The Legal Framework for Reproducible Scientific Research
Licensing and Copyright

Victoria Stodden

As computational researchers increasingly make their results available in a reproducible way, questions naturally arise regarding copyright, subsequent use and citation, and ownership rights in general. The author proposes the Reproducible Research Standard for all components of scientific researchers' scholarship, which should encourage replicable scientific investigation through attribution, facilitate greater collaboration, and promote engagement of the larger community in scientific learning and discovery.

In the US, when scientists put their original research on the Web, it automatically falls under copyright. However, copyright is an unsuitable legal structure for scientific works. Scientific norms guide scientists to reproduce and build on others' research, and default copyright law by its very purpose runs counter to these goals. In this article, I present a methodology for scientists to rescind copyright from their work in such a way that it realigns scientific information sharing with long-established scientific norms.

Options for Researchers

Computational research is becoming more pervasive across a growing number of

fields—for example, in the June 1996 issue of *The Journal of the American Statistical Association*, nine of 20 articles were computational, whereas a decade later, in the June 2006 issue, 33 of 35 were computational. Different journals have different agreements with article authors, but most require authors to relinquish their ownership rights to articles, including copyright. Thus authors have very little or—more typically—no say in how their work is used after publication; they also find that it's frequently bound away in journals that can be very expensive to access. This is especially tough for computational work because researchers often need more than just the published article to reproduce the results.

Although this trend seems to be changing, the typical journal publishing format doesn't allow for transmission of supporting files such as accompanying images, source code, or demonstrations of work to interested readers—although some journals are beginning to require that code or data be released as a precondition for publication, (for example, *Nature* [www.nature.com/authors/ editorial_policies/availability.html], *The Insight Journal* [www.insight-journal.org], and *Annals of Internal Medicine*]www.annals.org/cgi/content/ full/0000605-200703200-00154v1]). Evidence exists that reproducible research receives more citations than nonreproducible work,[1,2] and releasing research on the Web is a growing trend that seems to be gathering institutional support. On 12 February 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 (www.fas.harvard.edu/~secfas/ February_2008_Agenda.pdf, p. 3).

However, Stuart M. Shieber, the Harvard University computer science professor who proposed the new policy, said in a press release that the decision "should be a very powerful message to the academic community that we want and should have more control over how our work is used and disseminated" (www.news.harvard.edu/ gazette/2008/02.14/99-fasvote.html). Stanford's School of Education soon followed suit with a mandate for open access: All faculty members will deposit a copy of their published work in an open access repository as of 26 July 2008 (www. news.harvard.edu/gazette/2008/02.14/99-fasvote. html and www.earlham.edu/~peters/fos/2008/06/ oa-mandate-at-stanford-school-of-ed.html).

General concern clearly exists about ownership rights to scholarly research. Tension arises because the scientific ethos guides scientists to both reproduce previous results and build on them, thereby generating further scientific understanding. Copyright

stands as a bar preventing the open sharing, dissemination, and use of work. To rescind copyright on scientific research, you must actively choose to do so, and this is typically done with a license.

Licensing and rescinding copyright

Copyright is a set of rights that attach by default to "original works of authorship" although not to the underlying idea or discovery (www.copyright.gov/title17). Another option, quite different from copyright, is the patent, granted by the US Patent Office only after reviewing an invention to make sure it's relevant, useful, and nonobvious. Whereas authors don't need to apply for copyright protection because it "follows the author's pen across the page,"[3] inventors must apply for a patent. Just as a copyright protects an author from plagiarism, a patent's raison d'être is to open the knowledge publicly while granting the right to exclude others from making, using, offering for sale, or selling the invention in the United States (35 USC 154(a)(1)). However, copyright and patent laws work counter to prevailing scientific norms— copyright was intended to give authors of creative works (literature and music, for example) exclusive rights, such as the right to be credited, to determine who may adapt or perform the work, or who may benefit financially from it. A *derivative work* is one that's "based upon one or more preexisting works" and gives the original copyright holder the exclusive right to prepare such works (17 USC 106(2)). A scientific contribution is considered valuable if, among other things, researchers can reproduce the results successfully (verifiability), and the work is built upon it, thus uncovering new scientific discoveries. Copyright stands in

the way of both actions.

As explained earlier, open licenses offer an option to override default copyright law. Two of the most common types of open licenses that rescind copyright are those designed for code (for example, the GNU Public License or GPL and the Berkeley Software Distribution or BSD license) or media (for example, the family of Creative Commons licenses).

Licenses for Code

Because copyright extends to code, Richard Stall-man began the Free Software movement in the early 1980s to encourage programmers to release their source code along with the software compiled for end users (www.gnu.org/gnu/thegnu project.html). His license, which became the GPL, 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.

In brief, Stallman's license has a *viral* effect de- signed to propagate the release of all source code. This means that if you use GPL-licensed code in the development of another body of code, your entire work must also carry the GPL unless you negotiate an alternative with the original's copyright holders. This is the *Share Alike* provision of the license.

The Modified BSD license is an attribution license and doesn't contain the Share Alike provision. Software under a BSD license retains the original license when it's used

in a derivative work, but the entire derivative work doesn't necessarily become BSD-licensed unless the downstream author chooses to do so.

Typically, the computational researcher releases code that consists of instruction scripts for a proprietary compiled language, rather than a compiled binary, although it might require proprietary binary code to run. There has been a marked increase in the use of these types of quantitative programming environments such as Matlab, SPSS, SAS, R, and Stata for experimentation and data display across a variety of fields. To rescind copyright from these instruction scripts, a license for code would be used. But all of a computational scientist's research can be considered code—for example, figures, articles, data structures, and even pseudocode descriptions wouldn't be classified as code. Stallman created the GNU Free Documentation license to cover the documentation that accompanies code, but it wasn't intended to extend to media beyond text.

Creative Commons

In 2001, Larry Lessig founded Creative Commons to enable greater sharing of works and information. The Creative Commons attribution license (CC BY) was designed for nonsoftware works to "share your creations with others and use music, movies, images, and text online that's been marked with a Creative Commons license" (http://creative commons.org/learnmore). Creative Commons states explicitly that its licenses aren't intended to cover code (see www.fsf.org/licensing/license/ agpl-3.0.html). Because Creative Commons licenses are designed for media and not to code, they make no reference to source code. Among other things, this is so as not to create license incompatibilities and because of the lack of patent provisions in the Creative Commons

licenses (patent is not an issue generally associated with media). Many Creative Commons licenses also incorporate Stallman's notion of viral attachment through their own Share Alike concept:

> If you alter, transform, or build upon this work, you can distribute the resulting work only under the same or similar license to this one (see http:// creativecommons.org/licenses/by-nc-sa/2.0, for example). The Creative Commons CC BY license ensures attribution but doesn't contain the Share Alike component.

The Share Alike Provision in the Scientific Context

The Share Alike concept is inappropriate in the scientific context because it can impose limits on the use and reuse of others' work, which in the scientific context, should be avoided whenever possible. This would render the research unavailable for reuse by people who don't want 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. In the scientific context, the license should attach only to the derivative work's components that the original author actually carried out, which isn't possible under a Share Alike provision. 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 (sciencecommons.org/projects/ publishing/open-access-data-protocol/) embodies many of these arguments.

The Gaping Hole

The US National Science Foundation (NSF) funds a large proportion of scientific academic research: In 2006, federally funded science and engineering R&D comprised 63 percent of total academic research and development support.[4] The NSF requires that researchers make available any data and other supporting materials for the research it funds to other researchers at no more than incremental cost.[5] So how do we, as academics, comply with our funding mandate to release our research publicly? And less broadly, how do we, as computational researchers, comply with the validation and verification demands as part of the scientific method?

Jon Claerbout, a Stanford geophysics professor, established a principle that many prominent researchers are now following: "An article about computational science in a scientific publication isn't the scholarship itself, it's merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures."[6] David Donoho and his research group have practiced this idea of reproducible research for the past 15 years with the release of the Matlab toolboxes Wavelab (www-stat.stanford.edu/~wavelab), Beamlab

(www-stat.stanford. edu/~beamlab), Symmlab (www-stat.stanford. edu/~symmlab), and Sparselab (http://sparselab. stanford.edu/). These software packages let anyone with access to Matlab reproduce figures from their articles, inspect the source code, change parameters, and access their datasets. This isn't yet a common phenomenon, but researchers are increas ingly proceeding this way, including Jalal Fadili at ENSI Caen (www.greyc.ensicaen.fr/~jfadili/ software.html) and scientists at the Audio Visual Communications Lab (LCAV) at Ecole Polytechnique Fédérale de Lausanne, (www.lcavwww.epfl. ch/reproducible_research/).

A tenet of the scientific method holds that every research finding should be reproducible before it becomes accepted as a genuine contribution to human knowledge. But how to encourage this? One way could be for scientists to use a license that covers their entire research—for example, all the components required for reproducibility. As computational research becomes more pervasive, details of the work often remain unpublished, so 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, a key to progress in modern science, could end up undermining the scientific process and become "the last refuge of the scientific scoundrel."[7] Two blocks exist to truly reproducible research: the lack of reward for producing reproducible work (norms in the scientific community) and the legal obstacle to the full sharing of methodologies, writing, code, papers, and data (copyright law). I propose the Reproducible Research Standard (RRS) to address the second block:

problems of copyright and reproducibility in scientific research.[8]

The Reproducible Research Standard: Enabling Reproducibility

Computational research produces an entire research compendium,[9] which comprises:

1. *the research paper*—including all the source files from which the manuscript was built (for example, LaTex, Word, or WordPerfect files);
2. *the data*—including documentation completely describing the data (sources, components, and interpretation); a description of how the data was brought into the form used in the research; the code and instructions used to bring the data into the form used in the research; and documentation of any code used for data processing;
3. *the experiment*—the code and instructions used in the experiment, including all source code; documentation of any code used, including pseudocode; a clear listing of the parameters, settings, and operating system dependencies used to achieve the results described in the paper; and a clear description of the experimental methodology;
4. *the results of the experiment*—any figures, data, or the like produced from the experiment; any illustration source files; and documentation and explanation of the processing of the experimental results; and
5. *any auxiliary material*—materials used for presentation on the Web or an interface to the data or results; and documentation of any auxiliary code.

Typically, researchers release only the research paper, which is all that traditional journals usually publish. In contrast, to encourage scientists to release their entire compendium, there could be a licensing methodology that applies to every aspect except the data (discussed in the next section) and ensures attribution for any compendium elements used in derivative scientific research (meaning that any papers published using components of the research compendium must attribute the original author). Encouraging the release of the entire research compendium is the purpose of the RRS. I argue this is best done through ensuring attribution.

Because citations are important evidence of impact for scientists and often play a role in hiring and promotion decisions, the RRS is consistent with scientific norms. It can also offset researchers' fears that parts of their released compendium will be "stolen" or new publications made without attribution. The RRS requires attribution for any part of the upstream compendium used in derivative research and those (US) components of the downstream research carry the license. This is less restrictive than the GPL's Share Alike component or some of the Creative Commons licenses in that the RRS doesn't require all comingled works to carry the same license. One goal of the RRS is to ensure research attribution in any derivative compendium. The reason for not requiring the entire derivative compendium to carry the RRS, as would be required by the Share Alike component, is to encourage scientific research that builds on previous research without restriction and to remain consistent with the scientific ethos of attribution solely for work done.

Data under the RRS

Raw data aren't copyrightable, and thus it's meaningless to apply a copyright rescinding license to them. However, original *selection and arrangement* of the data are copyrightable, as are the original metadata associated with dataset production such as documentation, arrangement explanations, or data cleaning.[10,11] Thus, the RRS can rescind copyright from these aspects of the data process.

A license that applies to a database's selection and arrangement, in a virally attributive way, can encourage scientists to release the datasets they've compiled by providing a legal framework for attribution. A license that would protect their claim to authorship is an important tool to assuage researchers' concerns about loss of attribution and provide for greater transparency in dataset construction.

An adequate licensing structure doesn't exist that intentionally applies to the structures that house the data used. Although the raw facts aren't copyrightable, often researchers put a phenomenal amount of work into dataset preparation for research. Precisely how researchers generated or gathered the data, any processing they did to clean or verify them, and the data's current layout are all vital pieces of information for a scientist to reproduce or understand the final result. The fact that the RRS emphasizes the importance of transmitting dataset construction details dovetails neatly with Claerbout's aspirations of really reproducible research.[8]

The RRS Defined

The RRS suggests both licenses for the different aspects of the full research compendium and releasing data into the public domain. I believe that the appropriate choices in the scientific context are: attaching CC BY to the compendium's media

components, modified BSD to code components, and the Science Commons Database Protocol (http://commons.org/projects/publishing/ open-access-data-protocol) to the data if scientists choose to release their data to the public domain. The goal is to encourage scientists to release all components of their research, through viral attribution to help ensure that researchers receive credit for their work and provide a mechanism defining and promoting the idea of full compendium release.

As an umbrella licensing algorithm, the RRS is easier to use than the alternative—each time scientists release scholarship, they would have to fashion a combination of licenses from a spectrum of choices or accept the default full copyright status of their work. A corollary benefit to the RRS's relaxation of the Share Alike component is that it becomes easier for industry to employ research as part of a technology without having all the (possibly) proprietary work come under the RRS.

What Does This Mean for Scientific Researchers?

Rescinding copyright requires scientists to take active steps in securing the appropriate license for their compendia, but they can easily do this by placing a notification on their Web page that the compendia are under the RRS. They can also add machine-readable tags to the HTML that hosts the compendium's conponents to facilitate machine readability of attribution and other research facets, simplifying search and compendia element attribution.[12]

Copyright doesn't attach to data, so it doesn't make sense to attach a license rescinding copyright to the data themselves. However the RRS is a compilation of other

licenses and can treat individual compendium elements separately. This means, for example, that scientists wouldn't have to release private medical data, but they can still apply the appropriate license to release the other aspects of their work from copyright. Because the RRS is an amalgamation of commonly used existing licenses, there's no license proliferation with its introduction, and it's as compatible with other licenses as its component licenses.

Computational research is at a turning point—will it embrace the scientific values of reproducibility and verifiability? In such a young field, an immediate opportunity exists to set standards for quality and replicability in our work, but at the same time, we face the problem of working within a copyright structure that wasn't designed with scientific research in mind. The RRS is, in part, a tool to explain the meaning of reproducible research in the computational sciences and, as a result, help communicate scientific standards for acceptable practice. The RRS is also an important tool for encouraging scientific research by rescinding the aspects of copyright that prevent scientists from sharing important research information. If the scientific community adopts the license, we can solve the dual problems of standards for computational science and ensure that the scientific ethos in communicating and disseminating our work continues. A step forward for the computational science field would be if grant-giving agencies, such as the NSF, required the release of research compendia that qualify under the RRS. This would satisfy the requirement for researchers to make their publicly funded work available to the public as well as establish important structures and standards for the nature of computational research.

Acknowledgments

I'm very grateful for invaluable discussion with David Donoho, Danny Hillis, Larry Lessig, John Wilbanks, Wendy Seltzer, and Melanie Dulong de Rosnay. I also thank Peter Suber for his comments at the Yale Law School CyberScholar Series on 22 April 2008. Of course, any errors are mine alone.

References

1. D. Donoho, "How to Be a Highly Cited Author in the Mathematical Sciences," *in-cites*, Mar. 2002; www.in-cites.com/ scientists/DrDavidDonoho.html.

2. P. Vandewalle et al., "Experiences with Reproducible Research in Various Facets of Signal Processing Research," *Proc. IEEE Conf. Acoustics, Speech, and Signal Processing*, vol. 4, 2007, pp. 1253–1256.

3. E. von Hippel, *Democratizing Innovation*, MIT Press, 2005, p. 85.

4. Division of Science Resources Statistics, "Survey of Research and Development Expenditures at Universities and Colleges, FY 2006," US Nat'l Science Foundation, Sept. 2007; www. nsf.gov/statistics/infbrief/nsf07336.

5. "Grant General Conditions (GC-1) #38—Sharing of Findings, Data, and Other Research Products," US Nat'l Science Foundation, 1 June 2007; www.nsf.gov/pubs/policydocs/ gc1_607.pdf.

6. J. Buckheit and D. Donoho, *Wavelab and Reproducible Research*, tech. report, Dept. of Statistics, Stanford Univ., 1995.

7. R. LeVeque, "Wave Propagation Software, Computational Science, and Reproducible Research," *Proc. Int'l Congress Mathematicians*, M Sanz-Sole et al., eds., Aug. 22–30, 2006, p. 1227–1254

8. V. Stodden, *Enabling Reproducible Research: Open Licensing For Scientific Innovation*, International Journal of Communications Law and Policy, forthcoming 2009.

9. R. Gentleman and D. Lang, "Statistical Analyses and Reproducible Research," *J. Computational & Graphical Statistics*, vol. 16, no. 1, 2007 , p. 1–23.

10. *Feist Publications, Inc., v. Rural Telephone Service Company* , 499 U.S. 340, 1991, pp. 363–364.

11. M. Bitton, "A New Outlook on the Economic Dimension of the Database Protection Debate," *IDEA: The Intellectual Property Rev.*, vol. 47, June 2006, p. 93.

12. H. Abelson et al., "ccREL: The Creative Commons Rights Expression Language," Mar. 2008; http://wiki.creative commons.org/images/d/d6/Ccrel-1.0.pdf.

Victoria Stodden *is a research fellow at the Berkman Center for Internet and Society at Harvard University. Her current research includes understanding how new technologies and open source standards affect societal decision-making and welfare. Stodden has a PhD in statistics from Stanford University and an MLS from Stanford Law School. Contact her at vcs@ stodden.net; http://blog.stodden.net.*