Open Data, Code, and Computational Reproducibility

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

CMU Open Science Symposium
October 18, 2018
Agenda

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

“Empirical Reproducibility”

“Statistical Reproducibility”

“Computational Reproducibility”

V. Stodden, IMS Bulletin (2013)
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)
Enhancing reproducibility for computational methods

Data, code, and workflows should be available and cited

By Victoria Stodden, Marcia McNutt, David H. Bailey, Ewa Deelman, Yolanda Gil, Brooks Hanson, Michael A. Heroux, John P.A. Ioannidis, Michela Taufer

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 disciplines spanning computer science, computational science, and data science.

To understanding how computational results were derived and to reconciling any differences that might arise between independent replications (4). We thus focus on the ability to rerun the same computational steps on the same data the original authors used as a minimum dissemination standard (5, 6), which includes workflow information that explains what raw data and intermediate results are input to which computations (7). Access to the data and code that underlie discoveries can also enable downstream scientific contributions, such as meta-analyses, reuse, and other efforts that include results from multiple studies.

RECOMMENDATIONS

Share data, software, workflows, and details of the computational environment that generate published findings in open trusted repositories. The minimal components that enable independent regeneration of computational results are the data, the computational steps that produced the findings, and the workflow describing how to generate the results using the data and code, including parameter settings, random number seeds, make files, or other parameters.

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/2fWjPH). 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 other citation details.
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.
# Infrastructure Solutions

## Research Environments and Document Enhancement Tools

<table>
<thead>
<tr>
<th>StatTag.org</th>
<th>SHARE</th>
<th>Code Ocean</th>
<th>Jupyter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verifiable Computational Research</td>
<td>Sweave</td>
<td>Cyverse</td>
<td>NanoHUB</td>
</tr>
<tr>
<td>knitR</td>
<td>SOLE</td>
<td>Open Science Framework</td>
<td>Vistrails</td>
</tr>
<tr>
<td>Collage Authoring Environment</td>
<td>GenePattern</td>
<td>IPOL</td>
<td>Popper</td>
</tr>
<tr>
<td>Sumatra</td>
<td>torch.ch</td>
<td>Whole Tale</td>
<td>flywheel.io</td>
</tr>
</tbody>
</table>

## Workflow Systems

<table>
<thead>
<tr>
<th>Taverna</th>
<th>Wings</th>
<th>Pegasus</th>
<th>CDE</th>
<th>binder.org</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kurator</td>
<td>Kepler</td>
<td>Everware</td>
<td>Reprozip</td>
<td>Galaxy</td>
</tr>
</tbody>
</table>

## Dissemination Platforms

<table>
<thead>
<tr>
<th>ResearchCompendia.org</th>
<th>DataCenterHub</th>
<th>RunMyCode.org</th>
<th>ChameleonCloud</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occam</td>
<td>RCloud</td>
<td>TheDataHub.org</td>
<td>Madagascar</td>
</tr>
<tr>
<td>Wavelab</td>
<td>Sparselab</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
“Whole Tale” Project

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: 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**:

**Engage** researchers of all project scales

---

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

- 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.
Try it!

The first Whole Tale platform was released in July!

http://wholetale.readthedocs.io/users_guide/

Feedback is very welcome at feedback@wholetale.org and/or at https://github.com/whole-tale/whole-tale/issues
“ezDMP”

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


Helen M. Berman (Rutgers)
Kerstin Lehnert (Columbia)
Vicki Ferrini (Columbia)
Victoria Stodden (UIUC)
Maggie Gabanyi (Rutgers)
ezDMP Released!

Research progression:

- 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.

- Implemented into IEDA DMP Tool (“ezDMP”)

Try our prototype! [http://dev.ezdmp.org](http://dev.ezdmp.org) and we have a feedback rubric here [https://goo.gl/forms/CaEB3ddJ3iuUmpxS2](https://goo.gl/forms/CaEB3ddJ3iuUmpxS2)
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.
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?
Conclusions

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

- Scientific projects will become **massively more computing intensive**
- Research computing will become **dramatically more transparent**

These are *reinforcing* trends, whose resolution is essential for verifying and comparing findings.

The future is the *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.
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></th>
<th>Level 0</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Citation standards</strong></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><strong>Data transparency</strong></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><strong>Analytic methods (code) transparency</strong></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><strong>Research materials transparency</strong></td>
<td>Journal encourages materials 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><strong>Design and analysis transparency</strong></td>
<td>Journal encourages design and analysis transparency or says nothing.</td>
<td>Journal 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><strong>Preregistration of studies</strong></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 encourages 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><strong>Preregistration of analysis plans</strong></td>
<td>Journal says nothing.</td>
<td>Journal encourages preregistration of plans and provides link in article to registered analysis plans if it exists.</td>
<td>Journal encourages preregistration of plans and provides link in article and certification of meeting registered analysis plan badge 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><strong>Replication</strong></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>
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 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 op-ed edition had a column about the Science article was “Psychology’s Fears Confirmed: Replicated Studies Don’t Hold Up,” which ran in the New York Times on October 21, 2015.

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 in 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 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.

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.
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.