Computational Transparency: A Key Part of Reproducibility

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

Conceptualizing a Geospatial Software Institute (GSI)
Workshop 3: Strategic Plan and Governance of GSI
July 14-16, 2019
SESYNC and the Westin Annapolis, Annapolis, MD

Agenda

1. Computational Reproducibility: Impossible without Software Transparency?
2. Policy: 2016 REPS Recommendations
3. Policy: 2019 National Academies Report: “Reproducibility and Replicability in Science”

Assertion: Without transparency regarding the computational steps / software that generated results, we don’t have a reproducible scholarly record.

REPRODUCIBILITY

Enhancing reproducibility for computational methods

Data, code, and workflows should be available and cited

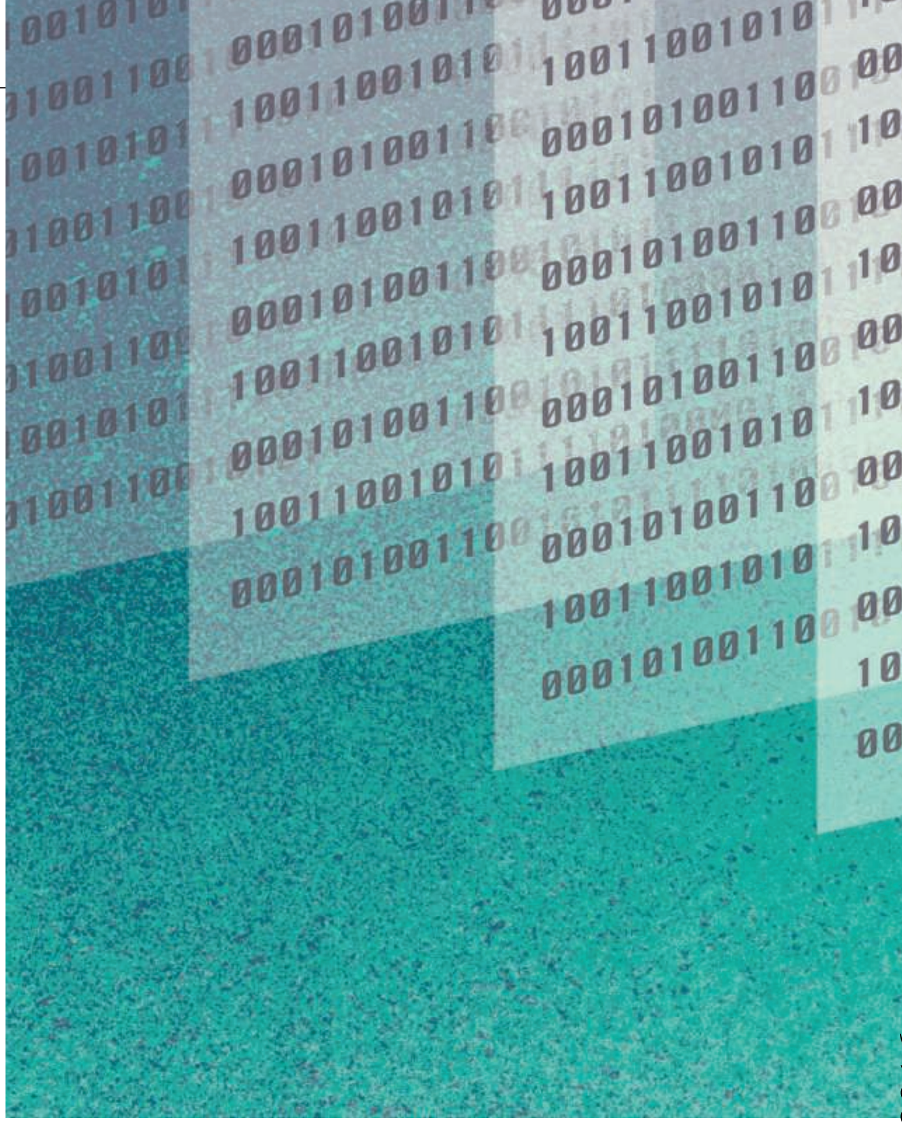
By **Victoria Stodden,[1] Marcia McNutt,[2] David H. Bailey,[3] Ewa Deelman,[4] Yolanda Gil,[4] Brooks Hanson,[5] Michael A. Heroux,[6] John P.A. Ioannidis,[7] Michela Taufer[8]**

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 repre-

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



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 addi-

Workshop Recommendations: “Reproducibility Enhancement Principles”

1. Share data, software, workflows, and details of the computational environment that generate published findings in open trusted repositories.

2. Persistent links should appear in the published article and include a permanent identifier for data, code, and digital artifacts upon which the results depend.

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

4. To facilitate reuse, adequately document digital scholarly artifacts.

Workshop Recommendations: "Reproducibility Enhancement Principles"

5. Use Open Licensing when publishing digital scholarly objects.

6. Journals should conduct a reproducibility check as part of the publication process and should enact the TOP standards at level 2 or 3.

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.

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

National Academies of Science, Engineering, and Medicine Consensus Report (2019): “Reproducibility and Replicability in Science”

Definitions

Reproducibility is obtaining ***consistent results using the same input data, computational steps, methods, and code, and conditions of analysis***. This definition is synonymous with "computational reproducibility," and the terms are used interchangeably in this report.

Replicability is obtaining ***consistent results across studies aimed at answering the same scientific question***, each of which has obtained its own data. Two studies may be considered to have replicated if they obtain consistent results given the level of uncertainty inherent in the system under study.

Key Recommendation

RECOMMENDATION 4-1: To help ensure the reproducibility of computational results, ***researchers should convey clear, specific, and complete information about any computational methods and data products that support their published results*** in order to enable other researchers to repeat the analysis, unless such information is restricted by non-public data policies. That information should include the data, study methods, and computational environment:

- *the input data* used in the study either in extension (e.g., a text file or a binary) or in intension (e.g., a script to generate the data), as well as intermediate results and output data for steps that are nondeterministic and cannot be reproduced in principle;
- *a detailed description of the study methods (ideally in executable form)* together with its computational steps and associated parameters; and
- *information about the computational environment* where the study was originally executed, such as operating system, hardware architecture, and library dependencies (which are relationships described in and managed by a software dependency manager tool to mitigate problems that occur when installed software packages have dependencies on specific versions of other software packages).

Key Recommendation

RECOMMENDATION 6-3:

Funding agencies and organizations should consider investing in research and development of open-source, usable tools and infrastructure that support reproducibility for a broad range of studies across different domains in a seamless fashion. Concurrently, investments would be helpful in outreach to inform and train researchers on best practices and how to use these tools.

Key Recommendation

RECOMMENDATION 6-5: In order to facilitate the transparent sharing and availability of digital artifacts, such as data and code, for its studies, the National Science Foundation (NSF) should:

- Develop a set of ***criteria for trusted open repositories*** to be used by the scientific community for objects of the scholarly record.
- Seek to ***harmonize with other funding agencies*** the repository criteria and data-management plans for scholarly objects.
- ***Endorse or consider creating code and data repositories*** *for long-term archiving and preservation of digital artifacts that support claims made in the scholarly record based on NSF-funded research*. These archives could be based at the institutional level or be part of, and harmonized with, the NSF-funded Public Access Repository.
- *Consider* ***extending NSF's current data-management plan to include other digital artifacts, such as software****.*
- Work with communities reliant on non-public data or code to ***develop alternative mechanisms*** **for demonstrating reproducibility**. Through these repository criteria, NSF would enable discoverability and standards for digital scholarly objects and discourage an undue proliferation of repositories, perhaps through endorsing or providing one go-to website that could access NSF-approved repositories.

Key Recommendation

RECOMMENDATION 6-6: ***Many stakeholders have a role to play*** in improving computational reproducibility, including educational institutions, professional societies, researchers, and funders.

- *Educational institutions* should educate and train students and faculty about computational methods and tools to improve the quality of data and code and to produce reproducible research.
- *Professional societies* should take responsibility for educating the public and their professional members about the importance and limitations of computational research. Societies have an important role in educating the public about the evolving nature of science and the tools and methods that are used.
- *Researchers* should collaborate with expert colleagues when their education and training are not adequate to meet the computational requirements of their research.
- In line with its priority for "harnessing the data revolution," the *National Science Foundation* (and other funders) should consider funding of activities to promote computational reproducibility.

Key Recommendation

RECOMMENDATION 6-9:

Funders should require a thoughtful discussion in *grant applications of* ***how uncertainties will be evaluated****, along with any relevant issues regarding replicability and computational reproducibility.*

Funders should introduce ***review of reproducibility and replicability guidelines and activities*** *into their merit-review criteria*, as a low-cost way to enhance both.

Conclusion

Reproducibility is now widely recognized as a pressing issue requiring a multi-faceted solution, of which computational infrastructure and software is one key part.

Infrastructure supporting transparency and reproducibility will be used not out of hygiene or ethics, but because it enables increasingly ambitious computational research.

A Convergence of Trends

- Scientific projects will become massively more computing intensive, and

- Scientific computing will become dramatically more transparent

Simultaneity: better transparency allows much more ambitious computational experiments. *And* better computational experiment infrastructure allows greater transparency.

Such a system is used not out of ethics or hygiene, but because this is a corollary of managing massive amounts of computational work, enabling *efficiency* and *productivity*, and *discovery*.

Legal Issues in Software

Intellectual property is associated with software (and all digital scholarly objects) e.g the U.S. Constitution and subsequent Acts:

"*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)

Copyright

- 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
- limited time: generally life of the author +70 years
- Exceptions and Limitations: e.g. Fair Use.

Patents

Patentable subject matter: "*new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof*" (35 U.S.C. §101) that is

1. *Novel*, in at least one aspect,
2. *Non-obvious*,
3. *Useful*.

USPTO Final Computer Related Examination Guidelines (1996) "A practical application of a computer-related invention is statutory subject matter. This requirement can be discerned from the variously phrased prohibitions against the patenting of abstract ideas, laws of nature or natural phenomena" (see e.g. Bilski v. Kappos, 561 U.S. 593 (2010)).

Bayh-Dole Act (1980)

- Promote the transfer of academic discoveries for commercial development, via licensing of patents (ie. Technology Transfer Offices), and harmonize federal funding agency grant intellectual property regs.

- Bayh-Dole 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.

- Hence, institutions such as universities charged with utilizing the patent system for technology transfer.

Legal Issues in Data

- 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)).

- Copyright adheres to raw facts in Europe.

- the possibility of a residual copyright in data (attribution licensing or public domain certification).

- Legal mismatch: What constitutes a “raw” fact anyway?

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 **MIT License** or similar,
- Release data to public domain (**CC0**) or attach attribution license.

➡ *Remove copyright's barrier to reproducible research and,*

➡ *Realign the IP framework with longstanding scientific norms.*

Infrastructure Solutions

Research Environments and Document Enhancement Tools

StatTag.org	SHARE	Code Ocean	Jupyter
Verifiable Computational Research	Sweave	Cyverse	NanoHUB
knitR	SOLE	Open Science Framework	Vistrails
Collage Authoring Environment	GenePattern	IPOL	Popper
Sumatra	torch.ch	**Whole Tale**	flywheel.io

Workflow Systems

Taverna	Wings	Pegasus	CDE	binder.org
Kurator	Kepler	Everware	Reprozip	Galaxy

Dissemination Platforms

ResearchCompendia.org	DataCenterHub	RunMyCode.org	ChameleonCloud
Occam	RCloud	TheDataHub.org	Madagascar
Wavelab	Sparselab		

“Quantitative Programming Environments”

- Define and create “Quantitative Programming Environments” to (easily) manage the conduct of massive computational experiments and expose the resulting data for analysis and structure the subsequent data analysis

- The two trends need to be addressed 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.