

# Re-use and Reproducibility: Opportunities and Challenges

Victoria Stodden  
Department of Statistics  
Columbia University

Open Repositories 2013 Keynote  
Prince Edward Island, Canada  
July 9, 2013

# Agenda

- Opportunities:
  - culture change regarding digital scholarly objects
  - federal directives regarding open data
  - tool development facilitating shareable objects
- Challenges:
  - defining “digital scholarly object”
  - providing appropriate sharing modalities, ie for code
  - intellectual property law
  - partnerships across academic silos, incentives to share

# 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

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

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%

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” inspired 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.” 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. “Breezy demos” of results.



# Openness in Science

- Science Policy must support scientific ends: Reliability and accuracy of the scientific record.
- Facilitate Reproducibility - the ability to regenerate published results (data and code availability, alongside results).
- Need infrastructure to facilitate (I):
  1. deposit/curation of versioned 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). Corollary: maximizes data reuse.
- Data and code availability at the time of publication.
- Public access. “With many eyeball, all bugs are shallow.”
- Need infrastructure/software tools to facilitate (2):
  1. 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.”
- Questions can be solely data-driven research to empirical research contained entirely in software (simulations).
- “Data” has very different meanings depending on the question behind the research.
- Empower communities to reach clearly specified goals that support science, with funds, deadlines, and enforcement (and community engagement in the process).

# Sharing: Funding Agency Policy

- NSF grant guidelines: “NSF ... **expects investigators to share** with other researchers, at no more than incremental cost and within a reasonable time, **the data**, samples, physical collections and other supporting materials created or gathered in the course of the work. It also encourages grantees to **share software** and inventions or otherwise act to make the innovations they embody widely useful and usable.” (2005 and earlier)
- NSF peer-reviewed Data Management Plan (DMP), January 2011.
- NIH (2003): “The NIH **expects** and supports the timely release and **sharing of final research data** from NIH-supported studies for use by other researchers.” (>\$500,000, include data sharing plan)

# NSF Data Management Plan

“Proposals submitted or due on or after January 18, 2011, must include a supplementary document of no more than two pages labeled ‘Data Management Plan.’ This supplementary document should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results.” (<http://www.nsf.gov/bfa/dias/policy/dmp.jsp>)

Software management plans appearing.. (BigData joint NSF/NIH solicitation)

# 2013: Open Science in DC

- Feb 22: Executive Memorandum directing federal funding agencies to develop plans for public access to data and publications.
- May 9: Executive Order directing federal agencies to make their data publicly available.

# Science Policy in Congress

- America COMPETES due to be reauthorized, drafting underway,
- Hearing on Research Integrity and Transparency by the House Science, Space, and Technology Committee (March 5).
- Reproducibility cannot be an unfunded mandate.

# National Science Board Report

NSB-11-79  
December 14, 2011

Prepublication Copy



## Digital Research Data Sharing and Management

December 2011

Task Force on Data Policies  
Committee on Strategy and Budget  
National Science Board

“Digital Research Data Sharing and Management,”  
December 2011.

[http://www.nsf.gov/nsb/publications/2011/  
nsb1124.pdf](http://www.nsf.gov/nsb/publications/2011/nsb1124.pdf)



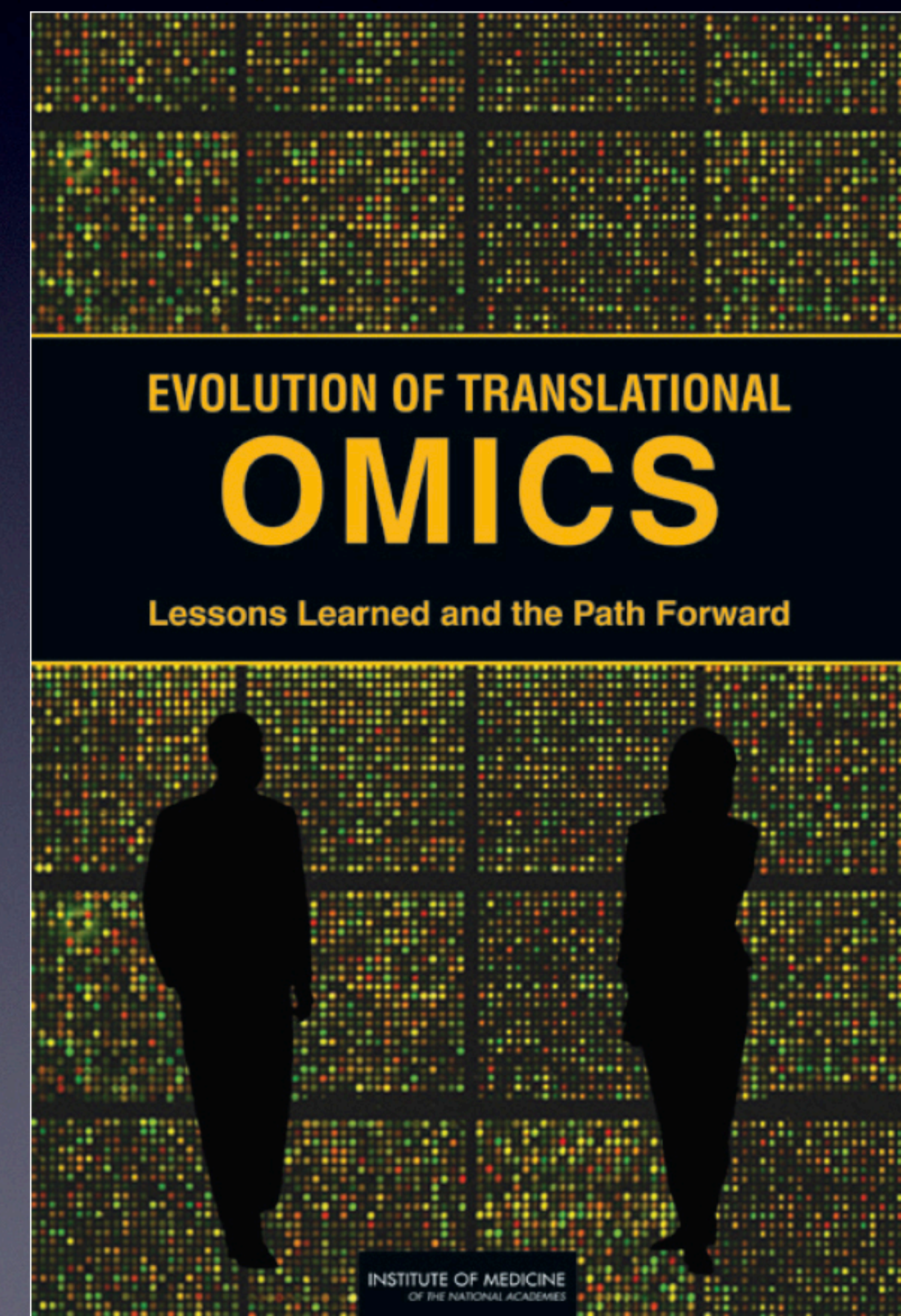
# NAS Data Sharing Report



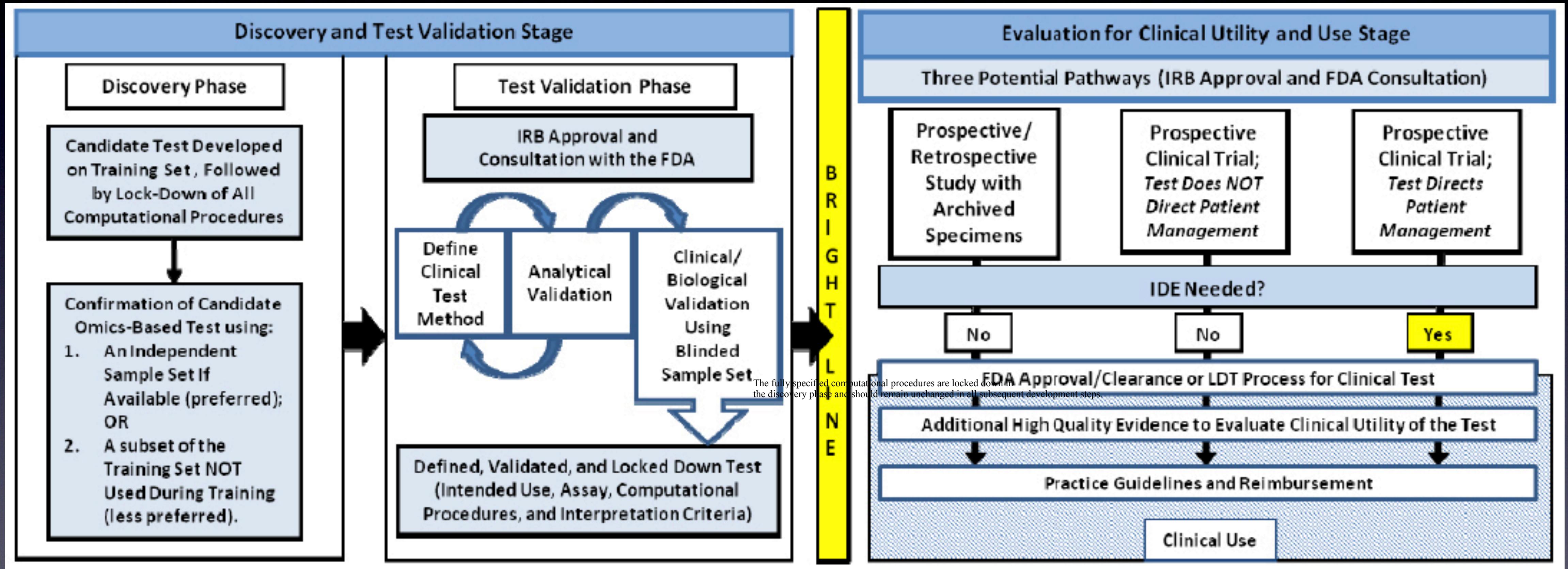
- Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences, (2003)
- “Principle 1. Authors should include in their publications the data, algorithms, or other information that is central or integral to the publication—that is, whatever is necessary to support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

# IOM “Evolution of Translational Omics: Lessons Learned and the Path Forward”

- March 23 2012, IOM releases report,
- Recommends new standards for omics-based tests, including a fixed version of the software, expressly for verification purposes.



# IOM Report: Figure S-1



“The fully specified computational procedures are locked down in the discovery phase and should remain unchanged in all subsequent development steps.”

# Legal Barriers: Copyright

“To promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries.” (U.S. Const. art. I, §8, cl. 8)

- Original expression of ideas falls under copyright *by default* (papers, code, figures, tables..)
- Copyright secures exclusive rights vested in the author to:
  - reproduce the work
  - prepare derivative works based upon the original

Exceptions and Limitations: Fair Use.

# Response from Within the Sciences

## The *Reproducible Research Standard (RRS)* (Stodden, 2009)

- A suite of license recommendations for computational science:
  - Release media components (text, figures) under CC BY,
  - Release code components under Modified BSD or similar,
  - Release data to public domain or attach attribution license.

➔ Remove copyright's barrier to reproducible research and,

➔ Realign the IP framework with longstanding scientific norms.

Winner of the Access to Knowledge Kultura Award 2008

# Copyright and Data

- Copyright adheres to raw facts in Europe.
- In the US raw facts are not copyrightable, but the original “selection and arrangement” of these facts is copyrightable. (Feist Publns Inc. v. Rural Tel. Serv. Co., 499 U.S. 340 (1991)).
- the possibility of a residual copyright in data (attribution licensing or public domain certification).
- Law doesn't match reality on the ground: What constitutes a “raw” fact anyway?

# Sharing: Journal Policy

- Journal Policy snapshots June 2011 and June 2012:
- Select all journals from ISI classifications “Statistics & Probability,” “Mathematical & Computational Biology,” and “Multidisciplinary Sciences” (this includes Science and Nature).
- $N = 170$ , after deleting journals that have ceased publication.

# 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)



# Journal Code Sharing Policy

	2011	2012	Change
Required as condition of publication, barring exceptions	6	6	0
Required but may not affect editorial decisions	6	6	0
Encouraged/addressed, may be reviewed and/or hosted	17	21	4
Implied	0	3	3
No mention	141	134	-7

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

# Findings

- Changemakers are journals with high impact factors.
- Progressive policies are not widespread, but being adopted rapidly.
- Close relationship between the existence of a supplemental materials policy and a data policy.
- Data and supplemental material policies appear to lead software policy.

# Barriers to Journal Policy Making

- Standards for code and data sharing,
- Meta-data, archiving, re-use, documentation, sharing platforms, citation standards,
- Review, who checks replication pre-publication, if anyone,
- Burdens on authors, especially less technical authors,
- Evolving, early research; affects decisions on when to publish,
- Business concerns, attracting the best papers.

# 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”
- ...

# References

- “Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals,” PLoS ONE, June 2013
- “Reproducible Research,” guest editor for Computing in Science and Engineering, July/August 2012.
- “Reproducible Research: Tools and Strategies for Scientific Computing,” July 2011.
- “Enabling Reproducible Research: Open Licensing for Scientific Innovation,” 2009.

available at <http://www.stodden.net>