Toward a Reproducible Scholarly Record

Victoria Stodden
School of Information Sciences
University of Illinois at Urbana-Champaign

Dataverse Community Meeting 2017
Harvard University

June 16, 2017
Agenda

1. Imagining a Future of Radical Transparency
2. Parsing Reproducibility
3. Tools and “Cyberinfrastructure”
4. Artifact Availability
Querying the Scholarly Record

• Show a table of effect sizes and p-values in all phase-3 clinical trials for Melanoma published after 1994;

• Name all of the image denoising algorithms ever used to remove white noise from the famous “Barbara” image, with citations;

• List all of the classifiers applied to the famous acute lymphoblastic leukemia dataset, along with their type-1 and type-2 error rates;

• Create a unified dataset containing all published whole-genome sequences identified with mutation in the gene BRCA1;

• Randomly reassign treatment and control labels to cases in published clinical trial X and calculate effect size. Repeat many times and create a histogram of the effect sizes. Perform this for every clinical trial published in the year 2003 and list the trial name and histogram side by side.

Courtesy of Donoho and Gavish 2012
Parsing Reproducibility

“Empirical Reproducibility”

“Statistical Reproducibility”

“Computational Reproducibility”

V. Stodden, IMS Bulletin (2013)
Empirical Reproducibility

Sorting Out the FACS: A Devil in the Details

William C. Hines, Ying Su, Irene Kuhn, Kornelia Polyak, and Mina J. Bissell

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 Coast 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 specimens, our two laboratories’ reproducibility was unable to replicate each other’s fluorescence-activated cell sorting (FACS) profiles of primary breast cells. This 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 Ulz, 2011). As with any exponential increase in data complexity, breast reduction mammaplasties. 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

Reproducibility Issues in Research with Animals and Animal Models

The ability to reproduce an experiment is one important approach that scientists use to gain confidence in their conclusions. Studies that show that a number of significant peer-reviewed studies are not reproducible has alarmed the scientific community. Research that uses animals and animal models seems to be one of the most susceptible to reproducibility issues.

Evidence indicates that there are many factors that may be contributing to scientific irreproducibility, including insufficient reporting of details pertaining to study design and planning; inappropriate interpretation of results; and author, reviewer, and editor abstracted reports, assessing, and accepting studies for publication.

In this workshop, speakers from around the world will explore the many facets of the issue and potential pathways to reducing the problems. Audience participation portions of the workshop are designed to facilitate understanding of the issue.
Statistical Reproducibility

- False discovery, p-hacking (Simonsohn 2012), file drawer problem, overuse and mis-use of p-values, lack of multiple testing adjustments.
- Low power, poor experimental design, nonrandom sampling,
- Data preparation, treatment of outliers, re-combination of datasets, insufficient reporting/tracking practices,
- Inappropriate tests or models, model misspecification,
- Model robustness to parameter changes and data perturbations,
- ...
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.
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.
Really Reproducible Research

Inspired by Stanford Professor Jon Claerbout, from 1992:

“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

Note the difference between: reproducing the computational steps and, replicating the experiments independently including data collection and software implementation. (Both required)
RECOMMENDATION SIX: Through their policies and through the development of supporting infrastructure, research sponsors and science, engineering, technology, and medical journal and book publishers should ensure that information sufficient for a person knowledgeable about the field and its techniques to reproduce reported results is made available at the time of publication or as soon as possible after publication.

RECOMMENDATION SEVEN: Federal funding agencies and other research sponsors should allocate sufficient funds to enable the long-term storage, archiving, and access of datasets and code necessary for the replication of published findings.
REPRODUCIBILITY

Enhancing reproducibility for computational methods
Data, code, and workflows should be available and cited

By Victoria Stodden,¹ Marcia McNutt,² David H. Bailey,³ Ewa Deelman,⁴ Yolanda Gil,⁴ Brooks Hanson,⁵ Michael A. Heroux,⁶ John P.A. Ioannidis,⁷ Michela Taufer⁸

Over the past two decades, computational methods have radically changed the ability of researchers from all areas of scholarship to process and analyze data and to simulate complex systems. But with these advances come challenges that are contributing to broader concerns over irreproducibility in the scholarly literature, among them the lack of transparency in disclosure of computational methods. Current reporting methods are often uneven, incomplete, and still evolving. We present a novel set of Reproducibility Enhancement Principles (REP) targeting disclosure challenges involving computation. These recommendations, which build upon more general proposals from the Transparency and Openness Promotion (TOP) guidelines (1) and recommendations for field data (2), emerged from workshop discussions among funding agencies, publishers and journal editors, industry participants, and researchers representing the scientific, engineering, and social science communities. REP requires that the full computational environment that generated published findings be made available in open and trusted repositories, including all software, data, workflows, and detailed descriptions. This is accompanied by detailed metadata about the software and workflow.

Sufficient metadata should be provided for someone in the field to use the shared digital scholarly objects without resorting to contacting the original authors (i.e., http://bit.ly/2fVwjPH). Software metadata should include, at a minimum, the title, authors, version, language, license, Uniform Resource Identifier/DOI, software description (including purpose, inputs, outputs, dependencies), and execution requirements.

To enable credit for shared digital scholarly objects, citation should be standard practice. All data, code, and workflows, including software written by the authors, should be cited in the references section (10). We suggest that software citation include software version information and its unique identifier in addition to relevant metadata. By standardizing on these best practices, we can build a robust, trustworthy digital scholarly record that enables reproducible science in the 21st century.
Reproducibility Enhancement Principles

RECOMMENDATION 1: To facilitate reproducibility, share the data, software, workflows, and details of the computational environment in open repositories.

RECOMMENDATION 2: To enable discoverability, persistent links should appear in the published article and include a permanent identifier for data, code, and digital artifacts upon which the results depend.

RECOMMENDATION 3: To enable credit for shared digital scholarly objects, citation should be standard practice.

RECOMMENDATION 4: To facilitate reuse, adequately document digital scholarly artifacts.
Reproducibility Enhancement Principles

RECOMMENDATION 5: Journals should conduct a Reproducibility Check as part of the publication process and enact the TOP Standards at level 2 or 3.

RECOMMENDATION 6: Use Open Licensing when publishing digital scholarly objects.

RECOMMENDATION 7: To better enable reproducibility across the scientific enterprise, funding agencies should instigate new research programs and pilot studies.
## Summary of the eight standards and three levels of the TOP guidelines

Levels 1 to 3 are increasingly stringent for each standard. Level 0 offers a comparison that does not meet the standard.

<table>
<thead>
<tr>
<th>Standard Type</th>
<th>Level 0</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citation standards</td>
<td>Journal encourages citation of data, code, and materials—or says nothing.</td>
<td>Journal describes citation of data in guidelines to authors with clear rules and examples.</td>
<td>Article provides appropriate citation for data and materials used, consistent with journal’s author guidelines.</td>
<td>Article is not published until appropriate citation for data and materials is provided that follows journal’s author guidelines.</td>
</tr>
<tr>
<td>Data transparency</td>
<td>Journal encourages data sharing—or says nothing.</td>
<td>Article states whether data are available and, if so, where to access them.</td>
<td>Data must be posted to a trusted repository. Exceptions must be identified at article submission.</td>
<td>Data must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.</td>
</tr>
<tr>
<td>Analytic methods (code)</td>
<td>Journal encourages code sharing—or says nothing.</td>
<td>Article states whether code is available and, if so, where to access them.</td>
<td>Code must be posted to a trusted repository. Exceptions must be identified at article submission.</td>
<td>Code must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.</td>
</tr>
<tr>
<td>(code) transparency</td>
<td>Journal encourages code sharing—or says nothing</td>
<td>Article states whether materials are available and, if so, where to access them.</td>
<td>Materials must be posted to a trusted repository. Exceptions must be identified at article submission.</td>
<td>Materials must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.</td>
</tr>
<tr>
<td>Research materials transparency</td>
<td>Journal encourages materials sharing—or says nothing</td>
<td>Article articulates design transparency standards.</td>
<td>Journal requires adherence to design transparency standards for review and publication.</td>
<td>Journal requires and enforces adherence to design transparency standards for review and publication.</td>
</tr>
<tr>
<td>Design and analysis</td>
<td>Journal encourages design and analysis transparency or says nothing</td>
<td>Journal articulates design transparency standards.</td>
<td>Journal requires preregistration of studies and provides link in article and certification of meeting preregistration badge requirements.</td>
<td>Journal requires preregistration of studies and provides link and badge in article to meeting requirements.</td>
</tr>
<tr>
<td>transparency</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preregistration of studies</td>
<td>Journal says nothing.</td>
<td>Journal encourages preregistration of studies and provides link in article to preregistration if it exists.</td>
<td>Journal requires preregistration of studies and provides link in article and certification of meeting preregistration badge requirements.</td>
<td>Journal requires preregistration of studies and provides link and badge in article to meeting requirements.</td>
</tr>
<tr>
<td>Preregistration of analysis</td>
<td>Journal says nothing.</td>
<td>Journal encourages preanalysis plans and provides link in article to registered analysis plan if it exists.</td>
<td>Journal requires preregistration of studies with analysis plans and provides link and badge in article to meeting requirements.</td>
<td>Journal requires preregistration of studies with analysis plans and provides link and badge in article to meeting requirements.</td>
</tr>
<tr>
<td>plans</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Replication</td>
<td>Journal discourages submission of replication studies—or says nothing.</td>
<td>Journal encourages submission of replication studies.</td>
<td>Journal encourages submission of replication studies and conducts blind review of results.</td>
<td>Journal uses Registered Reports as a submission option for replication studies with peer review before observing the study outcomes.</td>
</tr>
</tbody>
</table>
Actual question..

“[My Federal agency] is struggling with a lot of the same questions as the broader community around reproducible science. Are there particular groups that you would recommend we follow to keep track of progress that is being made?”
## Infrastructure Solutions

### Research Environments

- **Verifiable Computational Research**
  - knitR
- **Collage Authoring Environment**
  - Sumatra
  - Galaxy
- **SHARE**
  - Sweave
- **Code Ocean**
  - Cyverse
- **Jupyter**
  - NanoHUB
- **Collage Authoring Environment**
  - GenePattern
  - torch.ch
- **Open Science Framework**
  - IPOL
- **Vistrails**
  - Popper
- **ResearchCompendia.org**
- **DataCenterHub**
- **RunMyCode.org**
- **ChameleonCloud**
- **Occam**
- **Wavelab**
- **RCloud**
- **Sparselab**
- **TheDataHub.org**
- **ChameleonCloud**
- **Madagascar**

### Workflow Systems

- **Taverna**
- **Wings**
- **Pegasus**
- **CDE**
- **binder.org**
- **Kurator**
- **Kepler**
- **Everware**
- **Reprozip**
- **TheDataHub.org**
- **Whole Tale**
- **Cyverse**
- **GenePattern**
- **torch.ch**
- **Jupyter**
- **NanoHUB**
- **Vistrails**
- **Popper**
- **flywheel.io**

### Dissemination Platforms

- **ResearchCompendia.org**
- **DataCenterHub**
- **RunMyCode.org**
- **ChameleonCloud**
- **Occam**
- **Wavelab**
- **RCloud**
- **Sparselab**
- **TheDataHub.org**
- **ChameleonCloud**
- **Madagascar**
Looking Ahead

We see the convergence of two (ordinarily antagonistic) trends:

- Scientific projects will become massively more computing intensive
- Research computing will become dramatically more transparent
“Experiment Definition Systems”

- Create “Experiment Definition Systems” to (easily) manage the conduct of massive computational experiments and expose the resulting data for analysis and structure the subsequent data analysis.

- Address the two trends simultaneously: better transparency will allow people to run much more ambitious computational experiments. And better computational experiment infrastructure will allow researchers to be more transparent.
Three Principles for Cyberinfrastructure

1. **Supporting scientific norms**—not only should CI enable new discoveries, but it should also permit others to reproduce the computational findings, reuse and combine digital outputs such as datasets and code, and facilitate validation and comparisons to previous findings.

2. **Supporting best practices in science**—CI in support of science should embed and encourage best practices in scientific research and discovery.

3. **Taking a holistic approach to CI**—the complete end-to-end research pipeline should be considered to ensure interoperability and the effective implementation of 1 and 2.

The Whole Tale project seeks to leverage & contribute to existing cyberinfrastructure and tools to support the whole research story, and provide access to data and computing power.

Integrate tools to simplify usage and promote best practices

The Whole Tale
Merging Science and Cyberinfrastructure Pathways

Whole Tale will enable researchers to examine, transform, and then seamlessly republish research data that was used in an article. As a result, these "living articles" enable new discovery by allowing researchers to construct representations and syntheses of data.

NSF CC*DNI DIBBS awarded 2016: 5 Institutions for 5 Years ($5M total)
Whole Tale Project Goals

Expose existing digital resources to researchers
... through popular frontends (Jupyter, RStudio, ..)

Develop necessary “software glue”
... for seamless access to different CI-backend capabilities

Enhance conceptualization-to-publication lifecycle
... by empowering scientists to create computational narratives in their usual programming environments

Embed reproducibility and best/better practices in the digital research environment
Whole Tale: What’s in a Name?

(1) Whole Tale ⇔ Whole Story:
   Support (computational & data) scientists along the complete research lifecycle from experiment to publication and back!

(2) Whole Tale ⇔ Long Tail of Science

image from Ferguson et al. 2014 doi:10.1038/nn.3838
“Tales” are the final research output from a project, capturing the complete provenance of a particular activity within the system:

- capture full provenance of an analysis recorded transparently,
- easily sharable with others,
- publishable in repositories,
- associated with persistent identifiers,
- linked to publications,
- execute in the same state as it was when first published,
- acts as a starting point for research.
“the paper of the birch”

Researchers today are perhaps more like Lewis and Clark..

Jefferson’s instructions (1803):

“The object of your mission is to explore the Missouri river, & such principal stream of it, as, by its course & communication with the water of the Pacific ocean may offer the most direct & practicable water communication across this continent, for the purposes of commerce.”

Thanks to Dave Culler for the analogy
“Beginning at the mouth of the Missouri, you will take observations of latitude and longitude at all remarkable points on the river”

“The courses of the river between these points of observation may be supplied by the compass, the log-line & by time, corrected by the observations themselves. The variations of the compass too, in different places should be noticed”
“Your observations are to be taken with great pains & accuracy to be entered distinctly, & intelligibly for others as well as yourself, to comprehend all the elements necessary, with the aid of the usual tables to fix the latitude & longitude of the places at which they were taken, & are to be rendered to the war office, for the purpose of having the calculations made concurrently by proper persons within the U.S. Several copies of these as well as of your other notes, should be made at leisure times, & put into the care of the most trustworthy of your attendants, to guard by multiplying them against the accidental losses to which they will be exposed. A further guard would be that one of these copies be written on the paper of the birch, as less liable to injury from damp than common paper.”
“To provide, on the accident of your death, against anarchy, dispersion & the consequent danger to your party, and total failure of the enterprise, you are hereby authorised, by any instrument signed & written in your own hand, to name the person among them who shall succeed to the command”
Artifact Availability
Background: Open Source Software

• Innovation: Open Licensing
  ➡ 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
The Reproducible Research Standard

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.
“ezDMP”

NSF funded project to provide structured guidance for a second generation data management plan.


Helen M. Berman, Rutgers
Kerstin Lehnert, Columbia
Victoria Stodden, UIUC
Maggie Gabanyi, Rutgers
Vicki Ferrini, Columbia
exDMP Progress

- Examined selected data management plans to understand gaps, successes, and patterns of use in IEDA DMP Tool
- Reviewed the patterns exhibited by DMP creators using the IEDA DMP Tool
- GEO (under the lead of K. Lehnert & V. Ferrini)
- BIO (under the lead of H. Berman)
- SBE, MPS (under the lead of V. Stodden)
- Implement into IEDA ("ezDMP")
Open Questions

• Who funds and supports cyberinfrastructure?

• Who owns data, code, and research outputs?

• Who controls access and gateways?

• What are community standards around documentation, citation standards, best practices? Who enforces?

• Citation of cyberinfrastructure? What are the incentives? What should they be?
Currently there is a distribution of largely unconnected scholarly objects in various repositories, with different ownership structures.

Some repositories are institutional or federally funded, some are owned by publishers e.g. figshare, Mendeley.
Conclusions

Proactive design of a reproducible scholarly record.

• identification and linking of digital scholarly objects to the claims they support.

• facilitating use and re-use: discoverability and intellectual property.
Community Responses

Declarations and Documents:

- Yale Declaration 2009
- ICERM 2012
- XSEDE 2014
Government Mandates

- OSTP 2013 Open Data and Open Access Executive Memorandum; Executive Order.
- “Public Access to Results of NSF-Funded Research”
- NOAA Data Management Plan, Data Sharing Plan
- NIST “Common Access Platform”
- …
Federal Agencies

Reliable Science: The Path to Robust Research Results

September 8, 2015

These days, much discussion about the reproducibility of scientific results seems driven by critiques of research in biomedicine and psychology. Most recently, an article [1] in *Science* concluded that 60 percent of a collection of studies were not replicable. This result along with similar analyses of cancer research results have stimulated strong commentary. For example, the New York Times print edition headline about the Science article was “Psychology’s Fears Confirmed: Rechecked Studies Don’t Hold Up,” coverage that prompted a strong op-ed rebuttal titled, "Psychology Is Not In Crisis."

Issues that arise with human subjects or with other complex living systems do not plague physical science to the same degree. However, the notion of measuring the same value of a physical quantity or the same behavior of a physical system in different laboratories at different times is central to our concept of a valid scientific result. Often the approach is not simply to replicate an experiment, but rather to get at the same quantity via different paths. For example, we can measure the gravitational constant, \( G \), with approaches ranging from a torsional pendulum to atom interferometry.

Two of the cornerstones of science advancement are rigor in designing and performing scientific research and the ability to reproduce biomedical research findings. The application of rigor ensures robust and unbiased experimental design, methodology, analysis, interpretation, and reporting of results. When a result can be reproduced by multiple scientists, it validates the original results and readiness to progress to the next phase of research. This is especially important for clinical trials in humans, which are built on studies that have demonstrated a particular effect or outcome.

In recent years, however, there has been a growing awareness of the need for rigorously designed published preclinical studies, to ensure that such studies can be reproduced. This webpage provides information about the efforts underway by NIH to enhance rigor and reproducibility in scientific research.

Rigor and Reproducibility

Enhancing reproducibility through rigor and transparency: the information provided on this website is designed to assist the extramural community in addressing rigor and reproducibility in grant applications due on January 25, 2016, and beyond.

On This Page:
- News
- Goals

National Institutes of Health
Office of Extramural Research

RIGOR AND REPRODUCIBILITY

Two of the cornerstones of science advancement are rigor in designing and performing scientific research and the ability to reproduce biomedical research findings. The application of rigor ensures robust and unbiased experimental design, methodology, analysis, interpretation, and reporting of results. When a result can be reproduced by multiple scientists, it validates the original results and readiness to progress to the next phase of research. This is especially important for clinical trials in humans, which are built on studies that have demonstrated a particular effect or outcome.

In recent years, however, there has been a growing awareness of the need for rigorously designed published preclinical studies, to ensure that such studies can be reproduced. This webpage provides information about the efforts underway by NIH to enhance rigor and reproducibility in scientific research.
Journal Requirements

• Science: code data sharing since 2011.

• Nature: data sharing.

• ...

Merton’s Scientific Norms (1942)

**Communalism**: scientific results are the common property of the community.

**Universalism**: all scientists can contribute to science regardless of race, nationality, culture, or gender.

**Disinterestedness**: act for the benefit of a common scientific enterprise, rather than for personal gain.

**Originality**: scientific claims contribute something new.

**Skepticism**: scientific claims must be exposed to critical scrutiny before being accepted.
Skepticism -> Reproducibility

- Skepticism requires that the claim can be independently verified,
- This in turn requires transparency in the communication of the research process.
- Instantiated by Robert Boyle and the Transactions of the Royal Society in the 1660’s.