# ENABLING REPRODUCIBLE RESEARCH: OPEN LICENSING FOR SCIENTIFIC INNOVATION

*Victoria Stodden*[*]

ABSTRACT

*There is a gap in the current licensing and copyright structure for the growing number of scientists releasing their research publicly, particularly on the Internet. Scientific research produces more scholarship than the final paper: for example, the code, data structures, experimental design and parameters, documentation, and figures, are all important both for communication of the scholarship and replication of the results. US copyright law is a barrier to the sharing of scientific scholarship since it establishes exclusive rights for creators over their work, thereby limiting the ability of others to copy, use, build upon, or alter the research. This is precisely opposite to prevailing scientific norms, which provide that results be replicated before accepted as knowledge, and that scientific understanding be built upon previous discoveries for which authorship recognition is given. In accordance with these norms and to encourage the release of all scientific scholarship, I propose the Reproducible Research Standard (RRS) to both rescind the default copyright on scientific work and ensure attribution. Using the RRS on all components of scientific scholarship will encourage reproducible scientific investigation, facilitate greater collaboration, and promote engagement of the larger community in scientific learning and discovery.*

[*] Research Fellow, Berkman Center for Internet and Society, Harvard Law School; M.L.S. Stanford Law School; Ph.D., M.S. Stanford University (statistics); M.A. University of British Columbia (economics). I am very grateful for invaluable discussion with Miriam Bitton, David Donoho, Danny Hillis, Larry Lessig, John Palfrey, and John Wilbanks. Of course, any errors are mine alone.

CONTENTS

## INTRODUCTION

While researchers often publish papers in academic journals describing their work and summarizing their findings, it is rare they publish their entire research product. Most of the components necessary for reproduction of the results and for building upon the research – the code and parameters used, the dataset and acquisition details, documentation, and any meta-knowledge used in the experiment – usually remain unpublished. This may be due to strict space limitations in journals and conference proceedings, legal restrictions,[1] or perhaps a lag in academic expectations behind technological changes, but the problem is serious since this practice is counter to fundamental scientific principles which state that any finding be reproducible before it becomes accepted as a genuine contribution to our knowledge base.[2]

[1] See Victoria Stodden, "The Legal Framework for Reproducible Scientific Research: Licensing and Copyright," Computing in Science and Engineering, January, 11(1), 2009, p 35-40. Available at http://scitation.aip.org/getabs/servlet/GetabsServlet?prog=normal&id=CSENFA000011000001000035000001&idtype=cvips&gifs=Yes&type%20=ALERT (last accessed Jan 9, 2009).

[2] Jon Claerbout, Green Professor of Geophysics at Stanford, goes further and calls for research to be "*really* reproducible." Paraphrased by Donoho, Claerbout advocates that "[a]n article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures." Jonathan Buckheit and D. Donoho, "WaveLab and Reproducible Research,"

As computational research becomes more pervasive and details of the computations remain unpublished, the opportunity to hide poor scholarship increases. Without full publication of "a careful description of the methods used, in sufficient detail that others can attempt to repeat the experiment," computational research could end up undermining the scientific process and becoming "the last refuge of the scientific scoundrel."[3] Perhaps more importantly, reproducibility permits open access to scientific scholarship in such a way that allows researchers outside the traditional university setting to replicate results and understand their construction. This means the broadening of access to scientific know-how beyond the traditional researcher to anyone with an Internet connection and the ability to run the routines or understand the scripts.

Working reproducibly is not just an altruistic exercise for the researcher – very often there are many small details involved in the production of research results and without accurate documentation the researcher himself can forget how a particular outcome was reached. Reproducibility can also benefit the researcher by making her work fully accessible to potential co-

---

1995. Available at http://www-stat.stanford.edu/~donoho/Reports/1995/wavelab.pdf (last accessed Jan 5, 2009). See also http://sepwww.stanford.edu/doku.php?id=sep:research:reproducible (last accessed Jan 5, 2009).

[3] R. J. LeVeque, "Wave propagation software, computational science, and reproducible research," in Proc. International Congress of Mathematicians, Madrid, Spain, 2006. See also, P. Vandewalle, G. Barrenetxea, I. Jovanovic, A. Ridol, and M. Vetterli, Experiences With Reproducible Research in Various Facets of Signal Processing Research

authors, future collaborators or employers, students and post-docs and other researchers in the area.[4]

There is another reason to promote reproducibility: It is often required. In 2004 National Science Foundation (NSF) grants comprised 64% of total academic research and development support, and that proportion is increasing.[5] The NSF requires data and other supporting materials for any research it funds to be made available to other researchers at no more than incremental cost.[6] Publishing the complete research product will accelerate the pace of research in the field and benefit the scientist: open research is built upon and cited more frequently than work published in closed

---

http://infoscience.epfl.ch/record/97195/files/ (last accessed Jan 5, 2009).

[4] D. Donoho et al "Reproducible Research in Computational Harmonic Analysis," Computing in Science and Engineering, January, 11(1), 2009, p. 8-18. Available at http://scitation.aip.org/getabs/servlet/GetabsServlet?prog=normal&id=CSENFA000011000001000008000001&idtype=cvips&gifs=Yes (last accessed Jan 9, 2009). See also J. Ioannidis, "Why Most Published Research Findings are False" PLoS Med 2(8): e124, August 2005, Available at http://medicine.plosjournals.org/perlserv/?request=get-document&doi=10.1371%2Fjournal.pmed.0020124&ct=1 (last accessed Jan 4, 2009).

[5] Rhonda Britt, "Industrial Funding of Academic R&D Continues to Decline in FY 2004," National Science Foundation InfoBrief, NSF 06-315, April 2006. Available at http://www.nsf.gov/statistics/infbrief/nsf06315/nsf06315.pdf (last accessed Jan 5, 2009).

[6] **38. Sharing of Findings, Data, and Other Research Products**

a. NSF expects significant findings from research and education activities it supports to be promptly submitted for publication, with authorship that accurately reflects the contributions of those involved. It expects investigators to share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work. It also encourages grantees to share software and inventions or otherwise act to make the innovations they embody widely useful and usable.

National Science Foundation (NSF) Grant General Conditions (GC-1), June 1, 2007. Available at http://www.nsf.gov/pubs/policydocs/gc1_607.pdf (last accessed Sept. 4, 2007).

journals.[7]

Research based on computerized analysis is an increasingly important component of a growing number of fields, including computer science, statistics, many areas of engineering and the social sciences, as well as the traditional sciences such as biology or physics. For example, in the June 1996 issue of the flagship Journal of the American Statistical Association nine of twenty articles were computational, and in the June 2006 issue 33 of 35 were.

In this paper, I argue an appropriate licensing structure will encourage researchers to create fully reproducible research by allowing them to capture more of the credit for facilitating and expanding scientific understanding, while promoting the scientific ideal of reproducible research. I propose such a structure, called the Reproducible Research Standard or RRS.

Part I of this article establishes the current scientific landscape: Defining reproducible research, making clear precisely which research components are in need of protection, and describing the current mechanisms for research

[7] See e.g. Hajjem, C. and Harnad, S. "The Open Access Citation Advantage: Quality Advantage or Quality Bias?" available at http://eprints.ecs.soton.ac.uk/13328/ (last accessed July 17, 2008).

dissemination. Part II describes the need for a reproducible research standard that can align individual incentives for openness with society's interest in promoting scientific discovery, in particular how copyright can hamper the sharing of research, and how attribution can encourage it. Part III explains the components of the Reproducible Research Standard and discusses its rationale, including the importance of the attribution component of the licensing structure and the role of Share Alike in the scientific context. I also outline ongoing joint work with Science Commons to establish a recognizable Reproducible Research Standard. Part IV discusses the costs and benefits of the RRS.

## REPRODUCIBLE RESEARCH: A GOAL OF SCIENTIFIC INQUIRY

There is a groundswell of support for reproducible research and a discussion follows about how existing regulatory bodies support and adopt this notion, following a description of the concepts of the scientific research product and reproducible research.

### *A. The Scientific Research Product*

With ever cheaper computing power and disk space, and the increasingly ease at which we collect data, many research fields are turning to computational research to advance understanding of their subject.

Computational research can be as simple as standard statistical analysis of a well-understood dataset, or as detailed as the testing of complex new algorithms on comprehensive and standardized testbeds. Gentleman and Lang introduced the term compendium to describe all components of the research that are necessary for others to understand and replicate the research.[8] Computational research is widely varied but these research components remain the same. They are:

**a. The Research Paper.**

**a.1)** If included in a compiled format, such as pdf, then include the source files (TeX, Word, or WordPerfect files for example).

**b. The Data:**

**b.1)** The data itself.

**b.2)** Documentation completely describing the data so that a researcher in the same field could make use of it: Sources, components, and possibly interpretation.

**b.3)** A description of how the data was brought into the form used in the research, including any selection and arrangement of the data, cleaning methods, or processing of variables in preparation for

[8] Robert Gentleman and D. T. Lang, "Statistical Analyses and Reproducible Research," Bioconductor Project Working Papers, paper 2, 2004. Available at http://www.bepress.com/bioconductor/paper2 (last accessed Jan 5, 2009).

analysis.

**b.4)** The code and instructions used to bring the data into the form used in the research.

**b.5)** Documentation of any code used in this process.

**c. The Experiment:**

**c.1)** The code and instructions used in the experiment, including all source code.

**c.2)** Documentation of any code used, including pseudocode and algorithm descriptions.

**c.3)** A clear listing of the parameters, settings, and conditions under which the code was used to achieve the results described in the paper, including software, platform, and computing environment.

**c.4)** A clear description of the experimental methodology.

**d. Results of the Experiment:**

**d.1)** Any figures, data, or the like produced by the code from the experiment. These can appear in full, as produced by the experiment and described in the research paper, (ie. high resolution figures) since it is often not possible to include them in the research paper directly.

**d.2)** Documentation and explanation of the experimental results.

**e. Auxiliary material:**

**e.1)** Code used for presentation on the web or an interface to the data or results.

**e.2)** Documentation of auxiliary code.

Typically the compiled paper alone is all that is released. This makes it unnecessarily difficult for other researchers to fully reproduce and understand the published results, and thus build on scientific discoveries.

Releasing the data itself is vital to scientific progress but is typically not useful without a clear understanding of how the dataset was built and what methodologies were employed in the construction of the dataset (ie. points **2.2** – **2.5** above). I label these components meta-data: All information necessary to make clear how to replicate the data used in the generation of the new results. This includes providing the original sources and collection process for the data or code that generated the dataset, and the enumeration any changes made to the dataset. Although the data do not fall under copyright, such meta-data and original selection and arrangement of the data do,[9] and its protection is vital for the success of reproducible research.

*B. Current Mechanisms for Dissemination of Scientific Research*

For a scientific researcher, success is often measured by the impact his

or her research has on the community. This is usually gauged by the number of citations an author's work receives and the level of prestige of the journals in which the scientist's works appear. This means than scientists, especially young scientists seeking tenure, are under pressure to publish in top journals that spur a large amount of future research, and to do so frequently.

Most journals in which such a scientist would publish are closed, in the sense that they operate via subscription and cannot be accessed by those who have not paid subscription fees. Usually subscriptions costs are borne by libraries at academic institutions and a researcher's affiliation with an institution gives him or her access to the contents of these journals. Calling these journals "closed" is appropriate in the sense that people not affiliated with an academic institution with a subscription usually won't have access to the research papers.[10] This means that to access and contribute to scientific research, one effectively needs to be part of a, typically small, network of researchers established in academic institutions.

The Open Access Movement has started to change this dynamic – a number of new journals, most notably the Public Library of Science (PLoS)

---

[9] See Section II.

[10] The subscription fees for a journal are often in the tens of thousands of dollars per year. For example, the cost per published page in a closed for profit journal is 6 times than

series, operate under a different business model. They provide free access to the journal contents over the Internet, and charge each publishing author for the privilege of publishing in their particular journal. This fee to publish changes according to the prestige level of the journal. At the moment, the open access journals are not so numerous and don't command the level of prestige of some of the traditional closed journals. Some traditionally closed journals, such as Nature, Science, require the release of the data into the public domain upon publication. These gradual shifts in support of full revealing of the scientific research compendium seem to be gaining momentum, as does the shift on the part of scientists to publishing in open access journals.

This shift can also be seen in the popularity of the academic search tool Google Scholar. Google Scholar indexes scholarly literature, and if a link to the full text of the paper exists anyone with access to the Internet can download these papers. Although papers published in closed journals are usually not available through Google Scholar, the research landscape is changing in favor of those papers that are easily accessible.[11] The Social Science Research Network (SSRN) and the Berkeley Electronic Press

---

of an open non-profit journal. See carl and ted bergstrom.

[11] As one assistant professor at a prestigious research university recently told me, "If I can't access the paper using Google Scholar it doesn't exist."

(bepress) are common ways for legal and social science researchers to make preprints publicly available for comment and feedback.[12] In physics, and to a lesser degree in mathematics, computer science and electrical engineering, papers are routinely posted at arXiv.org with almost no peer review (there is a basic check for relevance and a decision is made as to which category to post the paper). In fact researchers in this area often post papers only to arXiv.org with no intention of ever publishing in a traditional journal. arXiv.org is widely read by researchers in the relevant fields often because results are available much more quickly than through conventional publication mechanisms. In October of 2008, arXiv.org announced that it posts roughly five thousand papers per month.[13]

But the change is slow – academic researchers continue to be rewarded for the prestige level of the journals where they publish and the number of citations they garner. In general scientists do not release their data or code and there are no widely accepted platforms for general code and data release.[14] Some mechanisms for data sharing exist but usually through the efforts of a small team or dedicated researchers in specialized areas, such as

[12] See http://www.ssrn.com and http://www.bepress.com (last accessed Dec 28, 2008).

[13] http://communications.library.cornell.edu/com/news/PressReleases/arXiv-milestone.cfm (last accessed Jan 5, 2009).

[14] Google's program to house author donated research datasets was cancelled Dec 18, 2008. See http://blog.wired.com/wiredscience/2008/12/googlescienceda.html and

the Stanford Microarray Database (SMD).[15] Paul Caron has predicted a "long tail" effect for legal scholarship – he notes a broadening of citations to more than the 40% of SSRN papers now cited, and a weakening of the concentration of download activity in the very top papers (new downloads in 2006 focus disproportionately on papers in the bottom third of total downloads).[16] It seems reasonable to hypothesize that this pattern extends more widely than the legal research community Caron discusses, suggesting a percolation of more new ideas into more readers' hands.

Some reasons for scientists' reticence to share their entire research compendium publicly appear to be that it takes time to prepare the code and data for release, it is not required for publication in the journal, and the fear that other scientists may be able to "steal" results from the released work before the author can publish them. As discussed later in the paper, the Reproducible Research Standard can address some of these concerns, helping to alter the landscape in favor of scientific openness.

---

http://blog.wired.com/wiredscience/2008/01/google-to-provi.html (last accessed Jan 5, 2009).

[15] See Janos Demeter et al. "The Stanford Microarray Database: implementation of new analysis tools and open source release of software", Nucleic Acids Res. 2007 January; 35(Database issue): D766–D770. (available online at http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1781111). See also ttp://smd.stanford.edu/ (last access Dec 28, 2008).

[16] Paul Caron, The Long Tail of Legal Scholarship, Yale Law Journal Pocket Part. Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=944233 (last accessed Dec 28, 2008).

### *C. Reproducible Research Defined*

Jon Claerbout, a Stanford geophysics professor, advocates that "[a]n article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures."[17] This encapsulates the idea of reproducible research: the notion that results are independently replicable, given an appropriate computer platform. Although the community of scientists who engage in reproducible research is small, one study has shown that papers available online are cited at three times the rate those not available online.[18]

### *D. The Groundswell*

The Internet is becoming the dominant way for researchers to communicate and publicize their research, and in light of the increasing pervasiveness of Internet publishing, the standards for scientific research are changing.

---

[17] Jonathan Buckheit and D. Donoho, "WaveLab and Reproducible Research," 1995. Available at http://www-stat.stanford.edu/~donoho/Reports/1995/wavelab.pdf (last accessed Jan 5, 2009). See also http://sepwww.stanford.edu/doku.php?id=sep:research:reproducible (last accessed Jan 5, 2009),

[18] S. Lawrence, "Free online availability substantially increases a paper's impact," *Nature*, vol. 411, no. 6837, pp. 521, 2001, http://www.nature.com/nature/debates/e-access/Articles/lawrence.html (last accessed Jan 5, 2009).

Demands for openness of data and research are growing. In June 2007, the OECD announced the Istanbul Declaration, calling for governments to make their data freely available online as a "public good." There is now an archive site for scientific research papers.[19] The Open Archives Initiative and Science Commons are proposing universal standards for data repositories to facilitate reproducibility and novel scientific research.[20] Companies such as Metaweb and Google are creating new web structures to unify the housing of complex data.[21] Some research labs carry out reproducible research as a policy and this number is growing.[22] Similarly a growing number of papers have been published recently calling for reproducible research.[23] In July of 2007, Microsoft held a Research Faculty Summit discussing reproducible research.[24] If passed, the Federal Research Public Access Act will require that 11 U.S. government agencies with annual extramural research expenditures over $100 million make manuscripts of

[19] http://www.arxiv.org/ Open access to 515,515 e-prints in Physics, Mathematics, Computer Science, Quantitative Biology and Statistics (last accessed Jan 5, 2009).

[20] http://www.openarchives.org/ (last accessed Jan 5, 2009).

[21] See http://www.freebase.com/ and http://www.google.com/base (both last accessed Jan 5, 2009).

[22] Although it is still very small. See http://sepwww.stanford.edu/, the Donoho group at http://www-stat.stanford.edu/~donoho, and http://lcavwww.epfl.ch/ for a few examples.

[23] See Gentleman, R., & Lang, D. T. Statistical analyses and reproducible research. Bioconductor Project Working Papers, May 2004; and Giovanni Baiocchi, Reproducible research in computational economics: guidelines, integrated approaches, and open source software, Computational Economics, Volume 30, Issue 1, August 2007.

[24] See http://research.microsoft.com/en-us/um/redmond/events/fs2007/agenda_mon.aspx (last accessed Jan 5, 2009).

journal articles stemming from research funded by that agency publicly available via the Internet within 6 months of publication. On February 12, 2008, Harvard University's faculty of arts and sciences adopted a policy that requires faculty members to let the university make their scholarly articles available freely online (rights are turned over to the university, nonexclusively):

> Each Faculty member grants to the President and Fellows of Harvard College permission to make available his or her scholarly articles and to exercise the copyright in those articles. In legal terms, the permission granted by each Faculty member is a nonexclusive, irrevocable, paid-up, worldwide license to exercise any and all rights under copyright relating to each of his or her scholarly articles, in any medium, and to authorize others to do the same, provided that the articles aren't sold for a profit.[25]

The faculty members must provide their manuscript to the university for deposit into the open access repository within a year of publication.[26] Stanford's School of Education followed suit with a mandate for open access: All faculty members must make a copy of their published work

---

[25] See http://www.fas.harvard.edu/~secfas/February_2008_Agenda.pdf, p. 3 (last accessed Jan 5, 2009).

[26] See also http://blog.stodden.net/2008/11/23/stuart-shieber-and-the-future-of-open-access-publishing/ (last accessed Dec 28, 2008).

available in an open access repository as of July 26, 2008.[27]

On September 20, 2007, the National Science Foundation (NSF) released a major new initiative on Cyber-enabled Discovery and Innovation (CDI).[28] The initiative is meant to foster American competitiveness through research contributing to "a new generation of computationally based discovery concepts and tools to deal with complex, data-rich, and interacting systems." The goals the NSF states encourage all of: Data mining of large sets; Interacting complex systems; High-performance computational experimentation; Virtual environments; and Educating researchers and students in computational discovery.

In 2007, the National Institutes for Health (NIH) mandated that research it funds becomes "available in a timely fashion to other scientists, health care providers, students, teachers, and the many millions of Americans searching the web to obtain credible health-related information."[29] The NIH envisions a searchable database of NIH funded publications.

Paul Huber has been advancing open access to research articles and their preprints, free of copyright and licensing restrictions.[30] He advocates

---

[27] See http://www.news.harvard.edu/gazette/2008/02.14/99-fasvote.html and http://www.earlham.edu/~peters/fos/2008/06/oa-mandate-at-stanford-school-of-ed.html (last accessed Dec 28, 2008).

[28] See http://mathinstitutes.org/cdi/ (last accessed Dec 28, 2008)

[29] http://grants1.nih.gov/grants/guide/notice-files/NOT-OD-04-064.html (last accessed Jan 5, 2009).

[30] http://www.earlham.edu/%7Epeters/fos/overview.htm (last accessed Jan 5, 2009).

the explicit use of Creative Commons licenses for the research papers or a similar licensing structure that allows the copyright holder to "consent in advance to let users "copy, use, distribute, transmit and display the work publicly and to make and distribute derivative works, in any digital medium for any responsible purpose, subject to proper attribution of authorship...."""[31] The Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities has been signed by 242 organizations including universities and advocacy groups such as the Open Society Institute.[32]

### *E. Reasons to Perform Reproducible Research*

Knowing your work will be fully open to inspection in the future creates an incentive for researchers to do better, more careful, science now. For example it will prevent any temptation, even unconscious, to substitute the more impressive looking figures into a paper that may be a slight mismatch with the accompanying methodological description.[33] Scientists operating under the principle of reproducible research will be able to reproduce their own work as they carry out further research and ensure the

[31] The Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities, Oct 20-22, 2003. http://oa.mpg.de/openaccess-berlin/berlindeclaration.html (last accessed Jan 5, 2009).

[32] http://oa.mpg.de/openaccess-berlin/signatories.html (last accessed Jan 5, 2009).

[33] See e.g. J. Young, "Journals Find Fakery in Many Images Submitted to Support Research" The Chronicle of Higher Education, May 29, 2008. Available at

accuracy of descriptions of their work. This could preserve valuable work: One researcher tells the story of losing figures that had not been created in a reproducible way before publication and, because of time constraints and expense, being forced to abandon publication of compelling results.[34]

Generation of results often requires a detailed knowledge of parameters and software invocation sequences. Without a clear description it can be next to impossible, even for the original scientist, to try the published methodology in a new setting or on a new dataset. Every publishing author hopes his or her new method will be of broader use than just that single published paper, and reproducible research helps ensure that possibility.

Building on research is more difficult without a full understanding of what has been done previously. Reproducible research makes it possible for researchers to communicate to others the difficulties they might be having in their work and for others to contribute to solutions.

By making the entire research compendium publicly available, scientists not in the immediate field of research can download, modify, and apply the work, thereby facilitating interdisciplinary research and collaboration. This

---

http://chronicle.com/free/2008/05/3028n.htm (last accessed July 18, 2008).

[34] Buckheit and Donoho, "WaveLab and Reproducible Research," at 2. Available at http://www-stat.stanford.edu/~donoho/Reports/1995/wavelab.pdf (last accessed Jan 5,

access to complete information may satisfy a basic need, or even "spiritual necessity," among independent scientists to understand scientific regions "as a whole, and to lend one another strength of that understanding."[35]

Science Commons suggests that "the legal questions – how can an author make her work available to the public, while taking comfort that she retains some rights to it - have yet to be answered."[36]

## THE RATIONALE FOR THE RRS (REPRODUCIBLE RESEARCH STANDARD):

## ALIGNING INCENTIVES

Open standards and open access alone are insufficient to promote the free discovery and development of science since scientific success is measured by citations, ie. the amount of subsequent work he or she engenders. This reward system can create short-sighted incentives not to release the full research compendium both in the belief that other scientists will "steal" future publications by building on released work or data without attribution, and because this takes time away from the next publication.[37]

---

2009).

[35] Norbert Weiner, Cybernetics, 2nd ed. MIT Press, 1965, at 3.

[36] Id.

[37] "The WIPO Copyright Treaty (WCT), concluded in 1996, recognizes "the need to maintain a balance between the rights of authors and the larger public interest, particularly education, research and access to information" in updating international copyright norms to respond to challenges arising from advances in information and communications technologies, including global digital networks.1" WIPO Copyright Treaty, CRNR/DC/96 (Dec. 20, 1996) (quoting preamble). http://people.ischool.berkeley.edu/~pam/papers/reversenoticebtljfinal.pdf at 2. (last accessed Jan 5, 2009).

An attribution license for all elements of the research compendium (excepting the raw data) that will perpetuate virally through all derivative works, thereby ensuring correct attribution for all parts of the research, can address a scientist's fear of theft of unpublished future results. By "virally" I mean that work, once licensed, retains the same license even when it is used in subsequent work. For example, if I develop code implementing my new image compression technique, and another researcher makes use of this code in his or her subsequent research, my code becomes part of his or her researcher compendium but retains the original license that attributes the code to me. Scientific research is, in effect, a re-mixing of different pieces of previous research, whether ideas, code, data, or other structures.

As a corollary to maintaining attribution, the Reproducible Research Standard provides a guide that makes the release of the complete research product as easy and as useful to others as possible. An appropriate attribution license, as suggested by the Reproducible Research Standard, will allow the scientist to rescind the default copyright status of their work and permit others to legally access and use the work.

### *A. The Production and Dissemination of Scientific Knowledge and Default Copyright*

Copyright is assigned automatically to the authors whenever a

creative work is produced: "copyright follows the author's pen across the page."[38] By default, control of both copying the work and using in another creative or scientific endeavor is retained by the original authors, thus hampering replication of original scientific results. Copyright stands in the way of the reuse of scientific scholarship in new scientific discoveries.

Articles released on the web by default have copyright assigned to the original authors, unless the authors have taken steps to rescind or transfer the copyright. Typically, published articles require the transfer of copyright to the publishing journal making it impossible to legally access the journal article without a subscription to the journal itself.

THE COMPONENTS OF THE REPRODUCIBLE RESEARCH STANDARD

The Reproducible Research Standard (RSS) sets forth the steps a scientist can take to ensure his or her work is recognized as really reproducible. The first component of the standard is applying an appropriate license to rescind copyright from elements of the research compendium. The RRS suggests using the **Creative Commons BY** license for the media components of the compendium, and the **Modified BSD** license[39] for code components, and if the scientist chooses to release his or

[38] E. von Hippel, Democratizing Innovation, MIT Press, 2005 at 85.

[39] Since the release of the "Modified" or "Simplified BSD License" in January 2008 the BSD license is roughly equivalent to the MIT License. See

her data to the public domain, attaching the Science Commons Database Protocol[40] to the data.

The CC BY license is designed for media: to "share your creations with others and use music, movies, images, and text online that's been marked with a Creative Commons license." If used on the entire research compendium, it is misapplied since it does not adequately cover code and, in fact, using the CC BY license for code is actively discouraged by Creative Commons.[41] Using the Modified BSD license alone for scientific compendia leaves the documentation, figures, final paper and other forms of text-based scholarship, the experimental design, GUIs interfacing with the algorithms, pseudocode, or dataset build methodologies for example, without an adequate license. But all of these works could be released appropriately under the CC BY license that ensures consistent viral attribution for the entire compendia.[42] Why the RRS recommends CC BY and the Modified BSD license is discussed in the forthcoming sections.

### *A. Attribution and a Principle for Scientific Licensing*

---

http://www.opensource.org/licenses/bsd-license.php (last accessed Jan 5, 2009).

[40] See http://sciencecommons.org/resources/faq/database-protocol/ (last accessed Jan 5, 2009).

[41] "[W]e do not recommend that you apply a Creative Commons license to software code." http://wiki.creativecommons.org/FAQ (last accessed Jan 5, 2009).

[42] See Section C: "Rescinding Copyright" for a discussion of various licensing

Attribution is a core mechanism by which scientific research progresses and it undergirds the traditional system under which ideas and research output are shared. Permitting an attribution component in the licensing structure of research compendia is aligned with these longstanding scientific values. I argue the attribution component is so closely aligned with scientific norms that it justifies not simply calling for the entire research compendia to be released to the public domain. Of course, a scientist is free to choose to release their compendia to the public domain, but creating an option for legal attribution would further encourage the release of scientific scholarship.[43] Research work qualifies as reproducible under the Reproducible Research Standard whether it is released into the public domain, or whether it uses the attributive licensing structure recommended in this paper.

The particular selection of licenses under the RRS allows for viral attribution in that any element of such a compendium that is reused in other's work, such as code that compresses images in a certain way or a particular method for cleaning a microarray dataset, must retain the original attribution. In practice this means that the reused code is still under the

---

options.

[43] For a discussion of the relationship between legal attribution and scientific citation see duLong and Stodden, 2009, forthcoming.

Modified BSD license and must be attributed to the original author. Furthermore, research that builds upon this subsequent work must also retain the Modified BSD license on the particular piece of code written by the original author, thus giving attribution to the original creator and contributor of the code no matter what downstream compendia it find itself in. This is the way in which the licensing structure is "viral," in that attribution propagates through downstream scientific research. This mirrors how scientific work is typically cited and built upon, with the exception that the attribution process is formalized in a legal license in a machine readable way.[44]

The RRS aligns the scientist's desire for citation with the interests of the larger community in the openness and availability of scientific research work. This also suggests a guiding principle for scientific licensing:

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

[44] See the Costs and Benefits Section for a discussion of the encoding of attribution for research compendia under the RRS.

Attribution provided by the CC BY and Modified BSD licenses satisfies such a principle. I further argue that the Share Alike aspect common to many licenses does not.

### *B. Share Alike in the Scientific Context*

The licensing structure under the RRS ensures each scientist is attributed only for the work he or she has authored. The Share Alike component is found some licenses (for example the GNU GPL license and a number of Creative Commons licenses) and specifies that the use of, say, GPL-licensed code, in the development of another body of code, will bring the entire derivative body of code under the GPL unless an alternative is negotiated with the original's copyright holders.[45] "If you alter, transform, or build upon this work, you may distribute the resulting work only under a license identical to this one."[46] Whenever work licensed under a Share Alike license is used, the resulting work must carry the same license. This means that only Share Alike or Public Domain content can now be incorporated into the work, since it must be distributed as Share Alike.

[45] If there are many authors, for example in a scientific work that builds on many previous results, this negotiation could become practically impossible. See Micheal Heller, The Gridlock Economy: How Too Much Ownership Wrecks Markets, Stops Innovation, and Costs Lives, Basic Books, 2008.

[46] http://fr.creativecommons.org/articles/sweden.htm (last accessed Jan 1, 2009).

The Share Alike concept is inappropriate in the scientific context because it can impose limits on the use and reuse of others' work. This would render the research unavailable for use by people who prefer not to use the same license as the original research for their own resulting research compendia. Ideally, downstream researchers will choose to license the original components of their compendia so that researchers, even those working within a proprietary context, can build upon the work without legal encumbrance, but scientific research should be encouraged over particular license use. Furthermore, expanding the license to cover the entire derivative work product makes attempts at attribution more difficult. Under Share Alike, it's no longer clear how to give credit to upstream work in a derivative product because a single attribution scheme could subsume and conflate work by different authors. Scientists should be free to license their compendia's components as they see fit, and they shouldn't be restricted in licensing, say, the figures from their work in a particular way because they used or modified another researcher's code to build them. The Science Commons Open Access Data Protocol embodies many of these arguments.[47]

[47] http://sciencecommons.org/projects/publishing/open-access-data-protocol/ (last accessed Jan 5, 2009).

An implication of the Principle of Scientific Licensing is that there must be the lowest bar possible to building upon previous scientific research. A corollary benefit to the RRS's relaxation of the Share Alike component is that it becomes easier for startups to employ the research as part of their technology without having all their (possibly) proprietary work come under a particular license.

### *C. Rescinding Copyright on Research Work: Licenses*

A plethora of licenses exist that allow authors to rescind copyright from components of his or her research compendium. This section highlights some of the most popular ones and discusses reasons for and against their use, in contrast with the Modified BSD and the CC BY licenses suggested by the RRS.

A popular license for code is the GNU General Public License (GPL).[48] The license has two main components:

1. If publicly distributed, all software subject to the license must also have its source code released, and
2. Once the license is attached to code, it also attaches to any body of code that uses the original code.

The license has the Share Alike provision and is therefore not suitable for

scientific works, as discussed in the previous section.

Less frequently used is the GNU Lesser General Public License (LGPL).[49] It was developed for code libraries and permits their use in proprietary packages, which the GPL does not.

The Apache 2.0 license is another common alternative method for rescinding copyright from code.[50] The Apache license does not contain the Share Alike provision and permits the exercise of patent rights that would otherwise only extend to the original licensor, provided the person modifying the code does not sue the original licensor for patent infringement.

The (Modified) Berkeley Software Distribution (BSD) license permits the downstream use, copying, and distribution of either unmodified or modified source code, so long as the license accompanies any distributed code and the previous authors' names are not used to promote modified downstream code.[51] There is no Share Alike provision, meaning that code licensed under the BSD can be incorporated into proprietary or Share Alike licensed work such as that licensed by the GPL, or co-mingled with closed

[48] http://www.gnu.org/licenses/gpl.html (last accessed Jan 1, 2009).

[49] Use of this license is discouraged by the Free Software Foundation who developed it. See http://www.gnu.org/licenses/lgpl.html (last accessed Jan 1, 2009).

[50] http://www.apache.org/licenses/LICENSE-2.0 (last accessed Jan 1, 2009).

[51] http://www.opensource.org/licenses/bsd-license.php (last accessed Jan 2, 2009).

code.[52]

### *D. Attribution of Data*

Collecting, cleaning, and otherwise preparing data for analysis is often a significant component of scientific research. Copyright law in the U.S. does not permit the copyrighting of "raw facts" but original products derived from those facts can be and are, in fact, copyrightable. In Feist Publications, Inc. v. Rural Telephone Service, the Court found that white pages telephone directories are not themselves directly copyrightable; copyrightable works must have creative originality:[53]

> . . . the copyright in a factual compilation is thin. Notwithstanding a valid copyright, a subsequent compiler remains free to use the facts contained in another's publication to aid in preparing a competing work, so long as the competing work does not feature the same selection and arrangement.[54]

---

[52] The term "Modified" refers to the January 9, 2008 version of the BSD license: The original BSD license contained an advertising or endorsement clause which required the licensee to acknowledge use of U.C. Berkeley code in any advertising of a product using that code. This clause was officially rescinded by the Director of the Office of Technology Licensing of the University of California on July 22nd, 1999. He stated that clause 3 is "hereby deleted in its entirety." See http://www.opensource.org/licenses/bsd-license.php (last accessed Jan 2, 2009).

[53] Feist Publ'ns Inc. v. Rural Tel. Serv. Co., 499 U.S. 340 (1991) at 363-364.

[54] Id. at 349. See also Bitton, Miriam, "A New Outlook on the Economic Dimension of the Database Protection Debate" and H. Zhu and S. Madnick, "One Size does not Fit All: Legal Protection for Non-Copyrightable Data" Working Paper CISL# 2007-04. Available at http://web.mit.edu/smadnick/www/wp/2007-04.pdf (last accessed Jan 4, 2009).

Currently the Court holds databases protectable.[55] A license that rescinds copyright from the original "selection and arrangement" of a database, in a virally attributive way, can encourage scientists to release the datasets they have compiled by providing a legal framework for attribution and reuse of their work. "Selection and arrangement" of a dataset may be formalized as either code or text descriptions and either the Modified BSD or the CC BY license respectively should be applied. Of course, since the raw facts themselves are not copyright able, it does not make sense to apply a copyright rescinding license to them. The RRS recommends the release of facts to the public domain (marked by the Science Commons CC0 standard[56]), accompanied by the attachment of the appropriate license (CC BY or Modified BSD whether text or code based, respectively) to any original selection and arrangement of the data to ensure legal attribution for dataset use. Using this licensing formulation for the original "selection and arrangement" of the data does not create a conflict with release of the raw facts to the public domain since these are separate elements of the dataset used in the scientific research.

### *E. The Reproducible Research Standard*

[55] Bitton, Miriam, "A New Outlook on the Economic Dimension of the Database Protection Debate" at 4.

[56] See http://creativecommons.org/weblog/entry/9071 (last accessed Jan 4, 2009).

The Reproducible Research Standard is a way for scientists to publicly mark their work as reproducible, meaning that certain conditions are satisfied:

1. The full compendium is available on the Internet,
2. The media components, including the original selection and arrangement of the data, are licensed under CC BY,
3. The code components are licensed under the Modified BSD license,
4. The data has been released into the public domain according to the Science Commons Open Data Protocol.

If a researcher satisfies these four criteria, his or her work can be marked as satisfying the "Gold Standard" of the RRS. The "Silver Standard" represents work that is not fully released, but the scientist promises to release the full compendium under the gold standard to anyone who asks him or her. The "Bronze Standard" is for any research work that satisfies some of the above criteria. For example, the code has been released under a Share Alike license or the data are not available.

Efforts are currently under way for the RRS to be an official mark of Science Commons. This would provide an easily identifiable logo and a clear definition for each level of reproducibility. A webpage would also be available at the Science Commons website for scientists to obtain

information on the possible licensing structures and html tags for their work.

An attribution based system of reproducible research holds the promise of encouraging scientists to release their entire research compendium on the Internet. With Science Commons' support this could become a standard for funding institutions, publishers, or collaborators who would like to require the public availability of the entire research product, as well as the scientist who wishes to make his or work visible for potential collaboration or citation. The RRS could provide cultural pressure that encourages reproducible research, and perhaps encourages journals to publish papers or grant giving organizations to fund work fully compliant with the RRS and principles of reproducible research.[57] As one researcher has pointed out, an advantage to open code and clarity of experimental method is publicity of the new work.[58]

The umbrella licensing structure of the RRS makes it easier for scientists to share their work than the alternative of each scientist evaluating all the different licensing possibilities. Without the RRS, each time he or she releases scholarship, the scientist would have to fashion together a

[57] See Jelena Kovacevic, "How to Encourage and Publish Reproducible Research" Available at http://lcav.epfl.ch/reproducible_research/ICASSP07/Kovacevic07.pdf (last accessed Jan 5, 2009).

[58] See http://infoscience.epfl.ch/record/97195/files/ (last accessed Jan 5, 2009)

combination of licenses from a spectrum of choices. Since the RRS suggests common existing licenses, there are no additional compatibility or interoperability issues with existing licenses.

## THE COSTS AND BENEFITS OF THE REPRODUCIBLE RESEARCH STANDARD

The NSF goal that publicly funded research be made publicly available achieves important objectives: accountability and oversight in the use of government funds; incentives for better research through the knowledge of future public release; promotion of scientific knowledge through both 1) direct conveyance and 2) facilitation of the opportunity to verify and improve upon answers to scientific questions. A licensing structure that can protect and promote these goals by aligning the scientific researcher's interests with society's interests in scientific research as a whole could dramatically improve participation by scientists in collaborative research, encourage citizen-scientists to actively engage in research, and institutionalize the web as the mode for release of the compendia associated scientific discovery. Such a licensing structure would also have the corollary effect of producing better science: a researcher who anticipates release of all his or her work to the public is apt to do a much

more careful job.[59]

The RRS provides a mechanism for scientists to license the meta-knowledge associated with the creation and perfecting of their data. The RRS also provides a mechanism through which this metadata that can encoded and associated with the research in a machine-readable way.[60] In fact, it provides a mechanism for handling detailed and complex multi-author attribution. Since attribution is a feature of a tag on the elements of the research compendia, adding an arbitrary number of authors in attribution tags becomes meaningful to computerized search and machine-readability.[61]

The RRS may encourage innovative ways to allow some reproducibility, such as providing an online system for other researchers to choose either algorithm parameters or specific sections of data and then just be returned processed results.[62] The RRS may encourage a change in the valuation of scientific work away from pure research results toward algorithm modification for useful purposes.[63] For example, industrial

[59] This is acknowledged by Richard Stallman when he suggests that if you develop code not under a free license, you "work on it only enough to write a paper about it, and never make a version good enough to release." http://www.gnu.org/philosophy/university.html (last accessed Jan 4, 2009).

[60] See http://wiki.creativecommons.org/CcREL (last accessed Jan 4, 2009).

[61] See duLong and Stodden, forthcoming 2009.

[62] Id.

[63] See Jelena Kovacevic, "How to Encourage and Publish Reproducible Research" Available at http://lcav.epfl.ch/reproducible_research/ICASSP07/Kovacevic07.pdf (last accessed Jan 5, 2009).

applications may become a vital part of research on the web and non-researchers may be able to use the scientific research more readily than under traditional publication methods.

Since the RRS facilitates the communication of research and ensures attribution, it avoids two of the stumbling blocks to very large scale collaboration. The Internet naturally suggests such collaboration and the RRS, by making entire research product coherently and consistently available, encourages this use of the Internet's potential. The RRS may facilitate Internet-based data sharing research models.

A researching scientist may have done more experimentation than is practical for a traditional research paper. Releasing the full research product allows for the reporting and attribution of more results, including negative results, and experimental configurations than would ordinarily be publishable.[64] This would permit scientists to see more of the full picture with regard to the research that has been done on the problem of interest.

As alluded to in the introduction, by ensuring open easy access to others' research, the RRS can stand as a bulwark against plagiarism and falsification of scientific results. If the potential exists for peers to verify all

[64] See John Ioannidis, "Why Most Published Research Findings are False," PLoS Med 2(8):e124, August 2005. Available at http://medicine.plosjournals.org/perlserv/?request=get-

your methodologies, the incentive to cheat is greatly reduced.

The RRS's licensing structure clarifies the role of third parties. This is especially important as the university is a common setting for computational research, and universities nearly always claim rights to work developed using university facilities, although are often amenable to open release of software.[65] The RRS publicly releases the compendium and appears to be compatible with university ownership rights. Most computational research work takes place in a university setting and many universities claim some ownership rights over the research product. On November 1, 2007 Katharine Ku, Director of the Office of Technology Licensing (OTL) at Stanford University, indicated the University's concern was not on copyright but focused primarily on patents. At least in Stanford's case, the OTL did not perceive any conflict between the rescinding of copyright as I am proposing and their interests as a university.

## CONCLUSION

Scientific computation appears to be taking a central role in the scientific method, but progress and credibility are stunted by the fact that

---

document&doi=10.1371/journal.pmed.0020124 (last accessed Jan 5, 2009).

[65] "… if a creator/inventor wants to put her software in the public domain so that no one has any intellectual property rights in the software, or if a creator/inventor wants to make the IP freely available, Stanford will be agreeable, so long as such an action does not conflict with any existing contractual obligations and does not create a conflict-of-interest issue." Computing Research News, Jan 2002, at 3, 8. Available at

the entire research compendia – including code and data – are not being routinely made available to others.[66] By addressing all components of the compendia, the Reproducible Research Standard encourages scientists to release all the computational details of their work. The RRS blends the attribution aspect of the Modified BSD license for the code, Creative Commons attribution protection for text, documentation, figures and other media, including dataset creation methodologies, and creates a clear standard for replicability of scientific work. The RRS licensing structure also rescinds the barrier to access that copyright creates as a default for released work. These twin pillars of the RRS are designed to achieve the scientific ideal of reproducibility and thus faciliate the extension of research results.

---

http://www.cra.org/CRN/articles/ku.html (last accessed Jan 5, 2009).

[66] See David Donoho et al, "Reproducible Research in Computational Harmonic Analysis," Computing in Science and Engineering, January, 11(1), 2009, p 8-18. Available at http://scitation.aip.org/getabs/servlet/GetabsServlet?prog=normal&id=CSENFA00001100 0001000008000001&idtype=cvips&gifs=Yes (last accessed Jan 9, 2009).