Open Access to Data: Policy Mandate and Scientific Imperative

Victoria Stodden
Department of Statistics
Columbia University

The Future of Scientific Publishing: Open Access to Manuscripts and Big Data
Stanford University
June 27, 2013

Two Parts

1. What does the OSTP mandate mean?
2. What does the implementation mean?
    - Opportunities
    - Challenges and caveats

Part 1: Data in the Memorandum

1. Digitally formatted data arising from federal grants should be stored and publicly accessible to search, retrieve, and analyze.

2. “[D]ata is defined, consistent with OMB circular A-110, 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”

Each Public Access Plan Shall...

a) Maximize access, by the general public and without charge, to digitally formatted scientific data created with Federal funds, while [respecting privacy, proprietary interests and IP, need for long-term preservation],

b) Ensure that all ... researchers receiving Federal grants ... develop data management plans, as appropriate, describing how they will provide for long-term preservation of, and access to, scientific data in digital formats resulting from federally funded research, or explaining why longterm preservation and access cannot be justified,

c) Allow the inclusion of appropriate costs for data management and access in proposals for Federal funding for scientific research,

d) Ensure appropriate evaluation of the merits of submitted data management plans,

e) Include mechanisms to ensure that intramural and extramural researchers comply with data management plans and policies,

f) Promote the deposit of data in publicly accessible databases, where appropriate and available,

g) 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,

h) Develop approaches for identifying and providing appropriate attribution to scientific data sets that are made available under the plan,

i) In coordination with other agencies and the private sector, support ... workforce development related to scientific data management, analysis, storage, preservation, and stewardship, and

j) 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.

Part 2: Implementation

Now from the scientist's perspective..

Open Data Crucial to Science Today

- not a new concept, rooted in *skepticism*
- Transactions of the Royal Society 1660's
- Transparency, knowledge transfer -> goal to perfect the *scholarly record*. Nothing else.
- Technology has changed the nature of experimentation, data, and communication.

Computation is Becoming Central to Scientific Research

1. enormous, and increasing, amounts of data collection:
   - CMS project at LHC: 300 "events" per second, 5.2M seconds of runtime per year, .5MB per event = 780TB/yr => several PB when data processed,
   - Sloan Digital Sky Survey: 9th data release (SDSS-III 2012), 60TB,
   - quantitative revolution in social science due to abundance of social network data (Lazier et al, *Science*, 2009)
   - *Science* survey of peer reviewers: 340 researchers regularly work with datasets >100GB; 119 regularly work with datasets >1TB (N=1700, Feb 11, 2011, p. 692)
2. massive simulations of the complete evolution of a physical system, systematically varying parameters,
3. deep intellectual contributions now encoded in software.

Credibility Crisis

JASA June	Computational Articles	Code Publicly Available
1996	9 of 20	0%
2006	33 of 35	9%
2009	32 of 32	16%
2011	29 of 29	21%

Ioannidis (2011): 9% of authors studied made data available.

Generally, data and code not made available at the time of publication, insufficient information in the publication for verification, replication of results. ***A Credibility Crisis***

Scientific Perspective

"Really Reproducible Research" pioneered by Stanford Professor Jon Claerbout:

"The idea is: An article about computational science in a scientific publication is *not* the scholarship itself, it is merely *advertising* of the scholarship. The actual scholarship is the complete ... set of instructions [and data] which generated the figures."

paraphrased by David Donoho, 1998.

Updating the Scientific Method

Argument: computation presents only a *potential* third branch of the scientific method (Stodden et al 2009):

- Branch 1 (deductive): mathematics, formal logic,
- Branch 2 (empirical): statistical analysis of controlled experiments,
- Branch 3,4? (computational): large scale simulations / data driven computational science.

The Ubiquity of Error

- The central motivation for the scientific method is to root out error:
  - Deductive branch: the well-defined concept of the proof,
  - Empirical branch: the machinery of hypothesis testing, structured communication of methods and protocols.

- Computational science as practiced today does not generate reliable knowledge. See e.g. Ioannidis, "Why Most Published Research Findings are False," PLoS Med, 2005.

Openness in Science

- Science Policy must support scientific ends: Reliability and accuracy of the scientific record.
- Facilitate Reproducibility - the ability to regenerate published computational results (data and code availability, alongside results).
- Need infrastructure to facilitate (1):
    1. deposit/curation of data and code,
    2. link to published article,
    3. permanence of link.

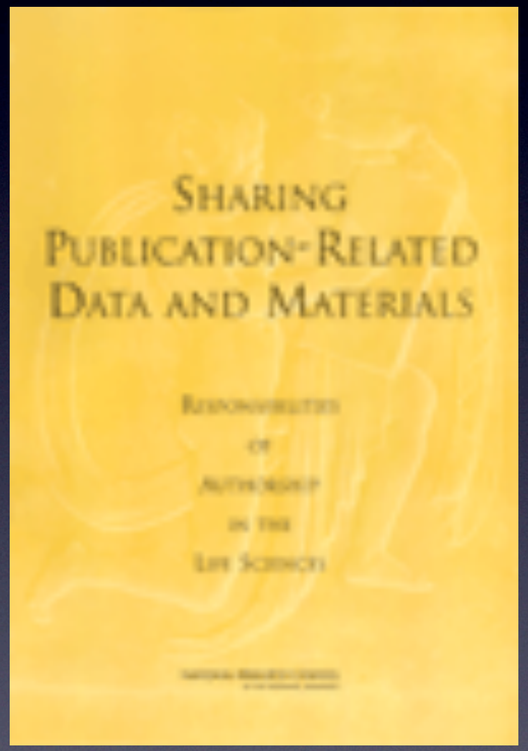
Science Policy

- "Open Data" is not well-defined. Scope: Share data and code that *permit others in the field to replicate published results*. (traditionally done by the publication alone).
- Data and code availability at the time of publication.
- Public access. "With many eyeballs, all bugs are shallow." Recall: primary goal of the scientific method to root out error.
- Need infrastructure/software tools to facilitate (2): Data/code suitable for sharing, created *during the research process*.

Scientific Research Varies Widely

- Different research questions call for different tools, solutions, and implementations to reach "really reproducible research."
- Spectrum from data-driven research to empirical research carried out out entirely in software (simulations).
- "Data" has very different meanings depending on the research.
- Overspecification of how to reach goals will not work, for either infrastructure or tools. Empower communities to reach clearly specified goals that support science, with funds, deadlines, and enforcement (and community engagement in the process).

NAS Data Sharing Report



- Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences, (2003)
- "Principle 1. Authors should include in their publications the data, algorithms, or other information that is central or integral to the publication—that is, whatever is necessary to support the major claims of the paper and would enable one skilled in the art to verify or replicate the claims."

Journal Data Sharing Policy

	2011	2012	Change
Required as condition of publication, barring exceptions	18	19	1
Required but may not affect editorial decisions	3	10	7
Encouraged/addressed, may be reviewed and/or hosted	35	30	-5
Implied	0	5	5
No mention	114	106	-8

Source: Stodden, Guo, Ma (2013) PLoS ONE, 8(6)

Tools for Computational Science

- Dissemination Platforms:

  RunMyCode.org | IPOL | Madagascar
  MLOSS.org | thedatahub.org | nanoHUB.org
  Open Science Framework

- Workflow Tracking and Research Environments:

  VisTrails | Kepler | CDE
  Galaxy | GenePattern | Paper Mâché
  Sumatra | Taverna | Pegasus

- Embedded Publishing:

  Verifiable Computational Research | Sweave
  Collage Authoring Environment | SHARE

A Grassroots Movement

- AMP 2011 "Reproducible Research: Tools and Strategies for Scientific Computing"
- Open Science Framework / Reproducibility Project in Psychology
- AMP / ICIAM 2011 "Community Forum on Reproducible Research Policies"
- SIAM Geosciences 2011 "Reproducible and Open Source Software in the Geosciences"
- ENAR International Biometric Society 2011: Panel on Reproducible Research
- AAAS 2011: "The Digitization of Science: Reproducibility and Interdisciplinary Knowledge Transfer"
- SIAM CSE 2011: "Verifiable, Reproducible Computational Science"
- Yale 2009: Roundtable on Data and Code Sharing in the Computational Sciences
- ACM SIGMOD conferences
- NSF/OCI report on Grand Challenge Communities (Dec, 2010)
- IOM "Review of Omics-based Tests for Predicting Patient Outcomes in Clinical Trials"
- ...