Enabling the Verification of Computational Results: An Empirical Evaluation of Computational Reproducibility

Victoria Stodden, Matthew S. Krafczyk, Adhithya Bhaskar
University of Illinois at Urbana-Champaign

https://github.com/ReproducibilityInPublishing/P-RECS-2018-Enabling-Verification

Introduction

We identify barriers and outline solutions to the dissemination of "really reproducible research," which means including with publications complete author-provided information that enables transparency and reproducibility for computational and data-enabled claims [Buckheit and Donoho 1995; Claerbout 1994; Donoho et al. 2009; Schwab et al. 2000]. We use the Claerbout definition of reproducibility, "computational reproducibility" [Stodden et al. 2013b], which refers to the verification of the computational steps, including input data, parameters, and other information, that generated the computational claims presented in the associated article.

Our study evaluates the sufficiency of information regarding availability of digital artifacts such as data and code, and seeks to procure artifacts from authors if they are not accessible via the article alone. We evaluate the ease at which we can then regenerate the associated published claims using the author artifacts.

The aim of this work is to better understand author and community needs regarding artifact availability to enable computational reproducibility.

Methods

1. Inspect and Classify

We chose a sample size of 300 and we started with Issue 322 of the Journal of Computational Physics and collected articles through Issue 331. Articles were inspected and classified according to how much information they disclosed about their code and data. The Table 'Classification of Articles' shows our collected data.

2. Request Code and Data

Articles which did not contain enough code and data for replication (all but 6), were sent a request by email (IRB #17329).

3. Evaluate Code and Data

We applied evaluation criteria to the 306 articles to assess the following categories:

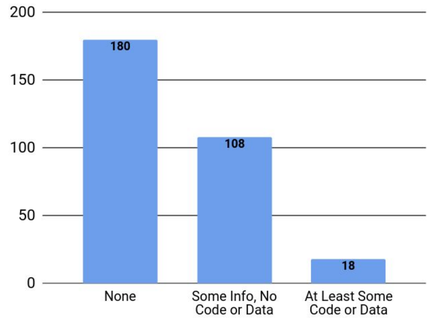
1. the level of information in the article enabling computational reproducibility;
2. the level of information provided on computational artifacts such as data and code that support the claims made in the article;
3. the facility at which that information and artifacts enabled the regeneration of the computational results in the article.

We used the evaluation criteria from "Best Practices for Publishing Research" [Bailey et al. 2013; Stodden et al. 2013a].

Results

Classification of Articles

- 306 articles with results based on code were inspected from volumes 322–331 of JCP
- 180 or 59% of articles gave no information about the code they used to get their results
- 108 or 35% of articles gave some information about their implementation such as library names, coding language, or hardware they used, but no actual co de
- 18 or 6% of articles gave or indicated they would give at least partial source code



Artifact Evaluation (n=55)

A precise statement of assertions to be made in the paper	100%
Full statement (or valid summary) of experimental results	100%
Salient details of data reduction & statistical analysis methods	73%
Necessary run parameters were given	86%
A statement of the computational approach, and why it rigorously tests the hypothesized assertions	100%
Complete statements of, or references to, algorithms and salient software details	63%
Discussion of the adequacy of parameters such as precision level and grid resolution	76%
Proper citation of all code and data used, including that generated by the authors	4%
Availability of computer code, input and output data, with some reasonable level of documentation	4%
Avenues of exploration examined throughout development, including negative findings	0%
Instructions for repeating computational experiments described in the article	79%
Precise functions were given, with settings	11%
Salient details of the test environment e.g. hardware, system software, and number of processors used	24%

Data documented to clearly explain what each part represents	40%
Data archived with significant longevity expected	27%
Data location provided in the acknowledgements	13%
Authors have documented use and licensing rights	29%
Software documented well enough to run it and what it ought to do	71%
The code is publicly available with no download requirements	27%
There was some method to track software changes, and some persistence of archiving	20%

Reproducibility Evaluation (n=55)

For the 55 articles with artifacts, we fully replicated none; partially replicated 32.7% (18); ran 54.5% (30); were able to build 3.6% (2); and had no progress on 9.1%.

Straightforward to reproduce with minimal effort	0%
Minor difficulty in reproducing	0%
Reproducible after some tweaking	9.1%
Could reproduce with fairly substantial skill and knowledge	16.4%
Reproducible with substantial intellectual effort	12.7%
Reproducible with substantial tedious effort	3.6%
Difficult to reproduce because of unavoidable inherent complexity	3.6%
Nearly impossible to reproduce	3.6%
Impossible to reproduce	50.9%

Recommendations

1. Community standards for documentation of computational artifacts.
2. Journals improve the communication of standards.
3. Appropriate use of open licensing for computational artifacts.
4. Cultural change and improved cyberinfrastructure for reporting negative results.
5. Greater investments in reproducibility research and cyberinfrasturcture and tools that support computational reproducibility.
6. Improved training in computational reproducibility practices.

Literature Cited

1. David H Bailey, Jonathan Borwein, and Victoria Stodden. 2013. Set the Default to 'Open'. Notices of the AMS (2013).
2. J. Buckheit and D. Donoho. 1995. WaveLab Architecture. Technical Report. Stanford University, http://www-stat.stanford.edu/wavelab.
3. J. Claerbout. 1994. Hypertext Documents about Reproducible Research. Technical Report. Stanford University, http://sepwww.stanford.edu.
4. David L Donoho, Arian Maleki, Inam Ur Rahman, Morteza Shahram, and Victoria Stodden. 2009. Reproducible research in computational harmonic analysis. Computing in Science & Engineering 11, 1 (2009), 8–18.
5. M. Schwab, N. Karrenbach, and J. Claerbout. 2000. Making scientific computations reproducible. Computing in Science & Engineering 2, 6 (2000), 61–67.
6. Victoria Stodden, Jonathan Borwein, and David H Bailey. 2013a. 'Setting the Default to Reproducible' in Computational Science Research. SIAM News (2013).
7. Victoria Stodden, Peixuan Guo, and Zhaokun Ma. 2013b. Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals. PloS one 8, 6 (2013), e67111.

Acknowledgements

We thank David Wong, Yantong Zhang, and Alex Dickinson for outstanding research assistance. We acknowledge support from NSF Award ACI-1659702 and NCSA SPIN.