A Future of Research Transparency: Enabling Reproducibility in Repository Design

Victoria Stodden
University of Illinois at Urbana-Champaign
victoria@stodden.net

2019

INTRODUCTION

The primary scientific rational for access to data and computational methods is to enable the verification and validation of published research findings (Donoho et al. 2009). Federal mandates and laws as well as ethical considerations regarding data and code disclosure, privacy, confidentiality, and ownership influence the ability of researchers to produce and share really reproducible research. This chapter addresses reproducibility and replication in the research context, then surfaces issues regarding reproducible computational research from the perspective of repository design. Some of these issues are extensions of traditional library concepts to the digital scientific research context (such as unique identifiers for artifacts), and other issues arise *de novo* and present new challenges to the repository community (such as the intimate and persistent linking based on artifact type). The novel scientific and research opportunities due to social network data and other forms of data digitization are readily apparent (Lazer et al. 2009) and with the reasoning recorded in the software scripts and code that analyze these data, new opportunities and responsibilities arise for sharing of research code and data. My point of departure for this discussion is the Principle of Scientific Data and Code Sharing (Stodden 2014a):

> **Principle of Scientific Data and Code Sharing**: Access to the data and methods associated with published scientific results should be maximized, only subject to clearly articulated restrictions interpreted in the most minimally restrictive way: such as privacy or confidentiality concerns; legal barriers including intellectual property or HIPAA regulations; or technological and cost constraints.

This principle is often implemented as "Default to Open," meaning that the digital scholarly objects that support research findings are made openly available in repository designed to be responsive to scientific reproducibility concerns (Claerbout and Karrenbach 1992; Buckheit and Donoho 1995;

Schwab, Karrenbach, and Claerbout 2000) and compelling reasons are necessary for restricting access to such artifacts (Bailey, Borwein, and Stodden 2013; Stodden, Borwein, and Bailey 2013). The Principle of Scientific Data and Code Sharing implies levels of access to artifacts that cannot be made openly available, due to legal or technological constraints for example. Such level could arise from the need to authorize access, perhaps via Institutional Review Board. Whether due to privacy concerns, technological barriers or another sources, restrictions on data and code availability do not necessarily imply absolute barriers. The repository design aspects addressed in this chapter from a reproducibility perspective include versioning and unique identifiers; artifact persistence; connecting code and results to data; citation; interoperability with other datasets, software, and repositories; acknowledging funding sources; providing appropriate metadata and documentation including that of the data generation mechanism; and licensing and terms of use. We also point to literature that evaluates the impact of repositories of social and economic data (Charles Beagrie Ltd and Centre for Strategic Economic Studies 2012) and encourage the creation of appropriate evaluation metrics for ISSOD.

Finally, the chapter explores how repository design can contribute to scientific integrity and discovery in a future of dramatically increased scientific transparency, including a discussion of certification of findings as reproducible. This chapter engages with the scientific research vision of ISSOD, and with community survey aspects as those produce research output.

REALLY REPRODUCIBLE RESEARCH AND THE RESEARCH COMPENDIA

The phrase "really reproducible research" refers to a concept first articulated by Jon Claerbout, a Stanford geophysicist, in approximately 1992 (Claerbout and Karrenbach 1992). The idea is that computational results are published with sufficient information that they can be destroyed and rebuilt from the information published with the result alone. This approach is described as follows in the Stanford Exploration Project (SEP) thesis template[1] :

> The markings [on each computational finding] ER, CR, and NR are promises by the author about the reproducibility of each [computational] result. Reproducibility is a way of organizing computational research that allows both the author and the reader of a publication to verify the reported results. Reproducibility facilitates the transfer of knowledge within SEP and outside of SEP.

- **ER** denotes **Easily Reproducible** and are the results of processing described in the paper. The author claims that you can reproduce such a figure from the programs, parameters, and makefiles included in the electronic document. The data must either be included in the electronic distribution, be easily available to all researchers ... Before the publication of the electronic document, someone other than the author tests the author's claim by destroying and rebuilding all ER [results]. Some ER [results] may not be reproducible by outsiders because they depend on data sets that are too large to distribute, or data that we do not have permission to redistribute but are in the SEP data library, or that the rules depend on commercial packages such as Matlab or Mathematica.
- **CR** denotes **Conditional Reproducibility**. The author certifies that the commands are in place to reproduce the [result] if certain resources are available. The primary reasons for the CR designation is that the data are not in the SEP library or the processing requires 20 minutes or more.

[1] See http://sepwww.stanford.edu/doku.php?id=sep:research:theses (last accessed October 8, 2018).

- **NR** denotes **Non-Reproducible**. SEP discourages authors from flagging their [results] as NR except for those that are used solely for motivation, comparison, or illustration of the theory, such as: artist drawings, scannings, or figures taken from SEP reports not by the authors or from non-SEP publications.

Since then this approach has been adopted and developed in computational research communities (Bailey, Borwein, and Stodden 2013; Stodden, Borwein, and Bailey 2013). Concurrently efforts emerged in the Political Science community to ensure computational reproducibility, where *The Replication Standard* was introduced: "The replication standard holds that sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party could replicate the results without any additional information from the author" (King 1995). Note the use of the word replication instead of reproducibility and the focus on inclusion of more information than what is needed for reproducing computational results (e.g., information detailing empirical experiments, if applicable). Starting in 2012 the preclinical life sciences and social psychology communities began discussions regarding reproducibility in earnest (Begley and Ellis 2012; *The Economist* 2013). These discussions focused on trust in research findings that included, but also extended beyond, computational transparency and reproducibility as described above. I have delineated three aspects of reproducibility to distinguish these conversations (Stodden 2013b):

- *Computational reproducibility*: transparency of any computational steps taken in the course of discovery;
- *Empirical reproducibility*: the traditional scientific notion of experimental researchers capturing descriptive information about (non-computational) aspects of their research protocols and methods, including information about the soundness and appropriateness of statistical methods applied.

Computational reproducibility has been summarized as "for an experiment to be reproducible, we need to have knowledge of at least the following information:

- Research data and metadata used
- Methods applied in the experiment
- Tools, software and execution environment used in the experiment" (Pröll and Rauber 2017)

In 2004 Gentleman and Temple Lang introduced the idea of the "Research Compendia" as "as both a container for the different elements that make up the document and its computations (i.e., text, code, data, ...), and as a means for distributing, managing and updating the collection." (Gentleman and Lang 2004) I will use the term Research Compendia to refer to the bundle of artifacts that support claims made in the article, and the article itself, in this chapter. For the purposes of this chapter I will focus on computational reproducibility, and use the term reproducibility as a synonym. I will not focus on what is often called replicability, which generally refers to the information needed to repeat the entire study independently beyond the computational aspects. Note that the computational information and artifacts associated with a study are needed to reconcile any differences in findings that might emerge from independent re-implementations of an experiments, indicating computational reproducibility is a necessary but not sufficient condition from independent replication.

FACETS REPOSITORIES NEED TO CONSIDER FOR REPRODUCIBILITY

In considering repositories designed to support the idea of really reproducible research discussed above, several facets deserve special consideration.

Research Compendia: Data and Code Must be Considered Together

The first step for repository design is to consider the article bundled along with supporting artifacts – the Research Compendia described above. There are lightweight ways of linking these objects, for example using DOI naming conventions and Crossref schema entries that refs to related objects (this is essential for all artifacts supporting published research results), to more involved linking such as digitally bundling artifacts in a container (Stodden and Miguez 2014). Container technology can be structured in a templated way designed to encourage reproducibility and the provision of digital artifacts associated with research results (Stodden, Wu, and Sochat 2018). A separate of the artifacts that support the claims from the claims and from each other leads to lack of discoverability and irreproducibility, and violates the integrity of the scholarly claim as the object of publication in the scholarly record.

Versioning and Persistent Unique Identifiers for Research Artifacts are Necessary

The Crossref service, mentioned above, has the primary mission of assigned and recording Digital Object Identifiers (DOIs) to scholarly objects. Its historical mission has included published scholarly articles, but has recently expanded to include scholarly object such as published dataset, software and code, and Research Compendia. This is one of the recommendations made as part of the "Reproducibility Enhancement Principles" REPS: Share data, software, workflows, and details of the computational environment that generate published findings in open trusted repositories (Stodden et al. 2016). This is important to enable persistence of links to these objects and to provide unique identifiers as artifact versions change as evolve. It is crucial to have unique identifiers, whether the DOI supplied by Crossref or another type such as Github hashes or a digital fingerprint (Altman 2008) or results identifier (Gavish and Donoho 2011). The important features of an identifier are that is it unique, persistent, and assigned at the appropriate level of granularity for citation. This is the second REPS Recommendation: To enable credit for shared digital scholarly objects, citation should be standard practice. New DOIs need to be issued when changes, corrections, or updates are made and published for any artifact. Note that DOI assignment, possible through the DataCite service for data, and registration via Crossref is an expense, both in obtaining the DOI and maintaining the persistence of the object to which the DOI pertains.

In short: Artifacts can change rapidly and subsets are frequently used, e.g., "researchers hardly use the whole data set at once, but rather create subsets of a larger data set. By a subset, we understand a selection of records and a projection of properties of the data set, based on defined parameters. Subsets are often specific for one particular study and contain implicit domain knowledge" (Pröll and Rauber 2017). Code and results must be connected with source data by identifiers supported by the repository and downstream research must be linked back to housed in ISSOD.

Appropriate Metadata and Documentation Must be Provided for all Artifacts

Metadata can include authorship, dates, version, unique identifier, funding source, and other information that depends on the type of artifact such as variable definitions, the data generation mechanism,

interoperability with other artifacts, dependencies such as other datasets or code. Funding sources for studies and for artifacts must be acknowledged in a standardized way, for example in a specific metadata field for data and/code or structured as part of the acknowledgements section and verified prior to deposit.

Emerging work regarding the structure of publishable Research Compendia sheds some light on standards. The Popper convention for structuring code and data compendia release (Jimenez et al. 2016; Jimenez et al. 2017) and CodeOcean.com's recent work describing their downloadable Compute Capsules (Green 2018) are important contributions. Both suggest structuring directories within the Compendia with separate folders for code, data, environment descriptions, overarching instructions for use, and attaching specific metadata fields to the Compendia via a .yml file. Here is CodeOcean's example metadata .yml file:

```
metadata_version: 1
name: Cape Feare
authors:
  - name: Sideshow Bob
affiliations:
  - name: The Krusty the Clown Show
corresponding_contributor:
name: Sideshow Bob
```

Such a metadata file would contain additional information such as artifact licensing for example.

Interoperability Must be Considered

The merging of datasets, re-use and combination of code, and the extension of research published in Research Compendia are all desirable activities enabled by the repository. Note that interoperability concerns appears at many levels: the Artifact level – data, code; the Research Compedia level; and even more granular level such as subsets of data and code.

Access to Artifacts Must be Maximized

In line with the principle articulated in the introduction, the benefits of reproducible research and maximized when access to artifacts is maximized (subject to legal barriers and other well-considered justifications) and supported with ways to report and correct errors by authors and by downstream users in general. There are several access issues when considering access to research artifacts that I discuss below.

Open Licensing Must be Used for Research Artifacts Intellectual Property law currents poses re-use barriers for researchers (Stodden 2014b, 2013a). Briefly, for authors and creators copyright protection adheres automatically when an original expression of an idea is rendered in fixed form. This applies to many, if not most, scientific activities, such as writing code to analysis data or prepare it for analysis or generating and a new dataset through the original selection and arrangement of data.[2] The default nature of copyright has the effect of creating an Intellectual Property framework odds with longstanding scientific norms in two key ways.[3] Firstly, it prevents copying the work, for example downloading

[2] See *Feist Publications Inc. v. Rural Telephone Servce Co.*, 499 U.S. 340 (1991) at 363-364.

[3] For a detailed discussion of copyright law and its impact on scientific innovation, see Stodden (2009a).

code to a computer, and thus creates a barrier to the legal reproduction and verification of results via running the code. Prior permission of the authoring researcher is needed (Stodden 2009b). Second, copyright establishes author rights regarding the creation of derivative works, such as re-use of software on a new dataset. Authors can follow the lead provided by open licensing practices in the open source software and artistic communities by realigning intellectual property rights with scientific norms by using the Reproducible Research Standard (RRS), a methods for applying appropriate open licenses that remove restrictions on copying and reuse of research works. Components of the Research Compendium have different features that necessitate using different licensing approaches. As such a principle for licensing scientific digital objects can guide choices:

> **Principle of Scientific Licensing**: Legal encumbrances to the dissemination, sharing, use, and re-use of scientific research compendia should be minimized, and require a strong and compelling rationale before their application.

This is followed on by another REPS recommendation: Use Open Licensing when publishing digital scholarly objects (Stodden et al. 2016).

For media components, the Reproducible Research Standard (RRS) suggests the Creative Commons attribution license (CC BY) which frees the work for replication and reuse without prior author approval, and provides for attribution to the original copyright holder. Use of the CC BY license for code is discouraged by Creative Commons.[4] For code components, the RRS recommends the Modified Berkeley Software Distribution (BSD) license, the MIT license, or the Apache 2.0 license, which permit the downstream use, copying, and distribution of either unmodified or modified source code, as long as the license accompanies any distributed code and the previous authors' names are not used to promote modified downstream code.[5]

Collecting, cleaning, and preparing data for analysis is typically a significant component of empirical research. Although "raw facts" are not copyrightable, in *Feist Publications, Inc. v. Rural Telephone Service*, the Court held that the original "selection and arrangement" of databases is copyrightable (Bitton 2006) and that "copyright protection extends only to those components of the work that are original to the author, not to the facts themselves…."[6] It may be possible to attaching an attribution license to the original "selection and arrangement" of a database.[7] These steps may be implemented in code or described in a text file accompanying the dataset, and an appropriate license would follow for

[4] See "Can I apply a Creative Commons license to software?" http://wiki.creativecommons.org/FAQ (last accessed Oct 22, 2018).

[5] http://opensource.org/licenses/bsd-license.php (last accessed Oct 22, 2018).

[6] *Feist Publications, Inc. v. Rural Telephone Service Co.*, Inc. 499 U.S. 340 at 340. The full quote is

> Although a compilation of facts may possess the requisite originality because the author typically chooses which facts to include, in what order to place them, and how to arrange the data so that readers may use them effectively, copyright protection extends only to those components of the work that are original to the author, not to the facts themselves. . . . As a constitutional matter, copyright protects only those elements of a work that possess more than de minimis quantum of creativity Rural's white pages, limited to basic subscriber information and arranged alphabetically, fall short of the mark. As a statutory matter, 17 U.S.C. sec. 101 does not afford protection from copying to a collection of facts that are selected, coordinated, and arranged in a way that utterly lacks originality. Given that some works must fail, we cannot imagine a more likely candidate. Indeed, were we to hold that Rural's white pages pass muster, it is hard to believe that any collection of facts could fail.

[7] See Sanders (2006) for a discussion of the international and WIPO statements of the legal status of databases.

mediu, for example the MIT License for code or the CC-BY license for text. Data itself ("raw facts") can be released to the public domain by using the CCo mark.[8]

A repository can check for appropriate licensing, offer defaults for example those that follow the RRS, and ensure that downstream derivative works confirm with licensing stipulations.

Data and Artifact Ownership Many entities often claim data ownership. Data collectors, curators, archivists, researchers involved in the preparation of data for analysis, just to name a few, can feel they have invested time and labor in the creation of data objects that confer ownership rights. Research subjects can feel ownership over "their" data. Traditional ideas of privacy protection may not confirm with wishes of human subjects in studies as they attempt to exert agency over data they perceive to be theirs. Some subject in studies would prefer that data about themselves, that might traditionally be considered worthy of privacy protection, be made more fully available.[9] As noted in the World Economic Forum Report, "[o]ne of the missing elements of the dialogue around personal data has been how to effectively engage the individual and give them a voice and tools to express choice and control over how data about them are used."[10,11] Funding agencies that support the research can be directive regarding ownership as can institutions and universities that have support research that generates data (Association of American Universities and Association of Public & Land-Grant Universities 2017). The story for code is similar since code can have multiple authors who have contributed to development sequentially over time for example.

Confidentiality in Data and Analysis Code Human subjects data are subject to myriad federal mandates and laws regarding disclosure, privacy, confidentiality, which influence the ability of researchers and repositories to share data. Linked data presents a challenging case for open data since linking can enable future privacy violations from data that are non-violating today (Nissenbaum and Barocas 2014). There are areas where solutions are emerging to maximize legal access including third party checks to enable trust in inaccessible artifacts; the creation data lakes for authorized access; and the use of methods that enhance trust, such as different groups working independently and cross checking findings that go to a particular hypothesis or discovery. This is an emerging and promising area of research that should be encouraged, following the final REPS recommendation: "To better enable reproducibility across the scientific enterprise, funding agencies should instigate new research programs and pilot studies."

The future linking of open data sets can allow individuals to be identified through via otherwise non-identifiable data. These potential future linkages cannot, by definition, be foreseen and can be of enormous benefit to research and discovery yet risk violated privacy norms and legal restrictions on data access. Data ownership can be difficult to construe, and individuals sometimes choose to release what might be considered private information by some.

[8] For details on the CCo protocol see http://creativecommons.org/press-releases/entry/7919 (last accessed Aug 21, 2013).

[9] Individuals may direct their data to be used for research purposes only, or to be placed in the public domain for broad reuse, for example. See, e.g., Consent to Research, http://weconsent.us, which supports data owner agency and informed consent for data sharing beyond the barriers dictated by traditional privacy protection.

[10] WEF Report p 12.

[11] Some restrictions on subject agency exist, see, e.g., *Moore v. Regents of University of California* 51 Cal.3d 120 Supreme Court of California July 9, 1990. This case dealt with ownership over physical human tissue, and not digital data, but the tissue could be interpreted as providing data for scientific experiments and research, in a similar role as that of data. See also the deal the National Institutes of Health made to continue research access to the Henrietta Lacks cell line, taking into account Lacks's family's privacy concerns (Callaway 2013).

Some concern about open code arises from the potential promulgation of misinformation and misuse as well as perceived privacy risks. In 2008, Taleb wrote about the dangers of using statistical methodology without having a clear understanding of the underlying techniques.[12] An example appeared on UCSF's EVA website, a repository of programs for automatic protein structure prediction. The UCSF researchers did not release their code publicly because "[w]e are seriously concerned about the 'negative' aspect of the freedom of the Web being that any newcomer can spend a day and hack out a program that predicts 3D structure, put it on the web, and it will be used." However a open dialog of an idea's merits is preferable to no dialog at all, and misinformation can be countered and exposed.

Really Really Big Data (and Code)

Dataset size, although generally not codebase size, can be a barrier to sharing in that it can require specialized computational infrastructure and tools. The Sloan Digital Sky survey creates different websites for the different data types, and provides different tools for access including SkyServer SQL search, CasJobs, and Schema Browser, each serving a different purpose.[13] This infrastructure also permits access to smaller subsets of the database. In some fields however even hundred of terabytes would not seem large. CERN director general Rolf Heuer said in 2008 that, "[t]en or 20 years ago we might have been able to repeat an experiment. They were simpler, cheaper and on a smaller scale. Today that is not the case. So if we need to re-evaluate the data we collect to test a new theory, or adjust it to a new development, we are going to have to be able reuse it. That means we are going to need to save it as open data."[14] Even 5 years ago, in March of 2013, the CERN data center exceeded 100 petabytes of stored data.[15]

Industry Collaborations

Collaboration with industry partners can be fruitful and important to the advancement of research. A company carrying out research may have different goals arising from their fiduciary responsibility to shareholders that may not include the goal of contributing their research findings to the public. Offering template agreements for data and code disclosure at the beginning of the collaboration can be helpful in aligning scientific priorities such as reproducibility.[16] Understanding the different research environment each group is subject to, industry research is not subject to IRB approval for example, can help enable productive collaborations that advance scientific research and understanding.

[12] N. Taleb, "The Fourth Quadrant: A Map of the Limits of Statistics." http://www.edge.org/3rd_culture/taleb08/taleb08_index.html (last accessed Oct 22, 2018).

[13] See http://www.sdss3.org/dr10/data_access/, including http://skyserver.sdss3.org/dr10/en/help/docs/sql/_help.aspx, http://skyserver.sdss3.org/CasJobs/ and http://skyserver.sdss3.org/dr10/en/help/browser/browser.aspx (last accessed Oct 22, 2018).

[14] "In search of the Big Bang," Computer Weekly, 2008. Available at http://www.computerweekly.com/feature/In-search-of-the-Big-Bang (last accessed Oct 22, 2018).

[15] "CERN data centre passes 100 petabytes," Cern Courier, Mar 28, 2013. Available at http://cerncourier.com/cws/article/cern/52730 (last accessed Oct 22, 2018). 100 petabytes is about 100 million gigabytes or 100,000 terabytes of data. This is equivalent to approximately 1500 copies of the Sloan Digital Sky Survey.

[16] See Kauffman chapter for further discussion of such template agreements.

EVALUATING REPOSITORY IMPACT

There exists a body of literature discussing the measuring and evaluation of the impact of repositories of social and economic data (Charles Beagrie Ltd and Centre for Strategic Economic Studies 2012). The creation of appropriate evaluation metrics for ISSOD will aid in assessing successes and shortcomings in the deployment of ISSOD, and perhaps more importantly focus stakeholders on key desirables. Ensuring downstream of artifacts made available through ISSOD by citation and licensing practices is important.

A FUTURE OF RADICAL SCIENTIFIC TRANSPARENCY

Imagining a future of sharing of artifacts, as routine aspect of carrying out and publishing research allows other affordances to be created. One can imagine visiting the following queries on the scholarly record as a starting point for research or as investigatory (Gavish and Donoho 2012):

- 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; and
- 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.

Journals and repositories can enhance trust by verifying computational results in Research Compendia themselves and certifying the claims. The Odum Institute at the University of North Carolina at Chapel Hill has developed a means of verifying computational results published in the American Journal of Political Science which has been used on 140 published AJPS articles (The Odum Institute 2018).

The development of computational research environments can also help capture and make available Research Compendia (Gowers et al. 2015). Some efforts include the Jupyter Notebook (Pérez and Granger 2007) and the WholeTale project at WholeTale.org (Brinckman et al. 2019). This is an exciting area for pilots projects and research and trying new ideas.

CONCLUSION

This chapter discusses the different notions of research reproducibility and how they may be instantiated at the level of repository design and research infrastructure development, with special attention paid to sensitive and proprietary data. The chapter suggests thinking of digital scholarly objects as linked to claims they support and explains the Research Compendia – a bundle of objects such as data and code that support published claims. It uses the Research Compendia as a way to frame repository

design and implement facets such as: Versioning and Persistent Unique Identifiers for Research Artifacts; Appropriate Metadata and Documentation for Artifacts; Interoperability between artifacts; and maximizing access to artifact to enable reproducibility via Open Licensing, Cyberinfrastrutcure design and managing partnerships. Finally, this chapter suggests metrics by which repository performance can be evaluated from the perspective of reproducibility and research and discovery enhancement.

References

Altman, Micah. 2008. "A Fingerprint Method for Scientific Data Verification." In *Advances in Computer and Information Sciences and Engineering*, edited by Tarek Sobh, 311–316. Dordrecht: Springer Netherlands. doi:10.1007/978-1-4020-8741-7_57.

Association of American Universities and Association of Public & Land-Grant Universities. 2017. *AAU-APLU Public Access Working Group Report and Recommendations.* Accessed June 11, 2019. https://www.aau.edu/sites/default/files/AAU-Files/Key-Issues/Intellectual-Property/Public-Open-Access/AAU-APLU-Public-Access-Working-Group-Report.pdf.

Bailey, David H., Jonathan M. Borwein, and Victoria Stodden. 2013. "Set the Default to "Open"." *Notices of the American Mathematical Society* 60 (6): 679. doi:10.1090/noti1014.

Begley, C. Glenn, and Lee M. Ellis. 2012. "Raise Standards for Preclinical Cancer Research." *Nature* 483 (7391): 531–533. doi:10.1038/483531a.

Bitton, Marion. 2006. "A New Outlook on the Economic Dimension of the Database Protection Debate." *IDEA: The Intellectual Property Law Review* 47 (2): 93–170.

Brinckman, Adam, Kyle Chard, Niall Gaffney, Mihael Hategan, Matthew B. Jones, Kacper Kowalik, Sivakumar Kulasekaran, et al. 2019. "Computing Environments for Reproducibility: Capturing the "Whole Tale"." *Future Generation Computer Systems* 94:854–867. doi:10.1016/j.future.2017.12.029.

Buckheit, Jonathan B., and David L. Donoho. 1995. "WaveLab and Reproducible Research." In *Wavelets and Statistics*, edited by Anestis Antoniadis and Georges Oppenheim, 103:55–81. New York, NY: Springer. Accessed June 11, 2019. doi:10.1007/978-1-4612-2544-7_5. http://link.springer.com/10.1007/978-1-4612-2544-7_5.

Callaway, Ewen. 2013. "Deal Done Over HeLa Cell Line." *Nature* 500 (7461): 132–133. doi:10.1038/500132a.

Charles Beagrie Ltd and Centre for Strategic Economic Studies. 2012. "Economic Impact Evaluation of the Economic and Social Data Service." Accessed June 11, 2019. https://esrc.ukri.org/files/research/research-and-impact-evaluation/economic-impact-evaluation-of-the-economic-and-social-data-service/.

Claerbout, Jon F., and Martin Karrenbach. 1992. "Electronic Documents Give Reproducible Research a New Meaning." In *SEG Technical Program Expanded Abstracts 1992*, 601–604. Society of Exploration Geophysicists, January. doi:10.1190/1.1822162.

Donoho, David L., Arian Maleki, Inam Ur Rahman, Morteza Shahram, and Victoria Stodden. 2009. "Reproducible Research in Computational Harmonic Analysis." *Computing in Science & Engineering* 11 (1): 8–18. doi:10.1109/MCSE.2009.15.

Gavish, Matan, and David Donoho. 2011. "A Universal Identifier for Computational Results." *Procedia Computer Science* 4:637–647. doi:10.1016/j.procs.2011.04.067.

———. 2012. "Three Dream Applications of Verifiable Computational Results." *Computing in Science & Engineering* 14 (4): 26–31. doi:10.1109/MCSE.2012.65.

Gentleman, Robert, and Duncan Temple Lang. 2004. *Statistical Analyses and Reproducible Research.* Bioconductor Project Working Paper 2. https://biostats.bepress.com/bioconductor/paper2.

Gowers, Timothy, Nicholas J. Higham, Ian Stewart, David L. Donoho, Victoria Stodden, David H. Bailey, Jonathan M. Borwein, and Heather Mendick. 2015. "Final Perspectives." In *The Princeton Companion to Applied Mathematics,* 897–962. Princeton, NJ: Princeton University Press. JSTOR: j.ctt1gr7dbs.13.

Green, Seth. 2018. "Exporting Capsules and Reproducing Results on Your Local Machine." Accessed June 11, 2019. http://help.codeocean.com/user-manual/sharing-and-finding-published-capsules/exporting-capsules-and-reproducing-results-on-your-local-machine.

Jimenez, Ivo, Michael Sevilla, Noah Watkins, Carlos Maltzahn, Jay Lofstead, Kathryn Mohror, Andrea Arpaci-Dusseau, and Remzi Arpaci-Dusseau. 2017. "The Popper Convention: Making Reproducible Systems Evaluation Practical." In *Proceedings 2017 IEEE International Parallel and Distributed Processing Symposium Workshops,* 1561–1570. Orlando / Buena Vista, FL, USA: IEEE. doi:10.1109/IPDPSW.2017.157.

Jimenez, Ivo, Michael Sevilla, Noah Watkins, Carlos Maltzahn, Jay Lofstead, Kathryn Mohror, Remzi Arpaci-Dusseau, and Andrea Arpaci-Dusseau. 2016. "Standing on the Shoulders of Giants by Managing Scientific Experiments Like Software." *USENIX; login* 41 (4): 20–26.

King, Gary. 1995. "Replication, Replication." *PS: Political Science and Politics* 28 (3): 444. doi:10.2307/420301.

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, et al. 2009. "Computational Social Science." *Science* 323 (5915): 721–723. doi:10.1126/science.1167742.

Nissenbaum, Helen, and Solon Barocas. 2014. "Big Data's End Run around Anonymity and Consent." In *Privacy, Big Data, and the Public Good: Frameworks for Engagement,* edited by Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nissenbaum, 44–75. New York, NY: Cambridge University Press.

Pérez, Fernando, and Brian E. Granger. 2007. "IPython: A System for Interactive Scientific Computing." *Computing in Science & Engineering* 9 (3): 21–29. doi:10.1109/MCSE.2007.53.

Pröll, Stefan, and Andreas Rauber. 2017. "Enabling Reproducibility for Small and Large Scale Research Data Sets." *D-Lib Magazine* 23 (1/2). doi:10.1045/january2017-proell.

Sanders, Anselm Kamperman. 2006. "Limits to Database Protection: Fair Use and Scientific Research Exemptions." *Research Policy* 35 (6): 854–874. doi:10.1016/j.respol.2006.04.007.

Schwab, Matthias, Martin Karrenbach, and Jon Claerbout. 2000. "Making Scientific Computations Reproducible." *Computing in Science & Engineering* 2 (6): 61–67. doi:10.1109/5992.881708.

Stodden, Victoria. 2009a. "Enabling Reproducible Research: Licensing for Scientific Innovation." *International Journal for Communications Law and Policy* 13. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1362040.

Stodden, Victoria. 2009b. "The Legal Framework for Reproducible Scientific Research: Licensing and Copyright." *Computing in Science & Engineering* 11 (1): 35–40. doi:10.1109/MCSE.2009.19.

———. 2013a. "Intellectual Property and Computational Science." In *Opening Science: The Evolving Guide on How the Internet Is Changing Research, Collaboration and Scholarly Publishing,* edited by Sönke Bartling and Sascha Friesike, 225–236. New York, NY: Springer.

———. 2013b. "Resolving Irreproducibility in Empirical and Computational Research." Accessed June 11, 2019. http://bulletin.imstat.org/2013/11/resolving-irreproducibility-in-empirical-and-computational-research/.

———. 2014a. "Enabling Reproducibility in Big Data Research: Balancing Confidentiality and Scientific Transparency." In *Privacy, Big Data, and the Public Good: Frameworks for Engagement,* edited by Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nissenbaum. New York, NY: Cambridge University Press.

———. 2014b. "What Computational Scientists Need to Know About Intellectual Property Law: A Primer." In *Implementing Reproducible Research,* edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng, 325–342. Boca Raton, FL: CRC Press.

Stodden, Victoria, Jonathan M. Borwein, and David H. Bailey. 2013. ""Setting the Default to Reproducible" in Computational Science Research." Accessed June 11, 2019. https://sinews.siam.org/Details-Page/setting-the-default-to-reproducible-in-computational-science-research.

Stodden, Victoria, Marcia McNutt, David H. Bailey, Ewa Deelman, Yolanda Gil, Brooks Hanson, Michael A. Heroux, John P. A. Ioannidis, and Michela Taufer. 2016. "Enhancing Reproducibility for Computational Methods." *Science* 354 (6317): 1240–1241. doi:10.1126/science.aah6168.

Stodden, Victoria, and Sheila Miguez. 2014. "Best Practices for Computational Science: Software Infrastructure and Environments for Reproducible and Extensible Research." *Journal of Open Research Software* 2, no. 1 (July 9): e21. doi:10.5334/jors.ay.

Stodden, Victoria, Xiaomian Wu, and Vanessa Sochat. 2018. "AIM: An Abstraction for Improving Machine Learning Prediction." In *Proceedings of the 2018 IEEE Data Science Workshop (DSW),* 1–5. IEEE. doi:10.1109/DSW.2018.8439914.

The Economist. 2013. "Trouble at the Lab." Accessed June 11, 2019. https://www.economist.com/briefing/2013/10/18/trouble-at-the-lab.

The Odum Institute. 2018. "Confirmable Reproducible Research (CoRe2) Environment: Linking Tools to Promote Computational Reproducibility." Accessed June 11, 2019. https://odum.unc.edu/2018/07/alfred-p-sloan-foundation-grant/.