Beyond Open Data: A Model for Linking Digital Artifacts to Enable Reproducibility of Scientific Claims

Victoria Stodden*
vcs@stodden.net
University of Illinois at Urbana-Champaign
Champaign, IL

ABSTRACT

The last few years has seen a substantial push toward "Open Data" by policy makers, researchers, archivists, and even the public. This article postulates that the value of data is not intrinsic but instead derives from its ability to produce knowledge; the extraction of which from data is not deterministic. The value of data is realized through a focus on the reproducibility of the findings from the data, which acknowledges the complexity of the leap from data to knowledge, and the inextricable interrelationships between data, software, computational environments and cyberinfrastructure, and knowledge. Modern information archiving practices have a long history and were shaped in a pre-digital world comprised of physical objects such as books, monographs, film, paper, and other physical artifacts. This article argues that "data," the modern collection of digital bits representing empirical measurements, is a wholly new entity and not a digital analog to any physical object. It further argues that a focus on the interrelationships between digital artifacts and their unique properties, instead of Open Data alone, will instead produce an augmented and more useful understanding of knowledge when it is derived from digital data. Data-derived knowledge, represented by claims in the scholarly record, must persistently link to immutable versions of the digital artifacts from which it was derived, including 1) any data, 2) software that allows access to the data and the regeneration of those claims that rely on the version of the data, and 3) computational environment information including input parameters, function invocation sequences, and resource details. In this sense the epistemological gap between data and extracted knowledge can be closed. Datasets and software are often subject to change and revision, sometimes even with high velocity, and such changes imply new versions with new unique identifiers. We propose considering knowledge, rather than data in isolation, with a schematic model representing the interconnectedness of datasets, software, and computational information upon which its derivation depends. Capturing the interconnectedness of these digital artifacts, and their relationship to the knowledge they generate, is essential for supporting the reproducibility, transparency, and cognitive tractability of scientific claims derived from digital data.

P-RECS '20, June 23, 2020, Stockholm, Sweden

ACM ISBN 978-1-4503-7977-9/20/06.
https://doi.org/10.1145/3391800.3398172

CCS CONCEPTS

• **General and reference** → **Verification**; **Cross-computing tools and techniques**; • **Information systems** → **Information systems applications**; *Information extraction*;

KEYWORDS

Open Knowledge; Reproducibility; Data Sharing; Code Sharing; Data Archiving; Software Archiving; Data Policy; Data Access; Open Data; Information Representation

ACM Reference Format:
Victoria Stodden. 2020. Beyond Open Data: A Model for Linking Digital Artifacts to Enable Reproducibility of Scientific Claims. In *3rd International Workshop on Practical Reproducible Evaluation of Computer Systems (P-RECS '20), June 23, 2020, Stockholm, Sweden.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3391800.3398172

1 INTRODUCTION

The library has traditionally served the important role of steward of the scholarly record. The traditional model from the print era saw libraries maintaining rich archives of journals and conference proceedings, accessible as bound paper volumes accessed in a physical library. Each article published in these volumes was presented as a novel contribution extending our stock of knowledge, and was self-contained in the sense that the information needed to understand and verify the scientific claims made was contained in or accessible through the published article itself. When research is computationally-and data-enabled, the scientific scholarly publication is no longer self-contained as traditionally understood. Access to data, verification of the computational steps and code that lead from data to inference, and exposition of the relevant details about the digital aspects of the research process do not yet have a broadly accepted and structured place within the traditional published article and scholarly record. Access to supporting digital scholarly objects and the importance of their long term stewardship is a facet of reproducibility discussions in the scientific community [12, 13, 28, 37, 38]. The 2019 National Academies of Science, Engineering, and Medicine report on "Reproducibility and Replication in Science" defined reproducibility as "obtaining consistent results using the same input data, computational steps, methods, and code, and conditions of analysis" and we follow this definition in this writing [26]. Note that reproducibility does not necessarily imply the ability to re-execute the software in perpetuity.

The "Open Data" effort recognizes that the majority of published scientific research today, perhaps nearly all, rely on digital scholarly objects, such as data, that are not typically included in the published article itself. This article traces previous work on the evolving scholarly record in the digital age, then argues that claims

extracted from digital datasets and data collections require special consideration with regard to archiving and access due to their unique relationships to the digital artifacts upon which they depend. Finally we present a model that links digital artifacts in support of "Open Knowledge" through reproducibility. This linked model indicates future directions in which the scholarly record is curated with persistent links to digital objects that support its claims.

2 PREVIOUS WORK

The effect of digitization on the scholarly record, including the digitization of books as well as scholarly publications in journals and conference proceedings, has a rich history of scholarly attention [4, 14, 20]. In addition, new scholarly objects are now routinely archived by libraries including datasets and software [5, 6, 11, 19]. The nature of data archiving itself is evolving. Digital data archives vary widely in organizational structure, mission, collection, funding, and relationships to their users and other stakeholders. Only relatively recently have forums such as the Research Data Alliance (founded 2013) and Force11 (founded 2011) coalesced to discuss common interests, policies, practices, and technologies that span research domains, countries, and communities. Kuny is the first to our knowledge to explicitly note the existence of inextricable relationships between digital scholarly objects from an archivist's perspective: "Some types of information, such as multimedia, are so closely linked to the software and hardware technologies that they cannot be used outside these proprietary environments" however this description does not refer to the essential epistemological connections between digital artifacts for knowledge extracted from data [17]. Borgman however makes this observation explicit for data and code: "data are inseparable from the software code used to clean, reduce, and analyze them." ([5] p. 106). However, there is "a broad-based movement toward publication practices that permit results to be readily reproduced, at least in the computational sciences." [22, 32, 33]. These efforts recognize the special and novel nature of the relationships between data, code, and published claims arising from data inference.

Researchers sometimes take steps on their own initiative to make digital artifacts that support their published claims available, frequently citing reproducibility, often using platforms and infrastructures developed outside the scholarly community and typically intended for other purposes [12]. For example, some researchers use the web-based platform GitHub.com for software development as well as software archiving and access [16, 42]. However GitHub.com is not designed for housing or archiving data as it is explicitly a software development platform (e.g. file size limitations of 2GB), and its use by the research community indicates an infrastructure divide between data and software from an archiving perspective. GitHub and other software development platforms have collaborative aspects in its infrastructure (e.g. the ability to commit, branch, fork, merge software contributions) that are not typically associated with infrastructure that archives and preserves datasets. There are entirely distinct infrastructure ecosystems for data and software and in the next section we proffer reasons as to why that might be, from differing policies to intrinsic attributes of data and code.

The research community has taken several steps to encourage, facilitate, and require data citation when data relied on to make claims in a publication, and when data are published independently. Some of the most notable examples are discussed in this section. As yet, broadly accepted citation standards are evolving and are not yet uniform or standardized. Publishers, journals, and repositories are providing predefined "badges" that can be used to kitemark a publication. An example is the "Open Data" badge defined by the Center for Open Science (see https://cos.io/our-services/open-science-badges/). The Association for Computing Machinery (ACM) Digital Library, in concert with its publishing arm, offers badges to indicate a publication's level of reproducibility including the "Artifacts Available" badge and the "Artifacts Evaluated" badge. Data is considered included in the definition of artifact [25]. Finally, a National Information Standards Organization (NISO) Taxonomy, Definitions, and Recognition Badging Scheme Working Group[1] is charged with advancing badging scheme standardization across the Computational and Computing Sciences (https://www.niso.org/standards-committees/reproducibility-badging). As discussed in the next section, journals and publishers are taking steps to encourage and require data availability for their articles, and data citation [36, 40]. Two of the most prominent examples of author requirements to make data and code that support the claims in the article openly available, prior to publication, come from *Science* and *Nature* [23, 29]. Finally, there are broad community-wide efforts to guide data access and citation practices, along with other artifacts. For example, the Center for Open Science hosts the Transparency and Openness Promotion (TOP) guidelines for journals and publishers,[2] which define four levels at which journals can comply with openness and transparency in their publication policies and practices (https://cos.io/top/) [24].

3 PRESERVING DIGITAL ARTIFACTS THAT SUPPORT SCIENTIFIC CLAIMS IS NECESSARY FOR REPRODUCIBILITY

Data preservation is motivated by many reasons: including re-use, reproducibility of scholarly claims, verification, and avoidance of duplication of effort to name a few [38]. The data preservation policies of research funding bodies, journals, and libraries can also influence what is preserved and why.

3.1 Policies Must Recognize the Relationships Between Digital Artifacts That Generate Knowledge

In 2011 the National Science Foundation began requiring two page Data Management Plans to be submitted with research proposals. Data in this context can mean datasets, software, workflow information, samples or other products or output of the funded research, however the use of the term data means the emphasis tends to be on datasets generated by the grant. Since 2011 the seven directorates comprising the National Science Foundation have developed their own guidance and requirements for the Data Management Plans submitted with proposal to their directorate.

[1] I am a member of the NISO Taxonomy, Definitions, and Recognition Badging Scheme Working Group.
[2] I am a member of the TOP steering committee.

In October of 2018 The National Institutes for Health released a Request for Information entitled "Request for Information (RFI) on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research" https://grants.nih.gov/grants/guide/notice-files/NOT-OD-19-014.html. Unlike the National Science Foundation, the National Institutes for Health has a long history of supporting its own repositories and infrastructure for specific types of output data from its research, e.g. NCBI BioCollections (https://www.ncbi.nlm.nih.gov/biocollections) and genotype data in the dbGaP repository (https://www.ncbi.nlm.nih.gov/gap). Research proposals to the Department of Energy have required Data Management Plans since 2014. This effort mentions data, meta data, and software: "A statement of plans for data and metadata content and format including, where applicable, a description of documentation plans, annotation of relevant software" and "Sharing and Preservation: any special requirements for data sharing, for example, proprietary software needed to access or interpret data, applicable policies, provisions, and licenses for re-use and re-distribution, and for the production of derivatives, including guidance for how data and data products should be cited" (https://science.energy.gov/funding-opportunities/digital-data-management/suggested-elements-for-a-dmp/). The Department of Energy has the most specific directive regarding software, stating in 2016 that software produced from grants must be at least 20% released open source (e.g. on GitHub.com): "establishes a pilot program that requires agencies, when commissioning new custom software, to release at least 20 percent of new custom-developed code as Open Source Software (OSS) for three years, and collect additional data concerning new custom software to inform metrics to gauge the performance of this pilot" (https://sourcecode.cio.gov/). Due May 6 2020, the Office of Science and Technology Policy in the Whitehouse requested information on "Public Access to Peer-Reviewed Scholarly Publications, Data and Code Resulting From Federally Funded Research" (https://www.federalregister.gov/documents/2020/03/31/2020-06622/request-for-information-public-access-to-peer-reviewed-scholarly-publications-data-and-code).

3.2 Policies Often Prioritize Open Data, and Open Code Lags

Recent research has shown a meaningful increase in the number of journals with data and/or code release policies associated with the claims in the manuscripts they choose to publish. Interestingly, journal policies regarding access to research code tend to come about several years after data policies [36].

Publishers, libraries, and independent organizations have developed efforts to coordinate the elements of the scholarly record. CrossRef (crossref.org), established in 1999 by a consortium of publishers, establishes a system to implement and recognize unique identifiers for articles in the scholarly record. Unique identifiers were developed shortly afterwards in 2000, called Digital Object Identifiers or DOIs (doi.org) and used by CrossRef to uniquely identify objects across a variety of publishers and other entities. DataCite, an organization in the same vein, was created in 2009 and acts as a registration agency for research data. The identifying information for both articles and data is collected in standardized schemas. We note that software is not explicitly included in the current management of digital objects, although software objects are often classified as data objects. The Scholix Framework (SCHOlarly LInk eXchange) was launched in 2016 [1] "to create an open global information ecosystem to collect and exchange links between research data and literature" (see http://www.scholix.org/about).

4 BEYOND DATA: INCLUDING ALL SUPPORTING DIGITAL ARTIFACTS

This section seeks to motivate the importance of linking published computational claims to all their supporting digital artifacts, typically software, data, and other information on the computational environment, function invocation sequences, and information on inputs, parameters and other settings. These supporting artifacts are inextricably related to each other, for example the software that extracted scientific inferences from the data, or generated the data. We argue that that digital infrastructure designed for data archiving is not a perfect fit for software artifacts or other information needed for computational reproducibility of scientific claims, and expansion in the conceptualization, design, and deployment of infrastructure is needed. What follows is an enumeration of challenges in realizing the vision of pervasively linked artifacts for reproducibility. As is evident by the challenges, advancing this vision requires the coordinated engagement of a variety of stakeholders, including researchers, scientific societies, publishers and journal editors, libraries and repositories, and funders. The solutions will differ for each challenge and our purpose is to motivate the landscape of differences between artifacts that give rise to their differential treatment when sharing, preserving and re-use, and the need to ensure these artifacts are linked and discoverable through the scientific claim they support.

4.1 Challenge 1: Data and Software Can Change Frequently

Datasets are subject to change and revision, often with high velocity, and such changes may require new unique identifiers for new versions of datasets that are derived from previous versions. The existence of these relationships between datasets imply relationships between object identifiers that embed information about these relationships. Similarly, software can be updated and changed (often for different reasons and in different ways than data) and unique identifiers may also be appropriate to assign to new versions. This mimics the current approach to the publication of findings, where new evidence is also published, allowing the community to revise their understanding of the underlying research (of course in the case of outright mistakes or fraud it should be possible to retract any digital scholarly object from the scholarly record, just as for publications).

4.2 Challenge 2: Data and Software Ownership and Rights Structures Are Different

The preceding discussion is predicated on the author having the ability and legal rights to share data and code associated with claims. When claims are self contained within articles – the model from the print era – rights over the text and figures in the article generally resided with the authors, who were then able to publish or

otherwise release their manuscript. Rights over data and code are not so clear cut: they may have different authors or creators than the manuscript, and data and code may have many different contributors at different points in time. Data is different to software in ways that alter the rights assigns and ability to publish. Data may be, for example, obtained or scraped from many sources each with its own terms of use. Software, like text and figures, is subject to copyright in the United States, whereas data generally are not [31]. Software may contain proprietary or potentially patentable algorithms. Each of these facets engenders different treatment regarding rights, accessibility, sharing, and re-use. In short, software cannot be considered as another form of data in scholarly communication with regard to sharing, re-use, and reproducibility.

This questions of artifact ownership and rights is complicated when viewed from a global perspective. Different countries and regions employ different Intellectual Property approaches and regimes which can shapes rights regarding research artifact sharing and re-use in different ways. For example in the United States Feist Pubs., Inc. v. Rural Tel. Svc. Co., Inc., 499 U.S. 340 (1991) established no copyright for raw facts up to "original selection and arrangement," a holding brought to bear for data, and the European Union has copyright protection for data. If there has been a substantial investment in obtaining, verifying or presenting the contents of a database, EU law may endow a sui generis database right to the database owner which lasts for 15 years and permits the owner the exclusive right to extract large or repeated small amounts of data from the database. These regional differences in rights and polices can introduce legal complications when, for example, merging and sharing data for difference regions. For more detail on Intellectual Property and computational science see e.g. [8, 34, 35].

4.3 Challenge 3: Data and Software Engender Different Preservation Strategies

Software, by definition, is meant to execute on a computational system. Data, by contrast, can exist in a fixed form, without adapting to computational systems. Software requires specialized maintenance to keep running and has different types of dependences to data since it relies on other software components to execute. Data, however can reach a size and scale that software is very unlikely to reach. Even a complex set of codes is unlikely to reach multiple GBs in size, whereas there are numerous examples of data at that scale and significantly larger. The CMIP5 climate science data set is over 3PB, and one set of products (DR12) in the Sloan Digital Sky Survey is over 116TB in size (https://www.sdss.org/dr12/data_access/volume/) for example.

Extensive work developing provenance standards for data (for example PROV-O, BPMN, DCAT, SDMX, DataCube, SSN/SOSA and Schema.org, see e.g. https://www.w3.org/TR/prov-overview/) are not directly applicable to code which evolves differently and comprises fundamentally different digital objects, nor were they designed with a focus on scientific results. Rather than measurements typically embodied in data, software is designed to execute on computational platforms and is more closely related to language than measurement requiring a rethinking of data best practices in the software context [2, 15, 39]. Data itself is not a monolithic concept. Different types of data also require different preservation strategies, such as digital images vs scanned images, human subjects data, derived vs collected data, and data generated by software simulation, just to name a few [41]. In addition, the effects of resolving the various preservation complexities associated with different artifacts may fall disproportionately on some fields vs others. Fields that rely on complex or large data and code bases, data with privacy concerns, or proprietary code for example, may have a need to undertake more extensive preservation and archiving approaches.

4.4 Challenge 4: Ethical Considerations Differ for Data and Code

Throughout this discussion, we have presumed digital artifacts are legally sharable. This may not always be the case and policy frameworks are developing around data sharing and access. Data may contain personally identifiable information when generated in the course of human subjects research or even through behavior such as internet browsing. Many types of data with private or confidential information are subject to federal rules controlling its release e.g. the Federal Acts HIPAA, FERPA [18]. There is no comparable set of privacy protecting regulations regarding the release of software, however software is subject to a set of Intellectual Property rights that data are not. These rights can create barriers to release, including copyright and potentially patentability. These aspects of data and software imply a different set of ethics around stewardship of each, for the collection of the data or writing of the code, right through to curation and archiving. Privacy in data is an area under rapid policy development and scholarly research and a full treatment is beyond the scope of this work [3, 10].

5 A LINKED STRUCTURE FOR DATA, CODE, AND PUBLISHED SCIENTIFIC CLAIMS IS NECESSARY

Links between digital scholarly objects imply the need for immutable versioning, for example through the assignment of a DOI, that are persistent and persistently connected both to claims in the scholarly record. This points to a corollary: the software associated with published claims, whether creating the claims itself or deriving claims from data, also needs a version and unique persistent identifiers as do data. These three identifiers: for example the published article containing the scholarly claims; the data supporting those claims; and the software that analyzed the data to discover or generate the claims, rely on each other to support the research. The metadata schemas associated with DOIs (e.g. https://schema.datacite.org/) contain relational information regarding the interconnectedness of digital scholarly objects, implying that such a change in information representation requires very little new infrastructure to recognize the inseparability of data and software, and the claims they support.

Use of DOI schema information for example can achieve persistent and discoverable connections between digital artifacts necessary for reproducibility of computationally- and data-enabled claims in the scholarly literature. Even if such a linked structure for the digital artifacts that support scientific claims became routine with publication, we still face other issues such as the appropriate repositories for stewardship of these sets of artifacts and support for their curation, definitions and standards for digital artifacts

intended to support reproducibility, and incentive structures for data and code producing researchers to make these artifacts available. Providing incentives to make reproducibility-enabling artifacts available for stewardship and curation implies fundamental changes in the research reward system. Such changes are underway, for example the National Science Foundation recently began recognizing relevant artifacts on its biosketch, a change from only permitting publications; and journals are increasingly implementing data and code availability policies [21, 36].

When considering next steps in advancing the vision of linked artifacts, each of the challenges in the previous section come to bear. A community effort, extending that under way for Open Data, is essential. There are some examples and efforts underway for example, considering and applying features used for software development and versioning in GitHub and/or BitBucket may be useful; learning from repositories such as Zenodo.org, DataVerse.org, and SoftwareHeritage.org regarding artifact preservation for the research community to enable effective preservation of data, software, and other artifacts and their linking to scientific findings; and examining pilot solutions for generating linked archival artifacts from the research process itself such as wholetale.org [7, 9] and the Open Science Framework [27]. Greater coordination and alignment of efforts around a vision of archiving research artifacts that are linked to scientific claims is essential.

A system of linked data and software opens new possibilities for automatic analysis of the scholarly record, including enabling different reproducibility checks. Versioning and artifact linking can enable, at least at the time of publication, the ability to understand the computational steps that lead to a particular scientific conclusions and, depending on computational details (such as dependencies and backward compatibility) could enable automated computational regeneration of results. With this linked publication model, changes to artifacts such as dataset updates or code bug fixes, can be tracked along with the revisions to scientific findings. One could imagine research suggestions aimed at extending published results to revised data or new methods. Linking may indicate future avenues and recommendations for replicability, e.g. new experiments, not only computational reproducibility, extending longstanding visions of artificial intelligence contributions to information retrieval [30].

Today human-in-the-loop intervention is a significant part of research and a researcher needs to understand the contents and implementation of reproducible packages to use them correctly. Linking may also require efforts to understand data and software in relation to each other as well as individual artifacts. For example provenance models do not translate to subsets of data or determine the parts of a research or data processing activity will affect data. With a linked model, traditional notions of provenance may be enhanced in these ways to become more effective or useful to the research community from a reproducibility point of view.

6 CONCLUSIONS

The use of computationally- and data-enabled methods for scientific discovery is expanding the scope of the scholarly record, by including digital scholarly objects upon which published claims rely. We intend to spark a conversation about the interconnected reliance of the claims on these artifacts and how infrastructures can enable persistence through linking and repositories that accommodate these artifacts as a collection, including software as well as data. Notions of provenance for datasets in the archival sense of a chain of custody and in the computing sense of transformations from original state are well developed in comparison to provenance for changes to software [5]. Incentives to make supporting artifacts available must come with the development of the knowledge infrastructures to support them. We present an argument that claims extracted from digital datasets and data collections require special consideration with regard to reproducibility due to their interrelated nature. Being able to access the data from which claims were extracted does not guarantee the ability to understand why those claims may or may not be correct, for any particular publication. We present an approach that is designed to allow a consumer of published research to access a collection of digital artifacts, including open data and evolving away from a sole focus on open data, that permit computational reproducibility, including transparency in how the claims were derived.

ACKNOWLEDGMENTS

We thank an anonymous colleague who provided valuable comments on an early draft, and very thoughtful anonymous reviewers. This material is based upon work supported by the National Science Foundation under Grants OAC-1839010 and OAC-1941443.

REFERENCES

[1] 2016. *RDA and ICSU-WDS Announce the Scholix Framework for Linking Data and Literature.* https://www.icsu-wds.org/news/news-archive/rda-and-icsu-wds-announce-the-scholix-framework-for-linking-data-and-literature

[2] P. Alliez, R. D. Cosmo, B. Guedj, A. Girault, M. Hacid, A. Legrand, and N. Rougier. 2020. Attributing and Referencing (Research) Software: Best Practices and Outlook From Inria. *Computing in Science & Engineering* 22, 1 (2020), 39–52.

[3] K. A. Bamberger and D. K. Mulligan. 2010. Privacy on the Books and on the Ground. *Stan. L. Rev.* 63 (2010).

[4] C. Borgman. 2007. *Scholarship in the Digital Age.* MIT Press.

[5] C. Borgman. 2015. *Big Data, Little Data, No Data: Scholarship in the Networked World.* MIT Press.

[6] C. L. Borgman, H. Van de Sompel, A. Scharnhorst, H. van den Berg, and A. Treloar. 2015. Who uses the digital data archive? An exploratory study of DANS. *Proceedings of the Association for Information Science and Technology* 52, 1 (2015), 1–4. https://doi.org/10.1002/pra2.2015.145052010096

[7] A. Brinckman, K. Chard, N. Gaffney, M. Hategan, M. B. Jones, K. Kowalik, S. Kulasekaran, B. Ludäscher, B. D. Mecum, J. Nabrzyski, V. Stodden, I. J. Taylor, M. J. Turk, and K. Turner. 2019. Computing environments for reproducibility: Capturing the "Whole Tale". *Future Generation Comp. Syst.* 94 (2019), 854–867. https://doi.org/10.1016/j.future.2017.12.029

[8] Michael W. Carroll. 2015. Sharing Research Data and Intellectual Property Law: A Primer. *PLOS Biology* 13, 8 (08 2015), 1–11. https://doi.org/10.1371/journal.pbio.1002235

[9] K. Chard, N. Gaffney, M. B. Jones, K. Kowalik, B. Ludäscher, J. Nabrzyski, V. Stodden, I. Taylor, M. J. Turk, and C. Willis. 2019. Implementing Computational Reproducibility in the Whole Tale Environment. In *Proceedings of the 2nd International Workshop on Practical Reproducible Evaluation of Computer Systems* (Phoenix, AZ, USA) *(P-RECS '19).* ACM, New York, NY, USA, 17–22. https://doi.org/10.1145/3322790.3330594

[10] D. Citron and D. Gray. 2013. The Right to Quantitative Privacy. *Minnesota Law Review* 98, 62 (2013).

[11] A. R. Diekema, A. Wesolek, and C. D. Walters. 2014. The NSF/NIH Effect: Surveying the Effect of Data Management Requirements on Faculty, Sponsored Programs, and Institutional Repositories. *The Journal of Academic Librarianship* 40, 3 (2014), 322 – 331. https://doi.org/10.1016/j.acalib.2014.04.010

[12] D. L. Donoho, A. Maleki, I. U. Rahman, M. Shahram, and V. Stodden. 2009. Reproducible Research in Computational Harmonic Analysis. *Computing in Science Engineering* 11, 1 (Jan 2009), 8–18. https://doi.org/10.1109/MCSE.2009.15

[13] R. Gentleman and D. Temple Lang. 2007. Statistical Analyses and Reproducible Research. *Journal of Computational and Graphical Statistics* 16, 1 (2007), 1–23.

https://doi.org/10.1198/106186007X178663

[14] T. Hahn. 2008. Mass Digitization: Implications for Preserving the Scholarly Record. *Library Resources & Technical Services* 52, 1 (2008), 18–26.

[15] D. Irving. [n.d.]. Best practices for scientific software. https://software.ac.uk/blog/2017-11-29-best-practices-scientific-software

[16] C. Kelty. 2008. *Two Bits: The Cultural Significance of Software.* Duke University Press.

[17] T. Kuny. 1998. The digital dark ages? Challenges in the preservation of electronic information. *International Preservation News* 17, 8 (1998).

[18] J. Lane, V. Stodden, S. Bender, and H. Nissenbaum. 2014. *Privacy, big data, and the public good: Frameworks for engagement.* Cambridge University Press.

[19] R. P. Light, D. E. Polley, and K. Börner. 2014. Open data and open code for big science of science studies. *Scientometrics* 101, 2 (November 2014), 1535–1551. https://doi.org/10.1007/s11192-014-1238-2

[20] B. Mak. 2014. Archaeology of a digitization. *Journal of the Association for Information Science and Technology* 65, 8 (2014), 1515–1526. https://doi.org/10.1002/asi.23061

[21] M. R. Munafò, B. A. Nosek, D. V. M. Bishop, K. S. Button, C. D. Chambers, N. Percie du Sert, U. Simonsohn, E. Wagenmakers, J. J. Ware, and J. P. A. Ioannidis. 2017. A manifesto for reproducible science. *Nature Human Behaviour* 1, 1 (2017), 0021. https://doi.org/10.1038/s41562-016-0021

[22] Edwards P. N., Jackson S. J., Chalmers M. K., Bowker G. C., Borgman C. L., Ribes D., Burton M., and Calvert S. 2013. Knowledge Infrastructures: Intellectual Frameworks and Research Challenges. *Ann Arbor: Deep Blue* (2013).

[23] Nature. 2020. *Reporting standards and availability of data, materials, code and protocols.* https://www.nature.com/nature-research/editorial-policies/reporting-standards

[24] B. A. Nosek, G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. Madon, N. Malhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. Levy Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E. J. Wagenmakers, R. Wilson, and T. Yarkoni. 2015. Promoting an open research culture. *Science* 348, 6242 (2015), 1422–1425. https://doi.org/10.1126/science.aab2374

[25] Association of Computing Machinery. 2018. *Artifact Review and Badging.* https://www.acm.org/publications/policies/artifact-review-badging

[26] National Academies of Sciences Engineering and Medicine. 2019. *Reproducibility and Replicability in Science.* The National Academies Press, Washington, DC. https://doi.org/10.17226/25303

[27] Open Science Framework (OSF). 2017. Open Science Framework (OSF). *Journal of the Medical Library Association : JMLA* 105, 2 (04 2017), 203–206. https://doi.org/10.5195/jmla.2017.88

[28] R. D. Peng. 2011. Reproducible Research in Computational Science. *Science* 334, 6060 (2011), 1226–1227. https://doi.org/10.1126/science.1213847

[29] Science. 2020. *Science Journals: editorial policies.* https://www.sciencemag.org/authors/science-journals-editorial-policies#data-deposition

[30] L. C. Smith. 1976. Artificial intelligence in information retrieval systems. *Information Processing & Management* 12, 3 (1976), 189 – 222. https://doi.org/10.1016/0306-4573(76)90005-4

[31] V. Stodden. 2009. The Legal Framework for Reproducible Scientific Research: Licensing and Copyright. *Computing in Science Engineering* 11, 1 (Jan. 2009), 35–40. https://doi.org/10.1109/MCSE.2009.19

[32] V. Stodden. 2010. Open Science: Policy Implications for the Evolving Phenomenon of User-Led Scientific Innovation. *Journal of Science Communication* 9, 1 (2010).

[33] V. Stodden. 2010. The Scientific Method in Practice: Reproducibility in the Computational Sciences. *MIT Sloan Research Paper* 4773-10 (2010).

[34] V. Stodden. 2014. Intellectual Property and Computational Science. In *Opening Science: The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing*, S. Bartling and S. Friesike (Eds.). Springer International Publishing, 225–235. https://doi.org/10.1007/978-3-319-00026-8_15

[35] V. Stodden. 2014. What Computational Scientists Need to Know about Intellectual Property Law: A Primer. In *Implementing Reproducible Research*, V. Stodden, F. Leisch, and R. D. Peng (Eds.). CRC Press.

[36] V. Stodden, P. Guo, and Z. Ma. 2013. Toward reproducible computational research: an empirical analysis of data and code policy adoption by journals. *PLOS ONE* 8, 6 (21 June 2013), e67111. https://doi.org/10.1371/journal.pone.0067111

[37] V. Stodden, F. Leisch, and R. D. Peng. 2014. *Implementing Reproducible Research.* CRC Press.

[38] V. Stodden, M. McNutt, D. H. Bailey, E. Deelman, Y. Gil, B. Hanson, M. A. Heroux, J. P.A. Ioannidis, and M. Taufer. 2016. Enhancing reproducibility for computational methods. *Science* 354, 6317 (2016), 1240–1241. https://doi.org/10.1126/science.aah6168

[39] V. Stodden and S. Miguez. 2013. Best Practices for Computational Science: Software Infrastructure and Environments for Reproducible and Extensible Research. *JORS* (Sep 2013). https://doi.org/10.2139/ssrn.2322276

[40] V. Stodden, J. Seiler, and Z. Ma. 2018. An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences* 115, 11 (2018), 2584–2589. https://doi.org/10.1073/pnas.1708290115

[41] A. Surkis and K. Read. 2015. Research data management. *Journal of the Medical Library Association : JMLA* 103, 3 (07 2015), 154–156. https://doi.org/10.3163/1536-5050.103.3.011

[42] J. C. Wallis, E. Rolando, and Borgman C. L. 2013. If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. *PLOS ONE* 8, 7 (2013), e67332.