What is Reproducibility?

Victoria Stodden

Journée du GdR ISIS
Reproducibility in signal and image processing
Telecom ParisTech, Paris

January 16, 2014
Agenda

• History and Rationale
• Solutions and Experiences: Tool Development
• Other supporting developments
  - culture change regarding digital scholarly objects
  - federal directives regarding open data
• Challenges:
  - defining “digital scholarly object”
  - providing appropriate sharing modalities, ie for code
  - intellectual property law
  - partnerships across academic silos, incentives to share
Reproducibility is an Old Concept

• rooted in skepticism

• Robert Boyle and the Transactions of the Royal Society 1660’s

• Transparency, knowledge transfer -> goal to perfect the **scholarly record**.

• Advances in the technology used for scientific discovery have changed how we effect reproducibility and standards for communication.
Defining Reproducibility

• Empirical Reproducibility: in meatspace
  - traditional notion of reproducibility (Boyle): sufficient description in the research paper for others in the field to replicate
  - access to reagents, cell lines, and other physical materials required to replicate the research
  - Begley, Nature “Try Harder”

• Computational Reproducibility: in silica
  - access to digital data and code used to generate findings
  - execution?
  - algorithm descriptions
Communication Standards: Data and Code

- Gentleman and Temple Lang proposed the *Research Compendia* (2003): computational scholarly communication as a triple of narrative, data, and code.


- Data sharing in genomics beginning in 1996.

- Gary King (1998) advocated reproducibility in political science and the social sciences.
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 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.
Persistently connect:
- data output,
- code,
- narrative/publications.

Goals:
- link research outputs for reproducibility and reuse,
- citation standards,
- open source.
A Comparative Study of Unit Root Tests with Panel Data and a New Simple Test


This Matlab code computes the Fisher (1932) type panel unit root tests, proposed by Choi (2001) and Maddala and Wu (1999). Both tests combine the significant levels obtained from individual ADF tests. The only required inputs is the \((T,N)\) matrix of data, where \(T\) is the time dimension and \(N\) is the cross section one. The user can choose the deterministic component: with no individual effects (model 1), with individual effects but no time trends (model 2), and with individual effects and time trends (model 3). The individual lag orders are determined according to the BIC information criteria or provided ...

A Constrained Random Demodulator for Sub-Nyquist Sampling

IEEE Transactions on Signal Processing (2013)

The code can be used to reproduce the simulations presented in the associated paper or to run similar simulations. The code uses SpaRSA to calculate the Lasso solution and YALL1 to calculate the basis pursuit solution to finding spectral coefficients.
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).
Openness in Science

• Need infrastructure to facilitate, at the time of publication, (at least):
  1. deposit/curation of versioned data and code,
  2. link to published article,
  3. permanence of link.

• Need infrastructure/software tools to facilitate:
  1. data/code suitable for sharing, created during the research process
  2. Public access. “With many eyeballs, all bugs are shallow.”
Mandates: 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)
“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.
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.
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

<table>
<thead>
<tr>
<th>Requirement</th>
<th>2011</th>
<th>2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>Required as condition of publication, barring exceptions</td>
<td>10.6%</td>
<td>11.2%</td>
</tr>
<tr>
<td>Required but may not affect editorial decisions</td>
<td>1.7%</td>
<td>5.9%</td>
</tr>
<tr>
<td>Encouraged/addressed, may be reviewed and/or hosted</td>
<td>20.6%</td>
<td>17.6%</td>
</tr>
<tr>
<td>Implied</td>
<td>0%</td>
<td>2.9%</td>
</tr>
<tr>
<td>No mention</td>
<td>67.1%</td>
<td>62.4%</td>
</tr>
</tbody>
</table>

Source: Stodden, Guo, Ma (2013) PLoS ONE, 8(6)
## Journal Code Sharing Policy

<table>
<thead>
<tr>
<th>Requirement</th>
<th>2011</th>
<th>2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>Required as condition of publication, barring exceptions</td>
<td>3.5%</td>
<td>3.5%</td>
</tr>
<tr>
<td>Required but may not affect editorial decisions</td>
<td>3.5%</td>
<td>3.5%</td>
</tr>
<tr>
<td>Encouraged/addressed, may be reviewed and/or hosted</td>
<td>10%</td>
<td>12.4%</td>
</tr>
<tr>
<td>Implied</td>
<td>0%</td>
<td>1.8%</td>
</tr>
<tr>
<td>No mention</td>
<td>82.9%</td>
<td>78.8%</td>
</tr>
</tbody>
</table>

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:
  - ResearchCompendia.org
  - MLOSS.org
  - Open Science Framework
  - IPOL
  - thedatahub.org
  - Madagascar
  - CDE
  - nanohub.org
  - RunMyCode.org

• Workflow Tracking and Research Environments:
  - VisTrails
  - Galaxy
  - Sumatra
  - Kepler
  - GenePattern
  - Taverna
  - Paper Mâché
  - CDE
  - Pegasus

• Embedded Publishing:
  - Verifiable Computational Research
  - Collage Authoring Environment
  - Sweave
  - 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


• “Reproducible Research,” guest editor for Computing in Science and Engineering, July/August 2012.


available at http://www.stodden.net