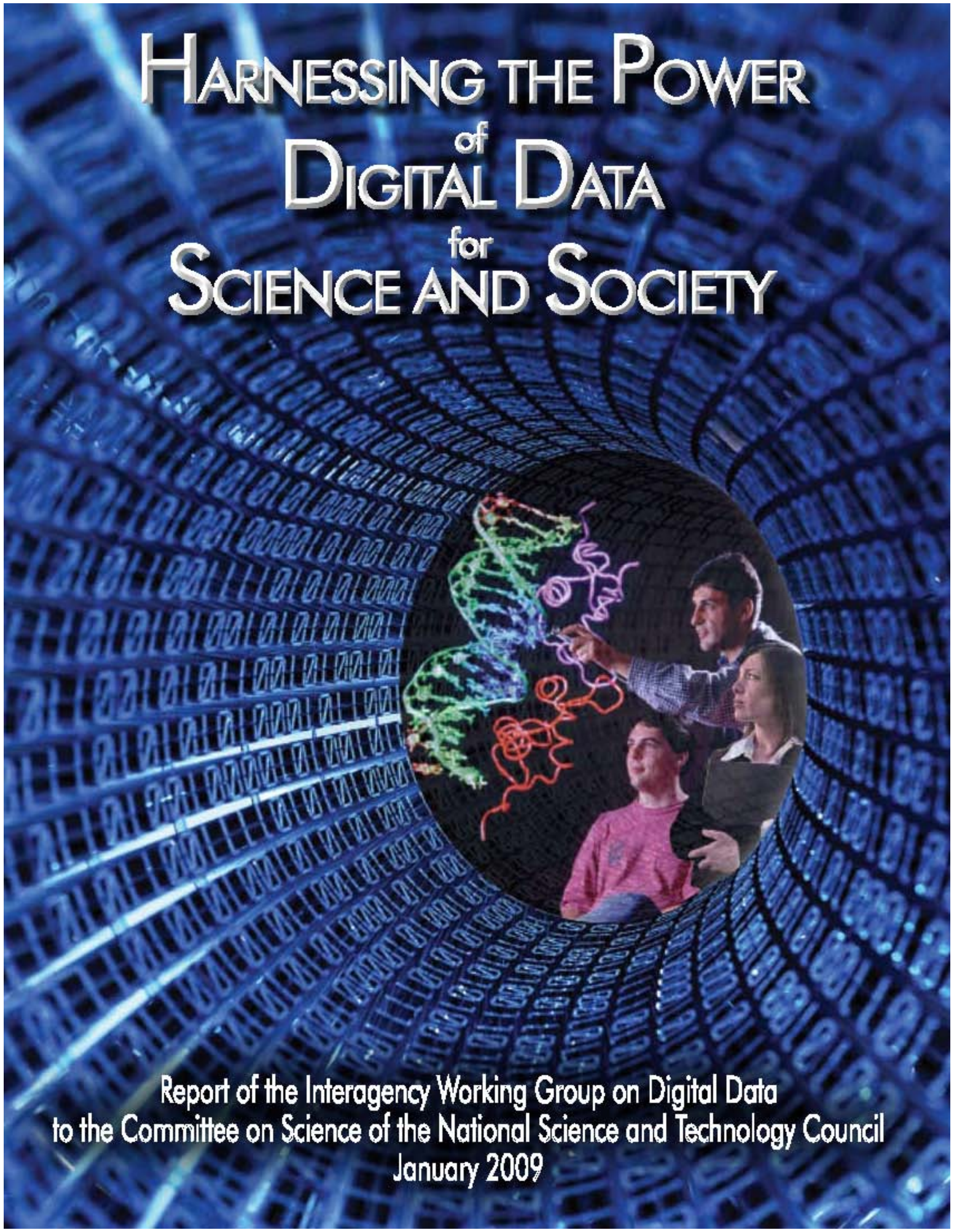
4) Sustained archiving and preservation are essential complements to reliable publishing methods.

5) The results of research need to be published and maintained in ways that maximize the possibilities for creative reuse and interoperation among sites that host them.

RECOMMENDATIONS

The Roundtable's core recommendation is:

Each federal research funding agency should expeditiously but carefully develop and implement an explicit public access policy that brings about free public access to the results of the research that it funds as soon as possible after those results have been published in a peer-reviewed journal.

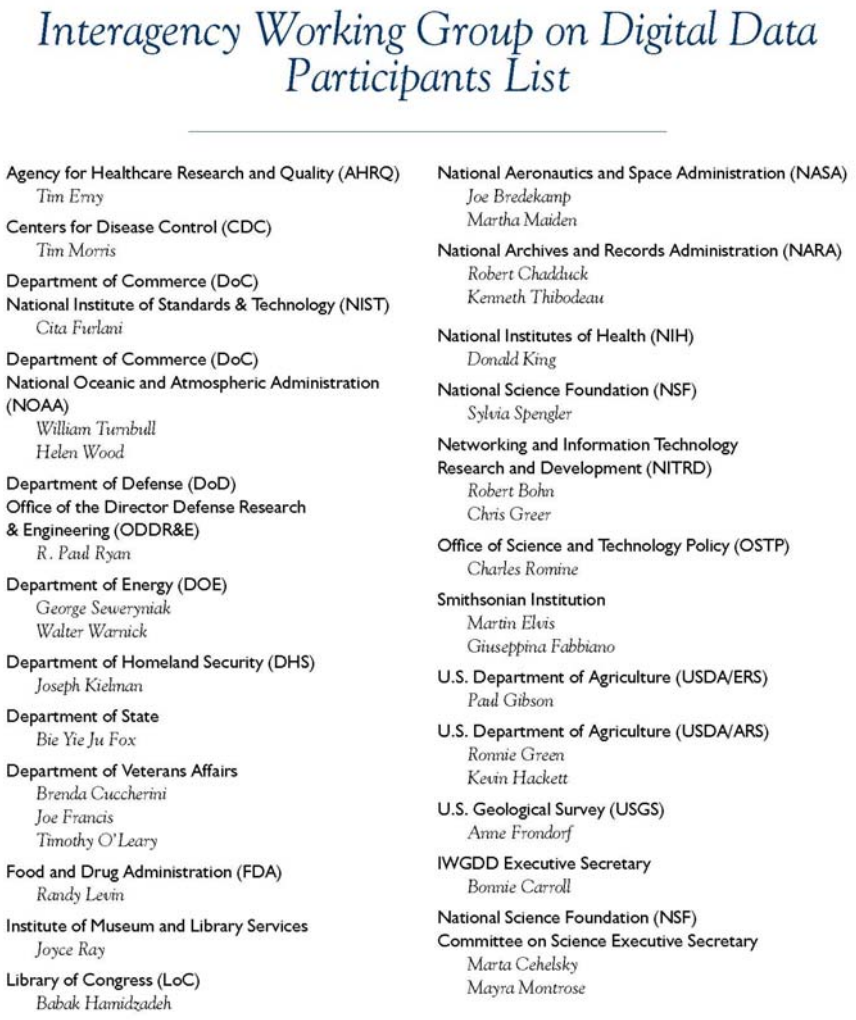


Harnessing the Power of Digital Data for Science and Society

Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council

January 2009

http://www.nitrd.gov/about/harnessing_power_web.pdf



Interagency Working Group on Digital Data Participants List

Agency for Healthcare Research and Quality (AHRQ)
Tim Erny

Centers for Disease Control (CDC)
Tim Morris

Department of Commerce (DoC)
National Institute of Standards & Technology (NIST)
Cita Furlani

Department of Commerce (DoC)
National Oceanic and Atmospheric Administration (NOAA)
William Turnbull
Helen Wood

Department of Defense (DoD)
Office of the Director Defense Research & Engineering (ODDR&E)
R. Paul Ryan

Department of Energy (DOE)
George Seweryniak
Walter Warnick

Department of Homeland Security (DHS)
Joseph Kielman

Department of State
Bie Yie Ju Fox

Department of Veterans Affairs
Brenda Cuccherini
Joe Francis
Timothy O'Leary

Food and Drug Administration (FDA)
Randy Levin

Institute of Museum and Library Services
Joyce Ray

Library of Congress (LoC)
Babak Hamidzadeh

National Aeronautics and Space Administration (NASA)
Joe Bredekamp
Martha Maiden

National Archives and Records Administration (NARA)
Robert Chadduck
Kenneth Thibodeau

National Institutes of Health (NIH)
Donald King

National Science Foundation (NSF)
Sylvia Spengler

Networking and Information Technology Research and Development (NITRD)
Robert Bohn
Chris Greer

Office of Science and Technology Policy (OSTP)
Charles Romine

Smithsonian Institution
Martin Elvis
Giuseppina Fabbiano

U.S. Department of Agriculture (USDA/ERS)
Paul Gibson

U.S. Department of Agriculture (USDA/ARS)
Ronnie Green
Kevin Hackett

U.S. Geological Survey (USGS)
Anne Frondorf

IWGDD Executive Secretary
Bonnie Carroll

National Science Foundation (NSF)
Committee on Science Executive Secretary
Marta Cehelsky
Mayra Montrose

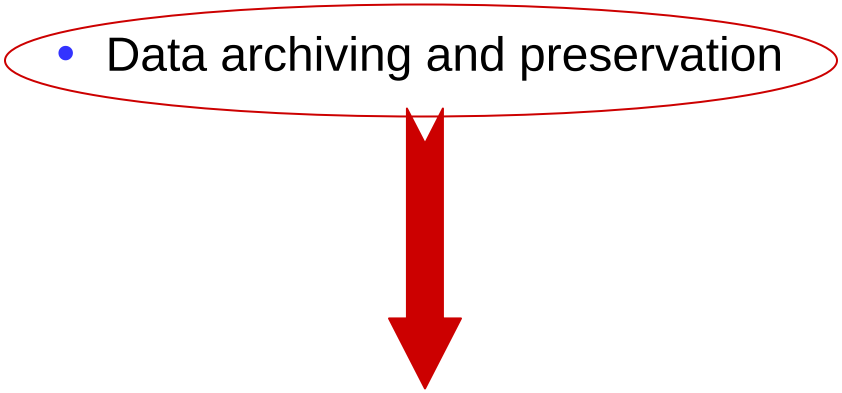
EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

Science
11 February 2011 | $10
scientific
information
climate
science
research
sharing
model
need
brain
challenges
analysis
new
researchers
visualization
health
many
access
human
example
knowledge
data
AAAS

11 Feb 2011

Dual and related drivers:

- “Open-science movement” – any research funded by the federal government should be available to the taxpayers (e.g. NIH)



- Data archiving and preservation

NSF Requirement for Data Management Plans

Data is an essential component of the process of science.

- Observations (**data collection**) followed by rational thinking

Data is now digital (not just paper in a file cabinet) and therefore can be easily retained and re-used (e.g. journal data deposition).

We must stop throwing data away.

Data Management = preservation and documentation

Data Life-cycle: *Create*
Use
Store
Share

Data Repositories: *The million-dollar question*

- Check for existing *(NIST may be able to advise)*
- Check with your institution – universities have started to create repositories for data deposition and archiving

Another huge challenge: "Data should be stored with sufficient information (metadata) that 'anyone' can use it"

Data Management Plans

NSF has offered good guidance

(http://www.nsf.gov/bfa/dias/policy/dmp.jsp see FAQs section

NIST ideas: (not formal policy, but well-thought out by a deliberative committee)

All digital scientific data must be covered by a data management plan.

Data management plans may differ, but each plan must have **critical elements** to ensure that the data deliver the greatest utility to the public.

For example:

- A plan that covers the end-to-end life cycle of the data.
- A plan must indicate how data will be retained and when it will be released.
- If data will not be released, the plan must indicate why such action(s) will not be taken.
- A plan should indicate what metadata is associated with each data document.
- A plan that includes procedures for providing archival storage.

Definitions

Data Disposition – A plan for the future of digital scientific data, i.e., a determination of how long data should be retained.

Data Management Services – a subset of Data Management that includes adherence to agreed-upon standards; ingesting data, developing collections, and creating products; maintaining databases; ensuring permanent, secure archives; providing both user-friendly and machine-interoperable access; assisting users; migrating services to emerging technologies; and responding to user feedback.

Digital Preservation – the active management of digital information over time to ensure its accessibility.

Digital Scientific Data – digitally recorded and derived observations and measurements of physical artifacts, properties, phenomena, as well as models, calculations, analyses, test results, and research compilations existing entirely on computers. Software, algorithms, instrument data, related documentation, and *metadata* are also included. Multiple media, including image, video, and audio recordings, along with visual renderings may be included.

Lifecycle Management - a comprehensive approach to managing digital scientific data through all the stages of its "life." It begins with planning for the creation or acquisition of data, continues through the refinement and use, and ends only when the data are transferred to another entity or destroyed. Lifecycle management functions are sequential but may go through certain stages of the lifecycle multiple times as the data are used by different groups or for different purposes.

Management Plan - a written plan that defines the lifecycle management processes with a clearly defined point of contact for the data, impact of the data, and description of the data.

Metadata - information that gives context to quantitative data. This information is critical to future interpretation of the data, and for relating the quantitative data to other data. Metadata give value to the data, and help to convey the quality of the data. Metadata can be a system description for a given system/device (hardware architecture, operating system, etc.), a content description (instrument data, text, image, audio, video, etc.), a procedure or protocol about a given instruction/instrument, a result from a computational process, etc.

Data management plans should:

1. **Identify the types of data associated with a project or activity**

   Data may be measurements, images, audio data, videos, computational results, professional correspondence, reports, publications, or presentations.

   Describe the format of each type of data and any software responsible for its creation and manipulation

2. **Identify the principal consumers of the data**

   Categorize data by significance to its consumers

3. **Identify relationships between and among data**

   To data in other projects, programs, or organizations

   To previously archived/released values for the same calculation/measurement

4. **Describe the process, technology, and infrastructure by which those data are archived**

   Describe metadata which will be stored along with the data and how minimum metadata requirements are satisfied

   Describe archival data and metadata intake process

   Describe storage strategy

5. **Specify a schedule for release of data products for public access**

   Provide information explaining why some data cannot be made publicly available

   Describe process and infrastructure used to release data to the public

6. **Identify a custodian of the plan –** *one who is responsible for its execution*

Extremely talented NIST colleagues who know much more and who have been thinking about data management for MANY years:

Dr. Donald Burgess
Editor-in-Chief, J. Physical & Chemical Reference Data
Chemical and Biochemical Reference Data Division
donald.burgess@nist.gov
Tel. (301)975-2614

Dr. Peter Linstrom
Editor-in-Chief, NIST Chemistry WebBook
Chemical and Biochemical Reference Data Division
peter.linstrom@nist.gov
Tel. (301)975-5422