Policy Update on Digital Scholarly Objects

Victoria Stodden
Graduate School of Information and Library Science
University of Illinois at Urbana-Champaign

NIH/NLM PMC Advisory Committee
Bethesda, MD
June 9, 2015
1. Empirical Reproducibility

The reproduction of results is the cornerstone of science; yet, at times, reproducing the results of others can be a difficult challenge. Our two laboratories, one on the East and the other on the West Coast of the United States, decided to collaborate on a problem of mutual interest—namely, the heterogeneity of the human breast. Despite using seemingly identical methods, reagents, and species, our two laboratories quite reproducibly were unable to replicate each other’s fluorescence-activated cell sorting (FACS) profiles of primary breast cells. Frustration of studying cells close to their context in vivo makes the exercise even more challenging.

Paired with in situ characterizations, FACS has emerged as the technology most suitable for distinguishing diversity among different cell populations in the mammary gland. Flow instruments have evolved from being able to detect only a few parameters to those now capable of measuring up to—and beyond—an astonishing 50 individual markers per cell (Cheung and Uytz, 2011). As with any exponential increase in data complexity, breast reduction mammoplasties. Molecular analysis of separated fractions was to be performed in Boston (K.P.’s laboratory, Dana-Farber Cancer Institute, Harvard Medical School), whereas functional analysis of separated cell populations grown in 3D matrices was to take place in Berkeley (M.J.B.’s laboratory, Lawrence Berkeley National Lab, University of California, Berkeley). Both our laboratories have decades of experience and established protocols for isolating cells from primary normal breast tissues as well as the capabilities required for...
2. Computational Reproducibility

Traditionally two branches to the scientific method:

- Branch 1 (deductive): mathematics, formal logic,
- Branch 2 (empirical): statistical analysis of controlled experiments.

Now, new branches due to technological changes?

- Branch 3,4? (computational): large scale simulations / data driven computational science.
Commonly believed...

“It is common now to consider computation as a third branch of science, besides theory and experiment.”

“This book is about a new, fourth paradigm for science based on data-intensive computing.”
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, appropriate statistical methods, structured communication of methods and protocols.

**Claim:** Computation presents only a potential third/fourth branch of the scientific method (Donoho, Stodden, et al. 2009), until the development of comparable standards.
The Impact of Technology

1. **Big Data / Data Driven Discovery**: high dimensional data, e.g. $p \gg n$,

2. **Computational Power**: simulation of the complete evolution of a physical system, systematically varying parameters,

3. Deep intellectual contributions now encoded only in software.

The software contains “ideas that enable biology...”
*Stories from the Supplement, 2013.*
3. Reproducibility as a Statistical Issue

- False discovery, chasing significance, p-hacking (Simonsohn 2012), file drawer problem, overuse and mis-use of p-values,

- Multiple testing, sensitivity analysis, poor reporting/tracking practices,

- Data preparation, treatment of outliers,

- Poor statistical methods (nonrandom sampling, inappropriate tests or models,..)

- Model robustness to parameter changes and data perturbations,

- Investigator bias toward previous findings; conflicts of interest.
Journal Requirements for Statistical Reproducibility

In January 2014 Science enacted new policies. The will check for:

1. a “data-handling plan” i.e. how outliers will be dealt with,

2. sample size estimation for effect size,

3. whether samples are treated randomly,

4. whether experimenter blind to the conduct of the experiment.

Statisticians added to the Board of Reviewing Editors.
Journal Policy on Data/Code

• Journal Policy setting study design:

• 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.

• Create dataset with ISI information (impact factor, citations, publisher) and supplement with publication policies as listed on journal websites, in June 2011 and June 2012.
## Journal Data Sharing Policy

<table>
<thead>
<tr>
<th>Policy Description</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.

• No statistically significant relationship between data and code policies and open access policy.

• Data and supplemental material policies appear to lead software policy.
Open Science from the Whitehouse

- Feb 22, 2013: Executive Memorandum directing federal funding agencies to develop plans for public access to data and publications.

- May 9, 2013: Executive Order directing federal agencies to make their data publicly available.

“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)
DOE Data Management Plan

“The Department is taking a phased approach to the implementation of requirements set forth by the OSTP memo. In particular, the Office of Science, which supports roughly two-thirds of the total R&D for the Department, plans to pilot a data management policy with the requirements described below by July 28, 2014. Other DOE Offices and elements with over $100 million in annual conduct of research and development expenditures will implement data management plan requirements that satisfy the requirements of the OSTP memo no later than October 1, 2015 in such a way that there is a single DOE policy for data management planning.” (DOE Public Access Plan 2014)
NSF Discussions

- ACCI WG on Data and Code Accessibility
- CyberInfrastructure discussions on data
Data / Code Sharing Practices

Survey of the NIPS community:

• 1,758 NIPS registrants up to and including 2008,
• 1,008 registrants when restricted to .edu registration emails,
• After piloting, the final survey was sent to 638 registrants,
• 37 bounces, 5 away, and 3 in industry, gave a final response rate was 134 of 593 or 23%.
• Queried about reasons for sharing or not sharing data/code associated with their NIPS paper.
## Sharing Incentives

<table>
<thead>
<tr>
<th>Code</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>91%</td>
<td>Encourage scientific advancement</td>
</tr>
<tr>
<td>90%</td>
<td>Encourage sharing in others</td>
</tr>
<tr>
<td>86%</td>
<td>Be a good community member</td>
</tr>
<tr>
<td>82%</td>
<td>Set a standard for the field</td>
</tr>
<tr>
<td>85%</td>
<td>Improve the calibre of research</td>
</tr>
<tr>
<td>81%</td>
<td>Get others to work on the problem</td>
</tr>
<tr>
<td>85%</td>
<td>Increase in publicity</td>
</tr>
<tr>
<td>78%</td>
<td>Opportunity for feedback</td>
</tr>
<tr>
<td>71%</td>
<td>Finding collaborators</td>
</tr>
</tbody>
</table>

Survey of the Machine Learning Community, NIPS (Stodden 2010)
## Barriers to Sharing

<table>
<thead>
<tr>
<th>Code</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>77%</td>
<td>54% Time to document and clean up</td>
</tr>
<tr>
<td>52%</td>
<td>34% Dealing with questions from users</td>
</tr>
<tr>
<td>44%</td>
<td>42% Not receiving attribution</td>
</tr>
<tr>
<td>40%</td>
<td>41% Possibility of patents</td>
</tr>
<tr>
<td>34%</td>
<td>38% Legal Barriers (ie. copyright)</td>
</tr>
<tr>
<td>-</td>
<td>35% Time to verify release with admin</td>
</tr>
<tr>
<td>30%</td>
<td>33% Potential loss of future publications</td>
</tr>
<tr>
<td>30%</td>
<td>29% Competitors may get an advantage</td>
</tr>
<tr>
<td>20%</td>
<td>29% Web/disk space limitations</td>
</tr>
</tbody>
</table>

Survey of the Machine Learning Community, NIPS (Stodden 2010)
Reproducibility in Computational and Experimental Mathematics (December 10-14, 2012)

Description
In addition to advancing research and discovery in pure and applied mathematics, computation is pervasive across the sciences and now computational research results are more crucial than ever for public policy, risk management, and national security. Reproducibility of carefully documented experiments is a cornerstone of the scientific method, and yet is often lacking in computational mathematics, science, and engineering. Setting and achieving appropriate standards for reproducibility in computation poses a number of interesting technological and social challenges. The purpose of this workshop is to discuss aspects of reproducibility most relevant to the mathematical sciences among researchers from pure and applied mathematics from academics and other settings, together with interested parties from funding agencies, national laboratories, professional societies, and publishers. This will be a working workshop, with relatively few talks and dedicated time for breakout group discussions on the current state of the art and the tools, policies, and infrastructure that are needed to improve the situation. The groups will be charged with developing guides to current best practices and/or white papers on desirable advances.

Organizing Committee
- David H. Bailey
  (Lawrence Berkeley National Laboratory)
- Jon Borwein
  (Centre for Computer Assisted Research Mathematics and its Applications)
- Randall J. LeVeque
  (University of Washington)
- Bill Rider
  (Sandia National Laboratory)
- William Stein
  (University of Washington)
- Victoria Stodden
  (Columbia University)
ICERM Workshop Report

Setting the Default to Reproducible

Reproducibility in Computational and Experimental Mathematics

Developed collaboratively by the ICERM workshop participants

Compiled and edited by the Organizers

V. Stodden, D. H. Bailey, J. Borwein, R. J. LeVeque, W. Rider, and W. Stein

Abstract

Science is built upon foundations of theory and experiment validated and improved through open, transparent communication. With the increasingly central role of computation in scientific discovery this means communicating all details of the computations needed for others to replicate the experiment, i.e., making available to others the associated data and code. The "reproducible research" movement recognizes that traditional scientific research and publication practices now fall short of this ideal, and encourages all those involved in the production of computational science — scientists who use computational methods and the institutions that employ them, journals and dissemination mechanisms, and funding agencies — to facilitate and practice really reproducible research.

Set the Default to “Open”

Reproducible Science in the Computer Age. Conventional wisdom sees computing as the "third leg" of science, complementing theory and experiment. That metaphor is outdated. Computing now pervades all of science. Massive computation is often required to reduce and analyze data; simulations are now employed in fields as diverse as climate modeling and astrophysics. Unfortunately, scientific computing culture has not kept pace. Experimental researchers are taught early to keep notebooks or computer logs of every work detail: design, procedures, equipment, raw results, processing techniques, statistical methods of analysis, etc. In contrast, few computational experiments are performed with such care. Typically, there is no record of workflow, computer hardware and software configuration, or parameter settings. Often source code is lost. While stripping reproducibility of results, these practices ultimately impede the researcher’s own productivity.

The State of Experimental and Computational Mathematics. Experimental mathematics — application of high-performance computing technology to research questions in pure and applied mathematics, including physics, legal scholars, journal editors, and funding agency officials representing academia, government labs, industry research, and all points in between. While

While it’s just a grumble things like inventories and payrolls, and it wants to make some break-throughs in astrophysics.

"It says it’s sick of doing things like inventories and payrolls, and it wants to make some breakthroughs in astrophysics."

Society for Industrial and Applied Mathematics

SIAM NEWS

“Setting the Default to Reproducible” in Computational Science Research

June 3, 2013

Following a late-2012 workshop at the Institute for Computational and Experimental Research in Mathematics, a group of computational scientists have proposed a set of standards for the dissemination of reproducible research.

Victoria Stodden, Jonathan Borwein, and David H. Bailey
reproducibility@XSEDE: An XSEDE14 Workshop

Overview

The reproducibility@XSEDE workshop is a full-day event scheduled for Monday, July 14, 2014 in Atlanta, GA. The workshop will take place in conjunction with XSEDE14 (conferences.xsede.org), the annual conference of the Extreme Science and Engineering Discovery Environment (XSEDE), and will feature an interactive, open-ended, discussion-oriented agenda focused on reproducibility in large-scale computational science. Consistent with the overall XSEDE14 conference theme, we seek to engage participants from a broad range of backgrounds, including practitioners whose computational interests extend beyond traditional modeling and simulation as well as decision-makers and other professionals whose work informs and determines the direction of computation-enabled research. We hope to help...
Supporting Computational Science

• Dissemination Platforms:
  ResearchCompendia.org  IPOL  Madagascar
  MLOSS.org  thedatahub.org  nanoHUB.org
  Open Science Framework

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

• Embedded Publishing:
  Verifiable Computational Research  SOLE  knitR
  Collage Authoring Environment  SHARE  Sweave
Goal: improve understanding of reproducible computational science, trace sources of error.

- link data/code to published claims,
- enable re-use,
- sharing guide for researchers,
- certification of results,
- large scale validation of findings,
- stability, sensitivity checks.
Is “Huh?” a Universal Word? Conversational Infrastructure and the Convergent Evolution of Linguistic Items

Mark Dingemanse, Francisco Torreira, N. J. Enfield, Johan J. Bolhuis

Code and Data Abstract

A word like Huh?—used as a repair initiator when, for example, one has not clearly heard what someone just said—is found in roughly the same form and function in spoken languages across the globe. We investigate it in naturally occurring conversations in ten languages and present evidence and arguments for two distinct claims: that Huh? is universal, and that it is a word. In support of the first, we show that the similarities in form and function of this interjection across languages are much greater than expected by chance. In support of the second claim we show that it is a lexical, conventionalised form that has to be learnt, unlike grunts or emotional cries. We discuss possible reasons for the cross-linguistic similarity and propose an account in terms of convergent evolution. Huh? is a universal word not because it is innate but because it is shaped by selective pressures in an interactional environment that all languages share: that of other-initiated repair. Our proposal enhances evolutionary models of language change by suggesting that conversational infrastructure can drive the convergent cultural evolution of linguistic items.

Compendium Type: article
Content License: CC0
Code License: MIT
Random survival forests for high-dimensional data

Hemant Ishwaran, Udaya B. Kogalur, Xi Chen, Andy J. Minn

Code and Data Abstract

Minimal depth is a dimensionless order statistic that measures the predictiveness of a variable in a survival tree. It can be used to select variables in high-dimensional problems using Random Survival Forests (RSF), a new extension of Breiman's Random Forests (RF) to survival settings. We review this methodology and demonstrate its use in high-dimensional survival problems using a public domain R-language package randomSurvivalForest. We discuss effective ways to regularize forests and discuss how to properly tune the RF parameters 'nodesize' and 'mtry'. We also introduce new graphical ways of using minimal depth for exploring variable relationships.


Code DOI: doi:10.7938/M1H41PBB.
Data DOI: doi:10.7938/M1CC0XMM.

Compendium Type: Journal or Magazine Articles
Primary Research Field: Computer and Information Sciences
Secondary Research Field: Mathematics
Content License: Public Domain Mark
Code License: MIT License

Verification

verification code

Recent Verifications

Verification run 41 created March 21, 2014, 9:43 a.m.
Verification run 40 created March 21, 2014, 4:19 a.m.
A proof of concept for a research compendia webapp http://researchcompendia.org — Edit

<table>
<thead>
<tr>
<th>Branch</th>
<th>Author</th>
<th>Date</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>develop</td>
<td>codersquid</td>
<td>30 minutes ago</td>
<td>bump revision</td>
</tr>
<tr>
<td>docs</td>
<td></td>
<td>10 days ago</td>
<td>removes instructions for envdir and bootstrap.sh, adds instructions for</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>envdir</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>requirements citation dialog and display for journals</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.gitignore adds vagrant and bootstrap starter</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 months ago</td>
<td>fixed broken doi service test and updated irc channel for travis</td>
</tr>
<tr>
<td>AUTHORS.rst</td>
<td></td>
<td></td>
<td>renaming project from tyler to researchcompendia</td>
</tr>
<tr>
<td>CITATION.bib</td>
<td></td>
<td></td>
<td>bump revision</td>
</tr>
<tr>
<td>CONTRIBUTING.rst</td>
<td></td>
<td></td>
<td>fixed thinko of 'comment' to 'commit'</td>
</tr>
<tr>
<td>HISTORY.rst</td>
<td></td>
<td></td>
<td>bump revision</td>
</tr>
<tr>
<td>LICENSE</td>
<td></td>
<td>4 months ago</td>
<td>release 1.0.0-alpha1</td>
</tr>
<tr>
<td>MANIFEST.in</td>
<td></td>
<td>5 months ago</td>
<td>making skeleton docs</td>
</tr>
</tbody>
</table>
The MIT License (MIT)

Copyright (c) 2013 Sheila Miguez, Victoria Stodden, Jennifer Seiler

Permission is hereby granted, free of charge, to any person obtaining a copy
of this software and associated documentation files (the "Software"), to
deal in the Software without restriction, including without limitation the
rights to use, copy, modify, merge, publish, distribute, sublicense, and/or
sell copies of the Software, and to permit persons to whom the Software is
furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in
all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR
IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY,
FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE
AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER
LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING
FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS
IN THE SOFTWARE.
We need:

Standards for reproducibility of big data findings:

1. data access, software access, persistent linking to publications.

2. innovation around data and code access for privacy protection and scale.

3. robust methods, producing stable results, emphasis on reliability and reproducibility.

References


available at http://www.stodden.net
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.
**Responses Outside the Sciences 1: Open Source Software**

- Software with licenses that communicate alternative terms of use to code developers, rather than the copyright default.

- Hundreds of open source software licenses:
  - GNU Public License (GPL)
  - (Modified) BSD License
  - MIT License
  - Apache 2.0 License
  - ... see [http://www.opensource.org/licenses/alphabetical](http://www.opensource.org/licenses/alphabetical)

Adapts the Open Source Software approach to artistic and creative digital works.
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 Kaltura 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?
Bayh-Dole Act (1980)

Promote the transfer of academic discoveries for commercial development, via licensing of patents (ie. Technology Transfer Offices),

Bayh-Dole Act gave federal agency grantees and contractors title to government-funded inventions and charged them with using the patent system to aid disclosure and commercialization of the inventions.

Greatest impact in biomedical research collaborations and drug discovery. Now, software patents also impact science.
Ownership of Research Codes

Patent and Copyright Agreement for Personnel at Stanford - SU18

I understand that, consistent with applicable laws and regulations, Stanford University is governed in the handling of intellectual property by its official policies titled Inventions, Patents and Licensing and Copyright Policy (both published in the Research Policy Handbook), and I agree to abide by the terms and conditions of those policies, as they may be amended from time to time.

Pursuant to those policies, and in consideration of my employment by Stanford, the receipt of remuneration from Stanford, participation in projects administered by Stanford, access to or use of facilities or resources provided by Stanford and/or other valuable consideration, I hereby agree as follows:

1. I will disclose to Stanford all potentially patentable inventions conceived or first reduced to practice in whole or in part in the course of my University responsibilities or with more than incidental use of University resources. I hereby assign to Stanford all my right, title and interest in such patentable inventions and to execute and deliver all documents and do any and all things necessary and proper on my part to effect such assignment. (See Inventions, Patents and Licensing for further clarification and discussion related to this paragraph.)

2. I am free to place my inventions in the public domain as long as in so doing neither I nor Stanford violates the terms of any agreements that governed the work done.

3. Stanford policy states that all rights in copyright shall remain with the creator unless the work:
   a. is a work-for-hire (and copyright therefore vests in the University under copyright law),
   b. is supported by a direct allocation of funds through the University for the pursuit of a specific project,
   c. is commissioned by the University,
   d. makes significant use of University resources or personnel, or
   e. is otherwise subject to contractual obligations.

I hereby assign or confirm in writing to Stanford all my right, title and interest, including associated copyright, in and to copyrightable materials falling under a) through e), above.

4. I am now under no consulting or other obligations to any third person, organization or corporation in respect to rights in inventions or copyrightable materials which are, or could be reasonably construed to be, in conflict with this agreement.

NOTE: An alternative to this agreement may be appropriate for personnel with a prior existing and conflicting employment agreement that establishes a right to intellectual property in conflict with Stanford policies. Personnel in this situation should contact the Office of the Vice Provost and Dean of Research.

5. I will not enter into any agreement creating copyright or patent obligations in conflict with this agreement.

6. This agreement is effective on the later of July 1, 2011 (on the one hand) or my date of hire, enrollment, or participation in projects administered by Stanford (on the other hand), and is binding on me, my estate, heirs and assigns.

Electronic Signature in AXESS
http://axess.stanford.edu

The signer should make a copy of this agreement for his or her own records, and hereby waives any objection to Stanford’s use of an electronic version of this agreement as a substitute for the original for any legally recognized purpose.

July 2011

Provider: Office of the Vice Provost and Dean of Research, Stanford University
Contact: Assistant Dean of Research
Last updated: July 2011
Disclosure of Research Codes

Claim: Codes would (eventually) be fully open in the absence of Bayh-Dole:

• Grassroots “Reproducible Research” movement in computational science (policy development, best practices, tool development),

• Changes in funding agency requirements

• Changes in journal publication requirements
Best Practices for Computational Science: Software Infrastructure and Environments for Reproducible and Extensible Research

Victoria Stodden
Columbia University - Department of Statistics
Sheila Miguez
Columbia University
September 6, 2013

Abstract:
Scholarly dissemination and communication standards are changing to reflect the increasingly computational nature of scholarly research, primarily to include the sharing of the data and code associated with published results. This paper presents a formalized set of best practice recommendations for computational scientists wishing to disseminate reproducible research, facilitate innovation by enabling data and code re-use, and enable broader communication of the output of digital scientific research. We distinguish two forms of collaboration to motivate choices of software environment for computational scientific research. We also present these Best Practices as a living, evolving, and changing document on wiki.
Other Legal Barriers to Open Code

HIPAA (Health Information Portability and Accountability Act) and privacy regulations,

Copyright (i.e. Reproducible Research Standard),

Collaboration agreements with industry,

Hiring agreements, institutional rules,

National security.
Executive Memorandum: “Expanding Public Access to the Results of Federally Funded Research”

- “Access to digital data sets resulting from federally funded research allows companies to focus resources and efforts on understanding and exploiting discoveries.”

- “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.”

- “digital recorded factual material commonly accepted in the scientific community as necessary to validate research findings”

- “Each agency shall submit its draft plan to OSTP within six months of publication of this memorandum.”
Executive Order: “Making Open and Machine Readable the New Default for Government Information”

- “The Director … shall issue an Open Data Policy to advance the management of Government information as an asset”
- “Agencies shall implement the requirements of the Open Data Policy”
- “Within 30 days of the issuance of the Open Data Policy, the CIO and CTO shall publish an open online repository of tools and best practices”
Request for Input:
“Strategy for American Innovation”

• “to guide the Administration's efforts to promote lasting economic growth and competitiveness through policies that support transformative American innovation in products, processes, and services and spur new fundamental discoveries that in the long run lead to growing economic prosperity and rising living standards.”

• “(11) Given recent evidence of the irreproducibility of a surprising number of published scientific findings, how can the Federal Government leverage its role as a significant funder of scientific research to most effectively address the problem?”

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)
• **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.”