Open Access to Data: Hopes and Fears From the Faculty Perspective

Bruce Clemens

Department of Materials Science and Engineering, Stanford University

Stanford Workshop on The Future of Scientific Publishing: Open Access to Manuscripts and Big Data
June 27, 2013
What is Changing and Why?

• In February OSTP issued a memo mandating open access to federally funded data
  – Federal agencies are to respond in August with a plan addressing objectives of the memo

• The goal is to promote innovation and technological development through maximizing distribution of scientific results

• What are implications for researchers?
  – So far the dialogue has mainly involved publishers, government offices and agencies, industry, and professional societies.
  – All of which have their own interests.
  – None of which solely represents the interests of the academic researcher.
OSTSP Memo

- To the extent feasible and consistent with applicable law and policy; agency mission; resource constraints; U.S. national, homeland, and economic security; and the objectives listed below, digitally formatted scientific data resulting from unclassified research supported wholly or in part by Federal funding should be stored and publicly accessible to search, retrieve, and analyze.

- Data is defined as the digital recorded factual material commonly accepted in the scientific community as necessary to validate research findings including data sets used to support scholarly publications, but does not include laboratory notebooks, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports, communications with colleagues, or physical objects, such as laboratory specimens.
OMB Circular A-110: Definition of “Data”

- Research data is defined as the recorded factual material commonly accepted by 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
  - Physical objects (e.g. laboratory samples)
  - Trade secrets, commercial information, materials necessary to be held confidential by a researcher until they are published, or similar information which is protected under law
  - Personnel or medical information and similar information the disclosure of which would constitute a clearly unwarranted invasion of personal privacy, such as information that could be used to identify a particular person in a research study
OSTP Mandate to Funding Agencies

• Maximize access, by the general public and without charge, to data created with Federal funds, while:
  – protecting confidentiality and personal privacy,
  – recognizing proprietary interests, business confidential information, and intellectual property rights and avoiding significant negative impact on intellectual property rights, innovation, and U.S. competitiveness, and
  – preserving the balance between the relative value of long-term preservation and access and the associated cost and administrative burden;

• Ensure that researchers develop data management plans, as appropriate, describing how they will provide for long-term preservation of, and access to, scientific data in digital formats, or explaining why long-term preservation and access cannot be justified;

• Allow the inclusion of appropriate costs for data management and access in proposals for Federal funding for scientific research;

• Ensure appropriate evaluation of the merits of submitted data management plans;

• Include mechanisms to ensure that intramural and extramural researchers comply with data management plans and policies;
OSTP Mandate to Funding Agencies

- Promote the deposit of data in publicly accessible databases, where appropriate and available;
- Encourage cooperation with the private sector to improve data access and compatibility, including through the formation of public-private partnerships with foundations and other research funding organizations;
- Develop approaches for identifying and providing appropriate attribution to scientific data sets that are made available under the plan;
- In coordination with other agencies and the private sector, support training, education, and workforce development related to scientific data management, analysis, storage, preservation, and stewardship; and
- Provide for the assessment of long-term needs for the preservation of scientific data in fields that the agency supports and outline options for developing and sustaining repositories for scientific data in digital formats, taking into account the efforts of public and private sector entities.
Hopes for the Academic Researcher

• Greater access to data can assist discovery and innovation in research
  – Example: the Materials Genome Database for 30,000 inorganic compounds from MIT and LBL

Example: Formation of oxynitride

\[ \text{La}_2\text{Zr}_2\text{O}_7 + 2\text{NH}_3 \rightarrow 2\text{LaZrO}_2\text{N} + 3\text{H}_2\text{O} \]

Reaction free energy

\[ \Delta G = \Delta G^\circ + RT \ln \left( \frac{P^3_{\text{H}_2\text{O}}}{P^2_{\text{NH}_3}} \right) \]

Equilibrium pressure

\[ P^\text{eq}_{\text{NH}_3} = P^3/2_{\text{H}_2\text{O}} e^{\Delta G^\circ / 2RT} \]

Knowledge of standard state free energy allows calculation of synthesis conditions for formation of desired phase

• Sharing data might increase its impact
  – Other researchers can discover aspects not studied in first pass
Fears for the Academic Researcher

Use of Archived Data

• How can the quality of the data be assured?
  – Metadata
    • That sample was grown on the substrate licked by my student’s dog!
  – Trust in experimental or computational approach and skill
    • Is peer review of datasets realistic in today’s time-constrained research environment?

• How can access to data be made clear, fast and efficient?
  – Searches, visualization, robust metrics
Fears for the Academic Researcher

Archiving Data Issues and Needs

• A clear vision of what data needs to be archived
  – Source data, all data – how far down on the data pyramid?
  – Some datasets are huge! Storing will represent a significant time and financial cost.
    • My data ate my grant!
• An efficient approach for organizing data and metadata
• Supported archive
• A clear standard for credit for archived data
  – Citations – can these be valued?
  – Co-Inventor for discoveries based on archived data?
• Ability to retain intellectual property
• PI gatekeeper power to reveal developing discoveries at appropriate time
MRS/TMS Survey on Big Data

Demographics of the 675 Respondents

• Institutions
  – 47% Academia
  – 14% Government
  – 12% Small business, Non-profit, other

• Positions
  – 40% Scientists/Engineers
  – 27% Industry
  – 26% University Professors
  – 11% Executive/Management
  – 7% Post-Doctoral Researchers
  – 8% Undergraduate and Graduate Students

• Areas
  – 48% Experimentalists
  – 32% Applied/Development
  – 13% Computational
  – 4% Theorist
MRS/TMS Survey on Big Data

- **Current Use of Software**
  - 82% For data analysis
  - 64% For data processing
  - 50% For data visualization and simulations

- **Current Use of Databases**
  - 43% Experiment design
  - 53% Interpreting/modeling of experimental data
  - 53% Input for calculations
  - 31% Model validation
  - 25% Data mining
  - 34% Materials selection
  - 20% Product engineering or design

- **Types of Data Desired**
  - 87% Physical and thermal property database
  - 62% Chemical properties database
  - Mechanical properties, electronic properties, microstructures etc
MRS/TMS Survey on Big Data

• Motivation to Share Data
  – 72% Increased visibility of research
  – 67% Receive feedback
  – 54% Opportunity to have others make discoveries

• Impediments to Sharing Data
  – 59% Proprietary/restricted nature of their data
  – 54% Intellectual property rules of organization
  – 42% Data stored in proprietary format

• Requirements for Sharing Data
  – 74% Willing to share data if encouraged as term or condition of funding
  – 46% Metadata standards and protocols should be established standards
  – 30% Metadata standards should be community-driven
MRS/TMS Survey on Big Data

- **What Data to Share for Publications**
  - 46.7% Release all data including data not used for publication
  - 42.8% Only release data used in publication
  - 10.5% Supporting data only, or it is situation-dependent, etc.

- **Ensuring Reproducibility of Data**
  - 57.3% Additional actions are needed to ensure a complete description of experimental and computational details
  - Establish standards and guidelines
  - Expand experimental/modeling sections
  - Verify data via peer review
  - Provide details/code for reproducibility

- **Wide spectrum of opinion expressed**
  - Enthusiastic embrace
  - Passionate objection
Acknowledgments

• **MRS** –
  – Todd Osman
  – Eileen Novak
  – Ron Kelly

• **TMS** - Sharing survey results

• **Co-Organizers** –
  – Tom Baer
  – Robert Byer
  – Amy Hodge
  – Julie Sweetkind-Singer
  – Bob Schwarzwalder

• **Sponsors**
  – Stanford School of Engineering
  – Stanford Dean of Research
  – Stanford Photonics Research Center
  – Stanford University Libraries
  – The American Physical Society

• Sara Lefort