INTERNATIONAL JOURNAL OF COMMUNICATIONS LAW & POLICY

ISSUE 13 WINTER 2009

ENABLING REPRODUCIBLE RESEARCH: LICENSING FOR SCIENTIFIC INNOVATION

Victoria Stodden

TABLE OF CONTENTS

ENABLING REPRODUCIBLE RESEARCH: LICENSING FOR SCIENTIFIC INNOVATION[†]

Victoria Stodden[*]

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 both 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) both to ensure attribution and facilitate the sharing of scientific works. 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.

I. INTRODUCTION

While researchers typically publish papers in academic journals describing their work and summarizing their findings, it is rare they make their entire research product available. Most of the components necessary for reproduction of the results and for building upon the research – for example, the code and data – usually remain unpublished. The problem is serious since this practice is counter to the fundamental scientific principle that any finding be reproducible before it becomes accepted as a genuine contribution to our knowledge base. The current laxity in reporting experimental details and validating results is creating a credibility gap for computational research just as massive computation is transforming scientific enterprise.[1]

Scientific computation is becoming a core component of modern scientific inquiry. In the June 1996 issue of the flagship *Journal of the American Statistical Association* nine of twenty articles were computational, while in the June 2006 issue 33 of 35 were, and this trend is not limited to statistics. In numerous fields the increasing prevalence of data collection, computerized simulations, and deep data exploration indicate that scientific computation is becoming central to the scientific method.[2] A credibility crisis is created since it is impossible to

---

[†] This paper is the winner of the *Kaltura Writing Competition*, issued in connection with the Yale Information Society Project's Third Conference on Access to Knowledge (A2K3), Sept. 8-10, 2008.

[*] Fellow, Berkman Center for Internet and Society, Harvard Law School; Fellow, Science Commons; 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, Thinh Nguyen, John Palfrey, and John Wilbanks. Of course, any errors are mine alone.

[1] *See* David L. Donoho, Arian Maleki, Morteza Shahram, Inam Ur Rahman & Victoria Stodden, *Reproducible Research in Computational Harmonic Analysis*, 11 COMPUTING IN SCIENCE AND ENGINEERING 8 (2009) [hereinafter Donoho et al.]. *See also* Christine Laine, Steven N. Goodman, Michael E. Griswold & Harold C. Sox, *Reproducible Research: Moving Toward Research the Public Can Really Trust*, 146 ANNALS OF INTERNAL MEDICINE 450 (2007), *available at* http://www.annals.org/cgi/reprint/146/6/450 (last visited Mar 14, 2009).

[2] *See, e.g,.* Chris Anderson, *The End of Theory: The Data Deluge makes the Scientific Method Obsolete*, WIRED MAGAZINE, June 23, 2008, *available at* http://www.wired.com/science/discoveries/magazine/16-

validate and verify most of the results computational scientists publish in papers and present at conferences.[3] This paper presents a solution to the problem Copyright law poses for verification of research results.

When scientists share their research on the web, the original expression of their ideas automatically falls under copyright. Copyright law is often understood as a tradeoff between providing incentives for the production of creative works by granting the author certain limited term exclusive rights over their work, and the public's desire to access the work. By blocking the ability of others to copy and reuse research, copyright law acts counter to prevailing scientific norms that encourage scientists to openly release their work to the community in exchange for citation.[4]

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. This standard is a mechanism not only for giving scientists the ability to determine terms of use associated with their released work, but it creates a framework for research funders, universities, and scientists to speak about research in terms of reproducibility and access.

Part II of this article defines reproducible research, describes today's scientific landscape, and recounts current mechanisms for research dissemination. Part III describes the need for a reproducible research standard that can align individual scientists' incentives for openness with society's interest in promoting scientific discovery, in particular how copyright can hamper the sharing of research, and how this can be remedied through appropriate licensing structures. Attribution only licensing is discussed as a potential vehicle for the establishment of a science information commons. I also outline ongoing joint work with Science Commons to establish a recognizable *Reproducible Research Standard*. Part IV discusses the Reproducible Research Standard in practice: comparing the RRS to release into the Public Domain or the Fair Use exception to copyright, discrepancies between legal attribution and academic citation, and impact on third parties.

II. REPRODUCIBLE RESEARCH: A NECESSARY GOAL OF SCIENTIFIC INQUIRY

With ever cheaper computing power and data storage capabilities, computing is becoming central to scientific endeavors, but the prevalence of relaxed practices with regard to the verification of results is causing a crisis of credibility. It is currently impossible to reproduce and validate most of the results that computational scientists publish in papers and present at conferences.[5]

07/pb_theory (last accessed January 30, 2009). *But cf.* The Reality Club: The End of Theory, http://edge.org/discourse/the_end_of_theory.html.

[3] Donoho et al., *supra* note 1, at 8.

[4] *See* ROBERT K. MERTON, THE SOCIOLOGY OF SCIENCE: THEORETICAL AND EMPIRICAL INVESTIGATIONS 223-80 (University of Chicago Press 1973).

[5] *See id.* and B.D. McCullough, *Got Replication? The Journal of Money, Credit, and Banking Archive*, 4 ECON JOURNAL WATCH 326 (2007), *available at* http://www.econjournalwatch.org/pdf/McCulloughEconomicsInPracticeSeptember2007.pdf (last visited Mar. 14, 2009). *See also* D.B. McCullough & H.D. Vinod, *Verifying the Solution from a Nonlinear Solver: A Case Study: Reply,* 94 AM. ECON. REV. 391 (2004), *available at* http://www.tau.ac.il/~rroonn/Papers/AER2004MV1.pdf (last visited Mar. 14, 2009).

All scientific research proceeds against the ubiquity of error -- the central motivation of the scientific method is to keep error from contaminating research conclusions. Even very disciplined and cultivated branches of science can even suffer from the problem of errors in final published conclusions.[6] Vagueness, wandering attention, forgetfulness, and confusion are scourges of human reasoning. Data are often noisy, apparent patterns can turn out to be nothing but randomness, and misinterpreting calculations or mislabeling data are easy to do.[7]

Longstanding branches of scientific inquiry have established methods to guard against error: standards of proof in mathematics, or the machinery of hypothesis testing in empirical research for example. Computational science is missing a crucial opportunity to establish verification of results as its response to the ubiquity of error in scientific research.

A. The Scientific Research Product

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 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.

---

[6] J.P.A. Ioannidis, *Why Most Published Research Findings are False*, 2 PLOS MEDICINE 696 (2004).

[7] *See, e.g.,* Greg Miller, *A Scientist's Nightmare: Software Problem Leads to Five Retractions*, 314 SCIENCE 1856 (2006), *available at* http://www.sciencemag.org/cgi/content/full/314/5807/1856 (last visited Mar. 13, 2009).

[8] Robert Gentleman & Duncan Temple Lang, *Statistical Analyses and Reproducible Research* (Bioconductor Project Working Paper, Paper No. 2, 2004), *available at* at http://www.bepress.com/bioconductor/paper2 (last visited Mar 14, 2009).

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, (i.e. 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.

d.3) Auxiliary material:

d.4) Code used for presentation on the web or an interface to the data or results.

d.5) Documentation of auxiliary code.

d.6) A description of the computing platform used.

Typically the compiled paper alone is all that is released. This is usually insufficient to allow other researchers to reproduce the published results, thus creating an encumbrance for building on scientific discoveries.

Releasing of data is important to scientific progress but is typically not useful without a clear understanding of what methodologies were employed in the construction of the dataset (i.e. points *b.2 – b.5* above). These components can be labeled *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 of any changes made to the dataset. Although the raw facts themselves do not fall under copyright, such meta-data and any original selection and arrangement of the data do,[9] and this can provide a hook to attach attribution to dataset release, thereby encouraging scientists to fully engage in reproducible research.[10]

B. Current Mechanisms for Dissemination of Scientific Research

For a scientist, success is often measured by the impact his or her work has in the research community. This is typically gauged by the number of citations an author's publications receive and the level of prestige of the journals in which the scientist's works appear. Thus scientists, especially young scientists seeking tenure, are under pressure to publish top journal articles that spur a large amount of future research, and to do so frequently.

Most of these journals operate via subscription and cannot be accessed by those who have not paid subscription fees. Usually subscription 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

[9] *See infra* sect. III.A.3.

[10] Meta-data that are mere statements of raw facts are not copyrightable and thus do not support licensing structures like those embodied by Creative Commons.

research papers.[11] To access and contribute to scientific communities, one traditionally needed to be part of a 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) 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 less numerous and don't command the level of prestige of some of the traditional closed journals. Some traditionally closed journals, such as the *Nature* family of journals[12], now require the release of the data to interested readers upon publication[13] and some, such as *The New England Journal of Medicine*,[14] now provide free online access six or 12 months after publication.[15] Support for the full revealing of the scientific research compendium seems to be gaining momentum, as does the shift on the part of scientists to publishing in open access journals.

This transformation can also be seen in the popularity of the academic search tool *Google Scholar.*[16] Google Scholar indexes scholarly literature, and if the full text of the paper is available online Google Scholar provides a link to download. 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.[17] The Social Science Research Network (SSRN) and the Berkeley Electronic Press (bepress) are common ways for legal and social science researchers to make preprints publicly available for comment and feedback.[18] 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 some researchers in these areas will post papers to arXiv.org with no intention of publishing in a traditional journal. arXiv.org is widely read by researchers in the relevant fields often because

[11] 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 of an open non-profit journal. *See* Carl T. Bergstrom & Theodore C. Bergstrom, *The Costs and Benefits of Library Site Licenses for Academic Journals,* 101 PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES 897 (2004), *available* http://www.econ.ucsb.edu/~tedb/archive/BergstromAndBergstrom04.pdf (last visited Mar. 14, 2009).

[12] Including *Nature*, *Nature Cell Biology*, *Nature Chemical Biology*, *Nature Genetics*, *Nature Methods*, *Nature Neuroscience*, and *Nature Genetics*.

[13] *See* Availability of data & materials: authors & referees @ npg, http://www.nature.com/authors/editorial_policies/availability.html.

[14] Provided since 2001, with no six month waiting period for developing countries. *See* Jeffrey Drazen & Gregory Curfman, *Public Access to Biomedical Research*, 351 NEW ENGLAND JOURNAL OF MEDICINE 1343 (2004), *available at* http://content.nejm.org/cgi/content/full/351/13/1343 (last visited Mar 7, 2009).

[15] For a discussion of data revealing and the role of journals in the neuroscience community, see Editorial, *Got Data?,* 10 NATURE NEUROSCIENCE 931 (2007), *available at* http://www.nature.com/neuro/journal/v10/n8/pdf/nn0807-931.pdf (last visited Mar 13, 2009). For the Oxford University Press's reasons for not giving free access after a time delay, see Martin Richardson, *Impacts of Free Access*, NATURE WEB DEBATES, April 5, 2001, *available at* http://www.nature.com/nature/debates/e-access/Articles/richardson.html (last visited Mar 13, 2009).

[16] Google Scholar, http://scholar.google.com.

[17] 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."

[18] *See* Social Science Research Network, http://www.ssrn.com and Berkeley Electronic Press, http://www.bepress.com.

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.[19]

But the change is slow – academic researchers continue to be rewarded for publication in prestigious closed journals 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.[20] Some mechanisms for data sharing exist but usually through the efforts of a small team or dedicated researchers in specialized areas, such as the Stanford Microarray Database (SMD),[21] WaveLab and SparseLab,[22] Sweave,[23] or the open-source software package Madagascar designed to facilitate reproducible research.[24] Paul Caron has predicted a "long tail" effect for legal scholarship – he notes a broadening of citations generally a weakening of the concentration of download activity in the very top papers.[25] It seems reasonable to hypothesize that this pattern extends more widely than the legal research community Caron discusses, suggesting an increasing percolation of more new ideas into more readers' hands.

C. Reproducible Research Defined

The views expressed by Jon Claerbout,[26] a Stanford geophysics professor, have been taken to convey the message 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

---

[19] *Online Scientific Repository Hits Milestone,* CORNELL UNIVERSITY LIBRARY NEWS, October 3, 2008, *available at* http://communications.library.cornell.edu/com/news/PressReleases/arXiv-milestone.cfm (last visited Mar. 14, 2009).

[20] Google's program to house author-donated research datasets was cancelled Dec 18, 2008. *See* Alexis Madrigal, *Google Shutters Its Science Data Service*, WIRED SCIENCE, December 18, 2008, *available at* http://blog.wired.com/wiredscience/2008/12/googlescienceda.html (last visited Mar. 14, 2009) and Alexis Madrigal, *Google to Host Terabytes of Open-Source Science Data,* WIRED SCIENCE, January 18, 2008, *available at* http://blog.wired.com/wiredscience/2008/01/google-to-provi.html (last visited Mar. 14, 2009). This also illustrates one of the dangers inherent in the use of a private company to house research data associated with society's stock of knowledge.

[21] *See* Stanford Microarray Database, http://smd.stanford.edu/ and Janos Demeter et al., *The Stanford Microarray Database: Implementation of New Analysis Tools and Open Source Release of Software*, 35 NUCLEIC ACIDS RESEARCH D766 (2007), *available at* http://nar.oxfordjournals.org/cgi/reprint/35/suppl_1/D766 (last visited Mar. 14, 2009). For another example of reproducibility in bioinformatics, see MDACC: Bioinformatics: Supplementary Material, http://bioinformatics.mdanderson.org/supplements.html.

[22] Both are software packages reproducing published works in signal processing and spare representation respectively. *See* WaveLab802, http://www-stat.stanford.edu/~wavelab and SparseLab, http://sparselab.stanford.edu.

[23] Sweave allows a user to embed reproducible research in slide decks. *See* Sweave, http://www.statistik.lmu.de/~leisch/Sweave/.

[24] Madagascar, http://www.ahay.org/wiki/Main_Page (software released in 2006). *See also* The Reproducible Research Archive, http://www.reproducible.org/; LCAV – Audiovisual Communications Laboratory, http://lcav.epfl.ch/reproducible_research/; Reproducible Research at Vanderbilt, http://www.reproducibleresearch.org/Vanderbilt_Biostatistics.html as well as Trident: Scientific Workflow Workbench for Oceanography, http://www.microsoft.com/mscorp/tc/trident.mspx (presenting software designed to allow oceanographers to keep track of data collection and processing steps)..

[25] Paul Caron, *The Long Tail of Legal Scholarship*, 116 YALE L. J. POCKET PART 36 (2006), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=944233 (last visited Mar 14, 2009).

[26] *See* Jon F. Claerbout & Martin Karrenbach - Seismology on CD-ROM, http://sepwww.stanford.edu/sep/jon/blurb.html and Jon Claerbout, Stanford University, Seventeen Years of Super Computing and Other Problems in Seismology (Oct. 2, 1994), *available at* http://sepwww.stanford.edu/sep/jon/nrc.html (last visited Mar. 30, 2009).

instructions which generated the figures."[27] This encapsulates the idea of *reproducible research*: The notion that results should be independently replicable, given an appropriate computer platform.[28] Although the community of scientists who engage in reproducible research is small, a number of studies have shown that making papers and data available online leads to higher citation levels.[29]

D. Moving Towards Replication of Scientific Findings

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." The Open Archives Initiative and Science Commons are proposing universal standards for data repositories to facilitate reproducibility and novel scientific research.[30] Companies such as Metaweb and Google are creating new web structures to unify the housing of complex data.[31] Some research labs carry out reproducible research as a policy and this number is growing.[32] Similarly an increasing number of papers have been published recently calling for reproducible research.[33] In July of 2007, Microsoft held a Research Faculty

---

[27] JONATHAN B. BUCKHEIT & DAVID L. DONOHO, WAVELAB AND REPRODUCIBLE RESEARCH (1995), *available at* http://www-stat.stanford.edu/~donoho/Reports/1995/wavelab.pdf (last visited Mar. 15, 2009). *See also* Matt Schwab & Jon Claerbout – Reproducible Electronic Documents, http://sepwww.stanford.edu/doku.php?id=sep:research:reproducible (last visited Mar. 15, 2009) [hereinafter Schwab & Claerbout].

[28] *See* Jelena Kovacevic, *How to Encourage and Publish Reproducible Research*, 4 PROC. IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING 1273 (2007), *available at* http://lcav.epfl.ch/reproducible_research/ICASSP07/Kovacevic07.pdf (last visited Mar. 15, 2009).

[29] *See, e.g.*, Steve Lawrence, *Free Online Availability Substantially Increases a Paper's Impact*, NATURE WEB DEBATES, May 31, 2001, *available at* http://www.nature.com/nature/debates/e-access/Articles/lawrence.html (last visited Mar. 15, 2009); Stevan Harnad & Tim Brody, *Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals*, D-LIB MAGAZINE, July 2004, *available at* http://www.tubonotubo.jp/webproxy/index.cgi?do=proxy&wpburl=http://eprints.ecs.soton.ac.uk%2F10207%2F01%2F06harnad.html (last visited Mar 15, 2009); Heather A. Piwowar, Roger S. Day & Douglas B. Fridsma, *Sharing Detailed Research Data Is Associated with Increased Citation Rate*, 2 PLOS ONE e308 (2007), *available at* http://www.plosone.org/article/fetchObjectAttachment.action?uri=info%3Adoi%2F10.1371%2Fjournal.pone.0000308&representation=PDF (last visited Mar. 15, 2009). *See also* CHAWKI HAJJEM & STEVAN HARNAD, THE OPEN ACCESS CITATION ADVANTAGE: QUALITY ADVANTAGE OR QUALITY BIAS? – OPEN ACCESS ARCHIVANGELISM (2007), *available at* http://eprints.ecs.soton.ac.uk/13328/2/moed.pdf (last visited Mar. 15, 2009).

[30] *See* Open Archives Initiative, http://www.openarchives.org/ and Science Commons - Protocol for Implementing Open Access Data, http://sciencecommons.org/projects/publishing/open-access-data-protocol/ [hereinafter Science Commons Open Access Data Protocol].

[31] *See* Freebase, http://www.freebase.com/ and Google Base, http://www.google.com/base.

[32] Although it is still very small. *See, e.g.,* Stanford Exploration project, http://sepwww.stanford.edu/; Dave Donoho, http://www-stat.stanford.edu/~donoho and LCAV – Audiovisual Communications Laboratory, http://lcavwww.epfl.ch/ for a few examples.

[33] *See* Gentleman & Lang, *supra* note 8 and Giovanni Baiocchi, *Reproducible Research in Computational Economics: Guidelines, Integrated Approaches, and Open Source Software*, 30 COMPUTATIONAL ECONOMICS 19 (2007). *See also* Artemus Ward, *How One Mistake Leads To Another: On the Importance of Verification/Replication*, 12 POLITICAL ANALYSIS 199 (2004), *available at* http://polmeth.wustl.edu/polanalysis/vol/12/ward.doc (last visited Mar. 15, 2009); Edward J. Kane, *Why Journal Editors Should Encourage the Replication of Applied Econometric Research*, 23 QUARTERLY JOURNAL OF BUSINESS AND ECONOMICS 3 (1984), *available at* http://www.qjbe.unl.edu/pdfs/Kane.pdf (last visited Mar. 15, 2009); John O'Loughlin, *The War on Terrorism, Academic Publication Norms and Replication*, 57 THE PROFESSIONAL GEOGRAPHER 588, *available at* http://www.colorado.edu/IBS/PEC/johno/pub/PG_COMMENTARY.pdf (last visited Mar. 15, 2009).

Summit discussing reproducible research.[34] 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 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.[35]

The faculty members must provide their manuscript to the university for deposit into the open access repository within a year of publication.[36] Stanford's School of Education followed suit with a mandate for open access: all faculty members must make a copy of their published work available in an open access repository as of July 26, 2008.[37]

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."[38] The NIH envisions a searchable database of NIH funded publications. In 2004, the NIH published a notice called "Enhanced Public Access to NIH Research Information" in the National Institutes of Health (NIH) guide.[39] In this notice, the NIH recommends that all publications that arise from NIH-funded research be made available free to the public within six months of publication.[40] So far, the NIH has been silent on the issue of copyright.

Paul Huber has been advancing open access to research articles and their preprints, free of copyright and licensing restrictions.[41] He advocates the explicit use of Creative Commons licenses for the research papers or a similar licensing structure that allows the copyright holder to

---

[34] *See* Faculty Summit 2007, http://research.microsoft.com/en-us/um/redmond/events/fs2007/agenda_mon.aspx.

[35] *February 2008 Agenda* 3 (Faculty of Arts and Sciences, Harvard University 2008), *available at* http://www.fas.harvard.edu/~secfas/February_2008_Agenda.pdf (last visited Mar. 15, 2009).

[36] *See also* Stuart Shieber and the Future of Open Access Publishing, http://blog.stodden.net/2008/11/23/stuart-shieber-and-the-future-of-open-access-publishing/ (Nov. 23, 2008).

[37] *See* Robert Mitchell, *Harvard to Collect, Disseminate Scholarly Articles for Faculty*, HARVARD UNIVERSITY GAZETTE ONLINE, February 13, 2008, *available at* http://www.news.harvard.edu/gazette/2008/02.14/99-fasvote.html (last visited Marc. 15, 2009) and Peter Suber, *OA Mandate at the Stanford School of Ed*, OPEN ACCESS NEWS, June 26, 2008, *available at* http://www.earlham.edu/~peters/fos/2008/06/oa-mandate-at-stanford-school-of-ed.html (last visited Mar. 15, 2009).

[38] Notice: Enhanced Public Access to NIH Research Information, http://grants1.nih.gov/grants/guide/notice-files/NOT-OD-04-064.html.

[39] *See id.*

[40] Jeffrey Drazen & Gregory Curfman, *Public Access to Biomedical Research*, 135 NEW ENGLAND JOURNAL OF MEDICINE 2879 (2004), *available at* http://content.nejm.org/cgi/content/full/351/13/1343 (last visited Mar. 17, 2009).

[41] Peter Suber – Open Access Overview, http://www.earlham.edu/%7Epeters/fos/overview.htm.

"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...."[42] The Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities advocates the promotion of "the Internet as a functional instrument for a global scientific knowledge base and human reflection and to specify measures which research policy makers, research institutions, funding agencies, libraries, archives and museums need to consider" and has been signed by 242 organizations including universities and advocacy groups such as the Open Society Institute.[43]

E. Reasons to Perform Reproducible Research

Individual scientists have incentives to engage in reproducible research, and society as a whole stands to benefit in several ways. A scientist's gains from working reproducibly derive from the follows areas.

Reputational gains through exposure and citation: Open research is built upon and cited more frequently than work published in closed journals.[44] Difficulty in reproducing computational results creates a barrier to other scientists' abilities to build on previously published work.[45] Verification and regeneration of results often requires a detailed knowledge of parameters and software invocation sequences and without a clear description it can be next to impossible, even for the original scientist, to try their methodology in a new setting or on a new dataset. Reproducibility makes a scientist's work fully accessible to potential co-authors, future collaborators, possible employers, students, post-docs and other researchers in the area.[46]

Preservation of valuable work: Very often there are many small details involved in the production of research results and without accurate documentation researchers can forget how a particular outcome was reached. Working reproducibly acts to 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 compelling results.[47]

Community membership: Access to complete information may satisfy a basic need, or even a "spiritual necessity," among independent scientists to understand scientific regions "as a whole, and to lend one another strength of that understanding."[48] Polanyi characterized scientists as community members bound by a commitment to truth and openness, thus creating what he

---

[42] The Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities, http://oa.mpg.de/openaccess-berlin/berlindeclaration.html. For background on Creative Commons, see Niva Elkin-Koren, *What Contracts Can't Do: The Limits of Private Ordering in Facilitating a Creative Commons*, 74 FORDHAM L. REV. 375 (2005), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=760906 (last visited Mar. 17, 2009).

[43] Open Access Conference – Signatories, http://oa.mpg.de/openaccess-berlin/signatories.html.

[44] *See, e.g.,* S. Lawrence, *supra* note 29 and Piwowar, Day & Fridsma, *supra* note 29. *See also* HAJJEM & HARNAD, *supra* note 29.

[45] For an example of the difficulty of replication in economics, see B.D. McCullogh, *Got Replicability? The Journal of Money, Credit, and Banking Archive*, 4 ECON JOURNAL WATCH 326 (2007), *available at* http://www.econjournalwatch.org/pdf/McCulloughEconomicsInPracticeSeptember2007.pdf (last visited Mar. 17, 2009). *See also* Gary King – Data Sharing and Replication Information, http://gking.harvard.edu/replication.shtml [hereinafter Gary King].

[46] Donoho et al, *supra* note 1.

[47] BUCKHEIT & DONOHO, *supra* note 27.

[48] NORBERT WEINER, CYBERNETICS 3 (2nd ed. MIT Press 1965).

termed "The Republic of Science."[49] Releasing the full research product allows for the reporting and attribution of more results and experimental configurations than would ordinarily be publishable, and can enhance a sense of full communication in a shared community.[50]

Good citizenship: Reproducibility is required: In 2004 National Science Foundation (NSF) grants comprised 64% of total academic research and development support, and that proportion is increasing.[51] 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.[52] An increasing number of journals are requiring the submission of the code and data supporting the results in the published work.[53]

Reputational gains through discipline: Working with the knowledge that the code you write, and that any modifications to the data you make will be public creates an incentive for a researcher to ready his or her work for wide scrutiny. A scientist will be less willing to work using his or her short-term memory of procedures or parameters used and more likely to record to steps carefully and so strive to produce better work.

Reproducibility also acts to increase social welfare through the following mechanisms.

Diminishing the credibility gap: Reproducibility permits the verification of results. Mistakes and self-delusion can creep into work anywhere and a scientist's effort is primarily expended in finding and controlling error. Before scientific computation can be recognized as a mature scientific endeavor, it must be practiced in a way that accepts the ubiquity of error, and works to identify and root out error.[54]

Fraud prevention: Knowing work will be fully open to inspection in the future creates an incentive for researchers to do better, more careful, science now. Openness prevents any desire, even unconscious, to modify results in such a way that departs from the paper's underlying methodology. For example, a researcher might be tempted to illuminate findings or otherwise clarify the exposition of results after they have been produced the way the paper describes.[55] Without full publication of "a careful description of the methods used, in sufficient detail that

---

[49] MICHAEL POLANYI, THE LOGIC OF LIBERTY 69 (The University of Chicago Press, 1951).

[50] *See* John Ioannidis, *Why Most Published Research Findings are False*, 2 PLOS MED e124 (2005), *available at* http://medicine.plosjournals.org/perlserv/?request=get-document&doi=10.1371/journal.pmed.0020124 (last visited Mar. 17, 2009).

[51] Rhonda Britt, *Industrial Funding of Academic R&D Continues to Decline in FY 2004*, NATIONAL SCIENCE FOUNDATION INFOBRIEF, April 2006, *available at* http://www.nsf.gov/statistics/infbrief/nsf06315/nsf06315.pdf (last visited Mar. 17, 2009).

[52]

> 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) 27 (National Science Foundation, June 1, 2007), *available at* http://www.nsf.gov/pubs/policydocs/gc1_607.pdf (last visited Mar 17, 2009).

[53] For further details of funding agency guidelines and journal policies, see Gary King, *supra* note 45.

[54] *See* Donoho et al, *supra* note 1.

[55] *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* http://chronicle.com/free/2008/05/3028n.htm (last visited Mar. 17, 2009).

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."[56]

Knowledge diffusion: Reproducibility implies a broadening of access to scientific know-how to anyone anywhere who has an Internet connection and the ability to run the routines or understand the scripts.[57] Researchers not in the immediate field of research can download, modify, and apply the work, thereby facilitating interdisciplinary research and collaboration.

Better quality science: Knowing their work will be open to full scrutiny, each scientist will strive for discipline that produces better science.[58] A researching scientist may have done more experimentation than is practical to report in a traditional research paper. Reproducible science also provides an opportunity for scientists to report the negative results of their research, furnishing a more complete picture of the state of knowledge with regard to a particular research problem.

F. Legal Impediments to Reproducibility

This paper focuses on impediments to scientific reproducibility found in our current intellectual property framework.[59] Copyright law acts against foundational scientific norms in two key ways. First, by preventing copying of the research work, it creates a barrier to the possibility of legally reproducing and verifying another scientist's results without the need to obtain prior permission from the authoring scientist.[60] Second, copyright also establishes rights for the owner over the creation of derivative works. A second scientific norm guides scientists to build on previous discoveries – using copyrighted work in derivative research typically requires obtaining the permission of the copyright holder, thus creating a block to the generation of new scientific discoveries. Although the alternative to copyright, secrecy, is of course a greater block to scientific progress.

---

[56] R. J. LeVeque, *Wave Propagation Software, Computational Science, and Reproducible Research*, 2006 PROC. INTERNATIONAL CONGRESS OF MATHEMATICIANS, *available at* http://www.amath.washington.edu/~rjl/pubs/icm06/icm06leveque.pdf (last visited Mar. 17, 2009). *See also* P. Vandewalle, G. Barrenetxea, I. Jovanovic, A. Ridol & M. Vetterli, *Experiences With Reproducible Research in Various Facets of Signal Processing Research*, 4 PROC. IEEE CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING 1253 (2007), *available at* http://infoscience.epfl.ch/record/97195/files/ (last visited Mar. 17, 2009) [hereinafter Vandewalle et al.].

[57] This includes researchers who may not have access to publications that require subscription access. For a discussion of copyright and idea flows across the digital divide, see Haochen Sun, *Copyright Under Siege: An Inquiry into the Legitimacy of Copyright Protection in the Context of Global Digital Divide*, 35 INT'L REV. INDUS. PROP. & COPYRIGHT L. 192 (2005), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=734623 (last visited Mar. 17, 2009).

[58] BUCKHEIT & DONOHO, *supra* note 27. *See also* Schwab & Claerbout, *supra* note 27.

[59] For more complete discussion of the disincentives facing scientists with regard to reproducible research, see Donoho et al, *supra* note 1, at 16. *See also* Partha Dasgupta & Paul A. David, *Toward a New Economics of Science, in* SCIENCE BOUGHT AND SOLD: ESSAYS IN THE ECONOMICS OF SCIENCE 219, 233 (P. Mirowski & E.-M. Sent eds., University of Chicago Press, 2002) ("… the reward system [in the sciences] sets up an immediate tension between cooperative compliance with the norm of full disclosure (to assist oneself and colleagues in the communal search for knowledge), and the individualistic competitive urge to win priority races.").

[60] *See* Victoria Stodden, *The Legal Framework for Reproducible Scientific Research: Licensing and Copyright*, 11 COMPUTING IN SCIENCE AND ENGINEERING 35 (2009).

III. COPYRIGHT AND THE SHARING OF SCIENTIFIC WORK

The Intellectual Property Clause of the Constitution has been interpreted to confer two distinct powers, the first providing the basis for copyright law: Securing for a limited time a creator's exclusive right to their original work;[61] and the second the basis for patent law: Giving an inventor a limited term exclusive right to their discoveries in exchange for disclosure of the invention.[62] With the advent of the Internet and the increase in computation in science, these two derived powers have come to create a false dichotomy with respect to the type of protection necessary for the "Progress of Science and Useful Arts."[63] Since key norms that promote the progress of science are the relinquishment of property rights and the ability to freely build upon others' work in exchange for attribution, Intellectual Property law that frames work as falling within one of the spheres of copyright or patent law, leaves scientific research without a natural legal home.[64]

Creative Commons was founded in 2001 by Larry Lessig to give creators of artistic works the ability to allow others to freely use and reuse their creation under terms they set.[65] Creative Commons provides a suite of licenses that give terms of use for work that differ from, and are usually more permissive than, the default copyright. This paper extends this approach to the scientific context to realign copyright with scientific norms, and to encourage really reproducible scientific research.

The Reproducible Research Standard (RRS) suggests an alignment of the Intellectual Property framework for scientific research with scientific norms. The first component of the standard is applying an appropriate license to remove restrictions on copying and reusing the scientific work, as well as adding an attribution requirement to elements of the research compendium.

A. Choosing to Free Research Work: Licenses

Components of the research compendium have different features that necessitate different licensing approaches. The effectiveness of a license, such as one of the Creative Commons licenses, is undergirded by copyright. Licenses do not remove or rescind copyright protection but

---

[61] For a discussion of the Copyright Act of 1976, see Pam Samuelson, *Preliminary Thoughts on Copyright Reform Project*, 3 UTAH L. REV. 551 (2007), *available at* http://people.ischool.berkeley.edu/~pam/papers.html (last visited Mar. 18, 2009). For a discussion of the Intellectual Property clause including relevant case law, see Yochai Benkler, *Constitutional Bounds of Database Protection: The Role of Judicial Review in the Creation and Definition of Private Rights in Information*, 15 BERKELEY TECH. L.J. 535 (2000), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=214973 (last visited Mar. 18, 2009).

[62] For further discussion, see Anselm Kamperman Sanders, *Limits to Database Protection: Fair Use and Scientific Research Exemptions*, 35 RESEARCH POLICY 854, 856 (2006).

[63] For further discussion of congressional power with respect to database protectability, see Benkler, *supra* note 61, at 2.

[64] For more on the incentives deriving from copyright law, such as the prevention of unfair competition, see Paul David, *The Economic Logic of Open Science and the Balance between Private Property Rights and the Public Domain in Scientific Data and Information: A Primer*, *in* THE ROLE OF SCIENTIFIC AND TECHNICAL DATA AND INFORMATION IN THE PUBLIC DOMAIN 19, 27-28 (J.M. Esanu & P.F. Uhlir eds. 2003), *available at* http://www.nap.edu/catalog.php?record_id=10785#toc (last visited Mar. 18, 2009). On the communitarian nature of scientific norms, see MERTON, *supra* note 4.

[65] For background on Creative Commons, see LAWRENCE LESSIG, CODE AND OTHER LAWS OF CYBERSPACE (Basic Books, 2000); LAWRENCE LESSIG, THE FUTURE OF IDEAS: THE FATE OF THE COMMONS IN A CONNECTED WORLD (Vintage, 2002) and LAWRENCE LESSIG, FREE CULTURE: THE NATURE AND FUTURE OF CREATIVITY (Penguin, 2005). *See also* Niva Elkin-Koren, *supra* note 42.

allow the creator to specify the conditions under which use of the work takes place. Licensing is given strength through rights created by underlying copyright law. Effectively, this means that even if a license fails, use of the work will remain subject to injunction and other remedies associated with copyright violation.[66]

With myriad options for licensing copyright-protected work, a principle for scientific licensing can serve to guide choices:

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

The goal of an Intellectual Property legal framework for scientific research must be to increase what Benkler terms "that most precious of all public domains -- our knowledge of the world that surrounds us."[67] This effort involves an alignment of the private incentives faced by a scientific researcher and the societal benefit of increasing our stock of public knowledge. Scientific norms have arisen to align these interests in practice, and an associated Intellectual Property structure should reflect these norms to allow scientific research to flourish.[68]

1. The Paper, Figures, and Other Media Files

Creative Commons licenses permit specific royalty-free uses of the licensed work.[69] CC BY is the most permissive license – requiring only that the copyright owner's attribution specifications are followed. Other optional conditions include a requirement that derivative works be licensed under the same terms (Share Alike), a restriction to non-commercial uses, and a prohibition on the creation of derivative works. In deciding which of these options he or she prefers, the copyright holder chooses from six licenses: "Attribution," (the CC BY license) "Attribution - No Derivative Works," "Attribution - Non-Commercial - No Derivative Works," "Attribution-Non-Commercial," "Attribution - Non-Commercial - Share Alike," and "Attribution - Share Alike."[70] Each CC license lasts the duration of the copyright protection associated with the work.

For media components of scientific work, alignment with scientific norms is most readily and simply achievable through use of the CC BY license which frees the work for reuse, with the condition that attribution must accompany any downstream use of the work.

---

[66] This recourse to copyright for enforcement may not be necessary: a recent case (Jacobsen v. Katzer, Fed. Cir., August 13, 2008, No. 2008-1001) found a software license to be enforceable like a copyright condition for which courts can apply the remedy of injunction.

[67] Benkler, *supra* note 61, at 3.

[68] For a description of the four scientific norms, see MERTON, *supra* note 4. Of particular interest to us is the *Communitarian* norm: that scientists relinquish ownership rights over their work in exchange for acknowledgement through citation or perhaps the naming of discoveries. This, in conjunction with the norm of *Skepticism* that establishes the close inspection and review of research work by the community, imply open access to scientific research, satisfying the interests of the larger community in the openness and availability of scientific research work. *See* David, *supra* note 64, at 21-22.

[69] For background on Creative Commons, see Niva Elkin-Koren, *supra* note 42.

[70] *See* Creative Commons - Choose a License, http://creativecommons.org/license. *See also* Michael Carroll, *Creative Commons and the New Intermediaries*, 2006 MICH. ST. L. REV. 45, 47.

2. The Code

A plethora of licenses exist that allow authors to set conditions of use for their code. In scientific research code can consist of scripts that are essentially stylized text files (such as Matlab or R scripts) or the code can have both a compiled binary form and a source representation (such as code written in C). Use of the CC BY license for code is actively discouraged by Creative Commons.[71] The license does not make a distinction between source or compiled forms of the work whereas licenses intended for use on software generally refer to source code as "the preferred form for making modifications."[72] This explicit reference to source code recognizes the fact that software code can exist in two forms, source and compiled, and for modification transmission of the binary form alone is not sufficient.[73]

Since default copyright extends to code, Richard Stallman began the Free Software movement in the early 1980's to encourage programmers to release their source code along with the software compiled for end users.[74] Stallman developed the GNU General Public License (GPL),[75] which 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.

This license contains the Share Alike provision – work that uses code under this license must carry the GPL. Less frequently used is the GNU Lesser General Public License (LGPL).[76] It was developed for code libraries and permits their use in proprietary packages, which the GPL does not due to the Share Alike provision.

The LGPL was designed for a code library that other modules link to. This type of code library development is not typical in computational scientific research. What is more typical is a group of scripts or code files that implement an algorithm or idea. The LGPL requires that for a work that uses an LGPL licensed library, the LGPL and the GPL licenses must be attached, reverse engineering of the library in not impeded (say, through compiling into binary form), and notice must be given that the library is a component of your work and well as installation information.[77]

---

[71] "[W]e do not recommend that you apply a Creative Commons license to software code." Frequently Asked Questions – CC Wiki, http://wiki.creativecommons.org/FAQ.

[72] *See, e.g.,* Apache License, Version 2.0, http://www.apache.org/licenses/LICENSE-2.0.html [hereinafter Apache License 2.0].

[73] It is conceivable that the CC BY license could apply to the script forms of code that have no binary compiled counterpart. There doesn't seem to be a benefit to doing this since these scripts can be licensed with attribution under a software license, thus keeping the intellectual separation between Creative Commons licenses for media and code-specific licenses for code and respecting Creative Commons desire that their licenses not be used for code.

[74] For more on the history of the GP, see The History of the GNU General Public License, http://www.free-soft.org/gpl_history and, for the history of the Free Software movement, see Richard Stallman – The GNU Project, http://www.gnu.org/gnu/thegnuproject.html.

[75] GNU General Public License, http://www.gnu.org/licenses/gpl.htm [hereinafter GPL].

[76] Use of this license is discouraged by the Free Software Foundation who developed it. *See* GNU Lesser General Public License, http://www.gnu.org/licenses/lgpl.html [hereinafter LGPL].

[77] *See* GPL, *supra* note 75. The LGPL also states that if a downstream author chooses to show copyright notices, the copyright notice for the library must also be included.

The (Modified) Berkeley Software Distribution (BSD) license permits the downstream use, copying, and distribution of either unmodified or modified source code, as long as the license accompanies any distributed code and the previous authors' names are not used to promote modified downstream code.[78] The license template is brief enough it can be included here:

> Copyright (c) <YEAR>, <OWNER>
>
> All rights reserved.
>
> Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:
>
> - Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
> - Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
> - Neither the name of the <ORGANIZATION> nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

This template is then followed by a disclaimer releasing the author from liability for use of the code.[79] There is no Share Alike provision, meaning that code licensed under the BSD can be incorporated into proprietary work.[80] The above copyright notice and list of conditions, as well as the disclaimer, must accompany derivative works. The Modified BSD license is very similar to the MIT license, with the exception that the MIT license does not include a clause forbidding endorsement.[81]

The Apache 2.0 license is another common method for developers to specify terms of use of their work.[82] Like the Modified BSD and MIT licenses, the Apache license does not contain the Share Alike provision and requires attribution. It differs from the previously discussed licenses in that it permits the exercise of patent rights that would otherwise only extend to the

---

[78] *See* Open Source Initiative OSI – The BSD License: Licensing, http://www.opensource.org/licenses/bsd-license.php.

[79] *See id.*

[80] 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 id.*.

[81] *See* Open Source Initiative OSI – The MIT License: Licensing, http://www.opensource.org/licenses/mit-license.php.

[82] *See* Apache License 2.0, *supra* note 72.

original licensor, meaning that a patent license is granted for those patents needed for use of the code. The license further stipulates that the right to use the work without patent infringement will be lost if the downstream user of the code sues the licensor for patent infringement. Attribution under Apache 2.0 requires that derivative works carry a copy of the license, with notice of any files modified. All copyright, trademark, and patent notices that pertain to the work must be included. Attribution can also be done in such a notice file.

3. The 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 are copyrightable. In *Feist Publications, Inc. v. Rural Telephone Service*,[83] the Court found that the white pages from telephone directories are not themselves directly copyrightable, since copyrightable works must have creative originality:[84]

> [T]he 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.[85]

Currently the Court holds *original* "selection and arrangement" of databases protectable:[86] The component falling under copyright must be original in that "copyright protection extends only to those components of the work that are original to the author, not to the facts themselves. . . ."[87] The extraction of facts from a database is permitted without violation of copyright. Attaching an attribution license to the original "selection and arrangement" of a database can encourage scientists to release the datasets they have created by providing a legal framework for attribution

---

[83] Feist Publications Inc. v. Rural Tel. Serv. Co., 499 U.S. 340 (1991) [hereinafter *Feist*].

[84] *Id.*, at at 363-364.

[85] *Id.*, at 349. *See also* Miriam Bitton, *A New Outlook on the Economic Dimension of the Database Protection Debate*, 47 IDEA: THE INTELLECTUAL PROPERTY LAW REVIEW 93 (2006) and H. Zhu & S. Madnick, *One Size does not Fit All: Legal Protection for Non-Copyrightable Data* (Sloan School of Management, Composite Information Systems Laboratory, Working Paper No. 2007-04), *available at* http://web.mit.edu/smadnick/www/wp/2007-04.pdf (last visited Mar. 30, 2009).

[86] Bitton, *supra* note 85, at 4.

[87] *Feist*, 499 U.S., at 340. The full quote reads:

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

Id. For a discussion of the Constitutional limits on Congress's ability to create property rights in facts, see Benkler, *supra* note 61.

and reuse of the original selection and arrangement aspect of their work.[88] Since the raw facts themselves are not copyrightable, it does not make sense to apply such a license to the data themselves. The selection and arrangement may be implemented in code or described in a text file accompanying the dataset, either of which can be appropriately licensed. Depending on the nature of the original "selection and arrangement" copyright protection may be dubious or nonexistent. Attribution requirements will only attach, through the license, when underlying copyright protection exists.

B. Revealing Research Compendia: The Reproducible Research Standard

Since the components of research compendia are varied, licenses should be applied as appropriate to each component in accordance with the Principle of Scientific Licensing. Providing attribution for use of research work is a cornerstone of scientific enterprise and this can be reflected in license choice. Using CC BY on the media components of the research, such as text and figures, permits other scientists to freely use and reuse this work provided the original author is attributed. The same result is obtained by using a software license that provides an attribution component for the code components, such as the Apache License 2.0, the Modified BSD License,[89] or the MIT License. The original selection and arrangement of the data can be similarly licensed depending on whether it takes a code or text format. Since an attribution license cannot be attached to raw facts, data can be released to the public domain by marking with the Science Commons Open Data Protocol[90] or CC0 standard.[91] A licensing structure that makes media, code, data, and data arrangements – the research compendium – available for reuse, possibly with attribution, is termed the *Reproducible Research Standard*.

If a researcher wishes to release his or her research compendium entirely to the public domain, this also complies with the Reproducible Research Standard. This goal is to make research compendia available for the verification of research results and to facilitate the building of further research upon these works. A licensing structure that adds an attribution component is designed to encourage scientists to release their work while reducing concern about recognition of their efforts.

The LGPL could be used for licensing of the code components of the RRS, although it's requirements for notice and attribution are much higher than the other three licenses (for example, including copies of both the LGPL and GPL licenses). In accordance with the Principle of Scientific Licensing, minimizing encumbrance to original and downstream researchers would suggest using one of the Apache 2.0 license, the MIT license, or the Modified BSD license.

C. Attribution in Scientific Licensing

Attribution is a core mechanism by which scientific research progresses and it underlies the traditional system under which ideas and research output are shared. For an individual scientist success is most often measured by citations, i.e. the amount of subsequent work he or

[88] For a discussion of the international and WIPO statements of the legal status of databases, see Kamperman Sanders, *supra* note 62, at 859.

[89] Creative Commons provides the BSD as a CC license. *See* Creative Commons BSD License, http://creativecommons.org/licenses/BSD/.

[90] *See* Science Commons – Database Protocol, http://sciencecommons.org/resources/faq/database-protocol/ and CC0 Beta/Discussion Draft 3, http://creativecommons.org/weblog/entry/9071 [hereinafter CC0 Standard].

[91] For details on the CC0 protocol, see CC0 Standard, *supra* note 90.

she engenders.[92] Including an attribution component in the licensing structure of research compendia aligns the RRS with these longstanding scientific values. Concern over loss of attribution is a reason scientists hesitate to release their full compendia publicly.

The particular selection of licenses under the RRS allows for "viral" attribution meaning that any element of such a compendium that is reused in others' work, such as new image compression code or a particular arrangement of microarray data, retains the original attribution. For example, software under the Modified BSD license is attributed to the original author, and research that builds upon this work must also retain the Modified BSD license on that particular piece of code written by the original author, thus maintaining attribution to the original creator of the code. Similarly, the CC BY license attaches notification of authorship to text and other media. The licensing structure of the RRS is therefore "viral" in that attribution propagates through downstream scientific research. This mechanism largely mirrors how scientific work is typically cited and built upon, with the difference that the attribution process is formalized in a legal license, as opposed to academic citation.[93]

The attribution aspect of licensing is so fundamental to scientific research I argue the benefit of providing attribution outweighs the encumbrance and satisfies the Principle of Scientific Licensing. I further argue that the Share Alike aspect common to many licenses does not.

D. Share Alike in the Scientific Context

The licensing structure under the RRS aims to ensure each scientist is attributed only for the work he or she has authored. The Share Alike component is found in some popular licenses (for example the GNU GPL license and a number of Creative Commons licenses) and specifies that the use of Share Alike licensed work in the development of another body of work, will bring the entire derivative work under the original license unless an alternative is negotiated with the original's copyright holders. Specifically, the Share Alike provision states that: "If you alter, transform, or build upon this work, you may distribute the resulting work only under a license identical to this one."[94] This creates the restriction that only Public Domain content or work capable of being licensed as Share Alike can now be incorporated into the research, since it must be distributed as Share Alike.

To take a specific example, suppose a body of code is licensed under the GNU Public License. Because of the Share Alike provision in the GPL license, using GPL licensed code in a new body of code requires the entire derivative code to come under the GPL. If the downstream author also wishes to incorporate code that does not have a Share Alike provision or a different license with a Share Alike provision, the downstream work cannot incorporate both pieces of code because of their incompatible, and equally rigid, licensing structures. Before scientific research becomes highly licensed, there is an opportunity to learn from the analogous experience in patent law and avoid the creation of "copyright thickets" preventing the reuse of scientific

[92] For a discussion of priority and the reward mechanism in scientific research, see Dasgupta & David, *supra* note 59, at 229.

[93] For a discussion of the encoding of attribution for research compendia under the RRS, see *supra* Part III.A..

[94] Mikael Pawlo, *What is the Meaning of Non-Commercial?, in* INTERNATIONAL COMMONS AT THE DIGITAL AGE 69, 76 (Danièle Bourcier & Mélanie Dulong de Rosnay eds., Romillat, Paris 2004), *available at* http://fr.creativecommons.org/iCommonsAtTheDigitalAge.pdf (last visited Mar. 30, 2009).

research.[95] 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[96], but scientific research should be encouraged over particular license use.

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. Restricting the licensing options of a scientist's figures, say, because they used or modified another researcher's code to build them creates an unnecessary bar to research. The Science Commons Open Access Data Protocol embodies many of these arguments.[97]

IV. THE REPRODUCIBLE RESEARCH STANDARD IN PRACTICE

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 or released to the public domain under CC0,
3. The code components are licensed under one of Apache 2.0, the MIT License, or the Modified BSD license, or released to the public domain under CC0,
4. The data have been released into the public domain according to the Science Commons Open Data Protocol.

In joint work with Science Commons we are developing a mechanism that would allow scientists to assert their compliance with these conditions, and publicly certify their work as reproducible. Since it may not be feasible for a scientist to satisfy all these conditions because of reasons beyond his or her control (such as privacy considerations for data)[98] we propose different levels of compliance. If the work is publicly released, for example on the Internet, and capable of being reproduced it is marked as "Verifiable." If the work has been verified by someone working independently it can be marked as "Verified." When the full compendium is not released, the work can be marked as "Semi-Verifiable" and if the code has been verified or the results are

[95] For a discussion of how patent thickets can operate to prevent use of patented work in innovation, see James Besson, *Patent Thickets: Strategic Patenting of Complex Technologies* (Research on Innovation Working Papers, March 2003), *available at* http://www.researchoninnovation.org/thicket.pdf (last visited Mar. 30, 2009) and MICHEAL HELLER, THE GRIDLOCK ECONOMY: HOW TOO MUCH OWNERSHIP WRECKS MARKETS, STOPS INNOVATION, AND COSTS LIVES (Basic Books, New York 2008).

[96] For a discussion of the sharing public results and "complementary externalities" in proprietary research, see David, *supra* note 64, at 23.

[97] Science Commons Open Access Data Protocol, *supra* note 30. For a discussion of the complementary nature of proprietary and academic research communities, see Dasgupta & David, *supra* note 59, at 227.

[98] There are many reasons a dataset may not be able to be made publicly available, such as confidentiality of the records, an author's lack of ownership of the data, security risks, for example. Recently Harvard's Berkman Center attempted the release of Facebook profile data scraped by a user within a particular network. Anonymizing these data has proven to be a challenge. *See* Berkman Center researcher publishes 1700 students' Facebook data: "We did not consult w/ privacy experts on how to do this, but we did think long and hard …" (DRAFT), http://www.talesfromthe.net/jon/?p=234 (Oct. 8, 2008, 23:32 EST).

achieved in a different dataset (depending on which components were not released), the work is marked as "Semi-Verified." Work that depends on streaming data, such as that provided by near-continuous sensor readings for example, could be marked as "Perpetually Verifiable."[99] [100] This schema of RRS levels does not disqualify a scientist who releases his or her compendia upon request and has another researcher independently verify his or her results from being labeled "verified."[101]

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. An identifying logo would foster a sense of community among scientists who work reproducibly. 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. The RRS delineates an easily communicable way for funding institutions, publishers, or collaborators to require the public availability of the entire research product. The RRS could provide cultural impetus to encourage reproducible research, and perhaps provide a mechanism enabling journals to publish RRS compliant papers and grant giving organizations to fund work in accordance with the RRS.[102] Compliance with the RRS is not costless. At minimum effort must be taken to post the research on the web, but there are probably larger costs in readying research for public inspection. The RRS can be a tool for university administrators to communicate a change in expectations for researchers. It can also act in the reverse: As a way for researchers to explain reproducibility to tenure committees or university administrators. Individual scientists also have reasons to mark their work with the RRS. As one researcher has pointed out, an advantage to open code and clarity of experimental method is publicity of the new work.[103]

A. Legal Attribution and Scientific Citation

The notion of attribution provided by many licenses, such as CC BY or the Modified BSD license, is not the same as scientific citation as routinely practiced by scientists. Each license creates an additional burden for scientists who then must attach the appropriate attribution information to their work. A scientist's incentives are currently structured so that credit for research is highly desirable,[104] but the process of legal attribution must be as natural as possible to the individual scientist according to the Principle of Scientific Licensing, while still achieving the welfare enhancing goal of the widest possible distribution of open scientific research.

---

[99] I suppose computational work without either code or data components could be marked as "Unverified."

[100] This marking schema is based on a suggestion from David Purdy at the Neyman Seminar given by the author at the UC Berkeley Statistics department February 18, 2009. It improves the original idea of Gold, Silver, and Bronze flavors of Reproducible Research I have previously suggested.

[101] This does not preclude data that has privacy restrictions or limited use through the legal requirement of a data use agreement. The RRS does not conflict with these requirements, although the work cannot be classified ad "verifiable." Gary King has developed the Universal Numerical Fingerprint (UNF) to identify data uniquely and associate it with certain results, even though the contents of the dataset are hidden. Attaching a UNF, while not permitting reproducibility, does provide a measure of confidence in the version of the dataset used. *See* The Standard – The Dataverse Network Project, http://thedata.org/citation/standard.

[102] *See* Kovacevic, *supra* note 28.

[103] *See* Vandewalle et al., *supra* note 56.

[104] Although more so for ideas than for data.

The terms of the CC BY license provide a hook for its adaptation to the scientific context. Section 4(b) refers to the obligation of a downstream user of a creative work to "keep intact all copyright notices for the Work and provide, reasonable to the medium or means You are utilizing… ."[105] The inclusion of such a clause appears to reflect the intention of the writers of CC BY license that it be applicable to forms of creative work that might appear on paper, clay, film, hard disk, or any other artistic media in order to allow the user to provide attribution in the form most natural for the work in question. "[R]easonable to the medium" can provide a hook in the reproducible research context by allowing legal attribution to occur in a form more recognizable as scientific citation. Here, "the medium," might be scientific research for which there are established standards of citation by field. Aside from minimizing the administrative burden on the scientist, the fact that citation norms vary by research area is another reason to consider a more norms-based and less strictly legal-based approach to attribution in the sciences.

Scientists may feel more comfortable turning over their attribution rights to a centralized body, such as the National Science Foundation.[106] This suggests a second possible solution to the burden of legal compliance. This central body could establish a list of best practices in citation and attribution by field or subfield, and scientists could then follow these standards to the best extent that they can. As with the first solution, this mechanism would allow citation standards in a field to evolve, especially with the introduction of new technologies for attribution.[107] This suggests establishing a copyright clearing house for scientific work similar to the Copyright Clearance Center to function as a central mechanism to handle rights associated with compendia.[108]

The benefit to allowing flexible and evolving standards of attribution goes beyond adaptation to new technologies. Without such a standard as the RRS, conflicts between the various copyright and licensing options could become a serious problem for reuse of scientific work, increasingly so as scientists begin to manage their own copyright rights. This could impose a cost on future scientists seeking to build upon these works, if rights are fragmented or in conflict.[109] 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. Since the RRS suggests using well-known licenses, there are no additional compatibility or interoperability issues with existing licenses.

The Creative Commons licenses use the Resource Description Framework (RDF) to encode meta-data concerning attribution and other specifications of the licenses. Similarly, the attribution parameter of the Modified BSD can be encoded as an html tag associated with the web page housing the released work. The RRS therefore enables a mechanism through which authorship data and other meta-data can be encoded and associated with the research in a

---

[105] For the complete terms, see Creative Commons Legal Code, http://creativecommons.org/licenses/by/3.0/legalcode.

[106] PNAS, the Proceedings of the National Academy of Science, has recently decided to return all rights associated with copyright to the original authors. *See* Randy Schekman, *PNAS Allows Authors to Keep Copyright*, 106 PNAS 3 (2009), *available at* http://www.pnas.org/content/106/1/3.full.pdf+html (last visited May 15, 2009).

[107] Both the ideas in the previous two paragraphs were developed in conversation with Thinh Nguyen, counsel for Science Commons.

[108] Copyright Clearance Center, http://www.copyright.com.

[109] For a discussion of impediments due to lack of standardization in the Creative Commons movement, see Niva Elkin-Koren, *supra* note 42.

machine-readable way.[110] Since attribution is a field in a tag on the elements of the research compendia, adding an arbitrary number of authors in tags becomes meaningful to computerized search and machine-readability.

B. The Role of Third Parties

The RRS's licensing structure clarifies the role of third parties. This is important as the university is a common setting for computational research, and universities nearly always claim rights to work developed using university facilities, although they are often amenable to open release of software.[111] 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 RRS and their interests as a university.

C. Why Not the Public Domain? Or Fair Use?

The scientific ethos generally disclaims ownership over scientific research by individual scientists. With copyright law assigning rights by default to the creators of original work, some action is needed on the part of the authoring scientist in order to comply with this scientific norm. If a scientist chooses to release all of his or her work to the public domain, foregoing the attribution element of the RRS, the scientist must still take steps to certify the work as part of the public domain. This can be done through the Creative Commons CC0 protocol, which provides a way for the scientist to assert that there are no legal restrictions attached to this compendium.[112] If a scientist does enter their work into the public domain, this work can still be branded as Reproducible under the RRS and may still carry the Reproducible Research logo.

Paul David has proposed the removal of legal barriers to sharing of scientific compendia by providing "those engaged in non-commercial scientific research and teaching with automatic "fair use" exemptions from the force of intellectual [property] law."[113] If such an exemption was created, the RRS would work in concert with this legal change.

---

[110] *See* ccRel – CC Wiki, http://wiki.creativecommons.org/CcREL. For a discussion of machine readability of Creative Commons licenses including search tools, see Carroll, *supra* note 70.

[111]

> [I]f 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.

Katharine Ku, *Software Licensing in the University Environment*, COMPUTING RESEARCH NEWS, Jan 2002, at 3, 8, *available at* http://www.cra.org/CRN/issues/0201.pdf (last visited Mar. 30, 2009).

[112] For details on the CC0 protocol, see Eric Steuer, *Creative Commons Launches CC0 and CC+ Programs* (December 17, 2007), *available at* http://creativecommons.org/press-releases/entry/7919 (last visited Mar. 30, 2009).

[113] David, *supra* note 64, at 29. For a discussion of branding and fair use to protect and encourage industrial research, see Kamperman Sanders, *supra* note 62, at 857. For a discussion of the use of the fair use exception in academic research, see posting of John Willinsky to *S*law, http://www.slaw.ca/2008/05/01/harry-potter-and-the-scholar%E2%80%99s-fair-use/ (May 1, 2008).

In the current legal framework, there are reasons why fair use alone may not be a sufficient palliative for scientists wishing to use copyrighted works in their research. Fair use in U.S. copyright law provides for the use of copyrighted works without the need to obtain the copyright holder's permission, in order to provide flexibility in balancing the interests of copyright holders and the public's desire to make use of copyrighted works. The copyright statute states that

> [T]he fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified by that section, for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright.[114]

Whether or not use of copyrighted material can be deemed fair use is fact specific and subject to a four factor test: the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes; the nature of the copyrighted work (whether entertainment or factual works); the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and the harm or potential for harm upon the potential market for or value of the copyrighted work. [115]

Each of these factors is subject to application to the facts of each case, making outcomes difficult to predict. It is an emerging practice for some copyright holders of artistic works to use the threat of a lawsuit to defend fair use as an incentive to extract a royalty fee, even when the use of the work seems to fall under the fair use exception. The Chilling Effects archive was created to document cease and desist letters from copyright holders and educate the public and their legal rights.[116] The purpose of the archive is to stand as a bulwark against the "chilling effect" these cease and desist letters may be having on free speech, in particular online speech. People may shy away from using the copyrighted scientific work, due to the ambiguity in classifying use of a work as fair, even when they are justified in doing so under a fair use exemption.

The express mention of fair use exceptions for research and scholarship in the statute might encourage scholars to use copyrighted work more readily, but there is no clear case law on reusing an entire copyrighted body of code for research, or a dataset selected and arranged by another researcher for the creation of scholarship and it seems dubious the courts would uphold these as fair use under the third factor. How far the fair use exception extends into entire research compendia is not clear since the contours of fair use of copyrighted scientific material are not clearly delineated. Although the scope of fair use is generally broader for scientific work, the *Texaco* case indicates that mere scientific use of copyrighted materials will not satisfy the Fair Use test.[117] Scientists may prefer to avoid copyrighted works due to the ambiguity of their legal status, which is less likely to be a concern if the copyright holders waive their rights with the use of a license such as CC BY or the Modified BSD.[118]

---

[114] 17 U.S.C. § 107 (2007).

[115] *Id. See also* Pamela Samuelson, *Copyright's Fair Use Doctrine and Digital Data*, 11 PUBLISHING RESEARCH QUARTERLY 27 (1995).

[116] Chilling Effects Clearinghouse, http://chillingeffects.org/

[117] American Geophysical Union v. Texaco, 37 F.3d 882 (2d Cir. 1994).

[118] This is reinforced by the upholding of a license similar to the Creative Commons licenses in Jacobsen v. Katzer, USCAFC No. 2008-1001, Aug 13, 2008.

V. CONCLUSION

The NSF goal that publicly funded research be made publicly available achieves important objectives: accountability and oversight in the use of government funds; the possibility of increased recognition and citation; promotion of scientific knowledge through both 1) direct conveyance of research details and 2) facilitation of the opportunity to verify and improve upon scientific results; and an addressing of the credibility crisis in modern research. A licensing structure that can protect and promote these goals by aligning the scientific researcher's interests with society's interests in furthering scientific research as a whole could improve participation by scientists in collaborative research, encourage citizen-scientists to actively engage in research, and institutionalize the 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.[119]

Scientific computation is developing a central role in the scientific method, but progress and credibility are stunted by the fact that the entire research compendia – including code and data – are not being routinely made available to others.[120] Legal barriers are not the only block to the sharing of research, but 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 open software licenses for the code, Creative Commons attribution protection for text, documentation, figures and other media, including dataset creation methodologies, and creates a standard for replicability of scientific work. The creation of a standard would allow policy makers, administrators, or grant-giving agencies such as the National Science Foundation to require the release of compendia that qualify under the RRS, making publicly funded research be publicly available. A common platform for data and code release, and tools to aid in the process of creating work that is geared toward facilitation of reproducibility are the next steps in encouraging verifiability.[121]

The RRS licensing structure frees the scientific research for verification and incorporation into other scientific projects. These twin pillars of the Reproducible Research Standard are designed to achieve the scientific ideal of reproducibility and promote the viability of computational research as a core element of scientific enterprise.

---

[119] 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." Richard M. Stallman, *Releasing Free Software if You Work at a University, in* FREE SOFTWARE, FREE SOCIETY. SELECTED ESSAYS OF RICHARD M. STALLMAN 63, 64 (Joshua Gay ed., Free Software Foundation, Boston 2002), *available at* http://www.gnu.org/philosophy/fsfs/rms-essays.pdf (last visited Mar. 30, 2009).

[120] *See* Donoho et al, *supra* note 1.

[121] For example, scientists may not have the resources to download or analyze large datsets. The Biocep project offers a way to view and interact with the data while it resides on a central server. *See* Biocep-R, Statistical Analysis Tools for the Cloud Computing Age, http://biocep-distrib.r-forge.r-project.org/. I am indebted to David Purdy for alerting me to this example.