The Legal Framework for Reproducible Research in the Sciences:
Licensing and Copyright

Victoria Stodden
Berkman Center for Internet and Society
Harvard Law School
Cambridge, MA 02138
vcs@stanford.edu

August 4, 2008

Abstract

As more and more computational researchers are making their results available in a reproducible way, and often outside the traditional journal publishing mechanism, questions naturally arise with regard to copyright, subsequent citation and use of the work, and ownership rights of the work in general. There is a gap in the current licensing and copyright structure for this growing number of scientists releasing their research publicly, particularly on the internet. Scientific research produces more than the final paper: the code, data structures, experimental design and parameters, documentation, figures, are all important for communication of the scholarship and replication of the results. I propose the Open Research License (ORL) for scientific researchers to use for all components of their scholarship. It is intended to encourage reproducible scientific investigation through attribution, facilitate greater collaboration, and promote engagement of the larger community in scientific learning and discovery.

Under American law, copyright automatically accrues to “original works of authorship” as soon as they are “in fixed form.” Since copyright is a mechanism designed to protect an author against unauthorized copying and distribution of his or her work for sale (think Disney’s copyright on Mickey Mouse or the copyright on the Happy Birthday song [5, 6]), it is a concept diametrically opposed to the fundamental premise of the academic world: that knowledge be openly released, shared, verified, and built upon. The ORL is a solution to this mismatch: it is designed to roll back certain aspects of the default copyright status for scientific works in order to allow for attribution and to encourage the creation of derivative works, i.e. further scientific discoveries.¹

1 Introduction

In this paper, I argue an appropriate license 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 ideal of reproducible research. Typically scientists hand ownership rights over their work to their publishing journal but with the spread of reproducible research more and more research is being communicated through non-traditional

¹I am very grateful for invaluable discussion with David Donoho, Danny Hillis, Larry Lessig, John Wilbanks, Wendy Seltzer, and Melanie Dulong de Rosnay. Of course, any errors are mine alone.
means, such as the internet, and thus can be subject to different legal frameworks, often up to the researchers to choose. This article outlines the copyright structures that can apply to researchers, various licensing options and their meanings, and identifies gaps that exist in the legal framework for scientists. The usual method of rescinding the default copyright that attaches to published works operates through the author’s adoption of a license for his or her work that modifies these rights. Typical options are the GNU Public License or the Berkeley Software Distribution License, both designed for code, or one of the Creative Commons licenses, designed for non-software works. Another option is the Open Research License, or ORL, designed specifically for scientific research that encompasses both code and media components [8]. In this paper I introduce the ORL and argue it is the appropriate way for scientists to release work.

2 Options for Researchers

Computational research is becoming more pervasive across a growing number of fields. For example, in the June 1996 issue of the flagship Journal of the American Statistical Association, nine of twenty articles were computational; and a decade later in the June 2006 issue, 33 of 35 were computational. Different journals have different agreements with article authors, but the majority require the authors to relinquish their ownership rights over articles, including all copyright. The authors thus have very little or, more typically, no say in how their work gets used after publication, and find it is frequently bound away in journals that can be very expensive for people to access. This is especially tough for computational work, since more than the published paper is often needed for reproducibility of the results.

Although there are signs this is changing, the typical journal publishing format doesn’t allow for such things as an interactive presentation of results or data and makes it difficult to transmit supporting files such as accompanying images, source code, or demonstrations of their work to interested readers.² There is some evidence that reproducible research receives more citations than non-reproducible work. [3, 9] Releasing research on the web is a trend that is growing and seems to have institutional support: On February 12 of this year, Harvard University’s Faculty of Arts and Sciences adopted a policy requiring faculty members to allow the university to make their scholarly articles available free 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 are not sold for a profit.³

After the Harvard decision, Stuart M. Shieber, the Harvard computer science professor who proposed the new policy, said in a news release that the decision “should be a very powerful

²Some journals are beginning to require the release of code and/or data as a pre-condition for publication. See e.g. Nature http://www.nature.com/authors/editorial_policies/availability.html, the Insight Journal http://www.insight-journal.org/, Annals of Internal Medicine http://www.annals.org/cgi/content/full/0000605-200703200-00154v1
message to the academic community that we want and should have more control over how our work is used and disseminated.”

There is clearly general concern about ownership rights over scholarly research. A tension is created since the scientific ethos is to reproduce previous results and build on them to generate further scientific understanding, not prevent the open sharing, dissemination, and use of work, as copyright does. To rescind copyright on scientific research, *one must actively choose to do so*. Typically this is done with a license.

### 2.1 Licensing and Rescinding Copyright

Copyright is a set of rights that attach to “original works of authorship” by default, but not to the underlying idea or discovery. Authors need not apply for copyright protection: it “follows the author’s pen across the page.” Copyright is a very different legal framework to a patent, which is obtained only after application and review by the Patent Office and determination that the invention is relevant, useful, and non-obvious. Patents are designed to encourage disclosure of knowledge publicly by providing legal protection to the inventor upon dissemination of the invention. A patent’s *raison d’être* is to grant the right to exclude others from making, using, offering for sale, or selling the invention in the United States. This approach is not well-suited to the academic sciences, where one mark of success is that others have built upon one’s contributions to scientific understanding. But doing nothing still brings at least part of a scientific compendium into the legal sphere since copyright is attached as a default to any work once it is “in fixed form.”

Copyright law works counter to prevailing scientific norms. It was intended to give the author of creative works, such as literature or music, exclusive rights like the right to be credited, to determine who may adapt or perform the work, or who may benefit financially from it. In scientific research, a contribution is understood as valuable if, among other things, other researchers are able to copy it successfully (verifiability) and if it encourages researchers to build on it and uncover new scientific discoveries. Both are prohibited by copyright. Copyright law defines a *derivative work* as one that is “based upon one or more preexisting works” and gives the original copyright holder the exclusive right to prepare such works (17 U.S.C. 106(2)). By contrast, scientific research norms require others to be able to access and build upon previous work without legal encumbrance.

Two of the most common types of open licenses using copyright to enforce openness rather than exclusivity, are those designed for code (for example, the GNU Public License (GPL) and the Berkeley Software Distribution license (BSD)), and those designed for media (the family of Creative Commons licenses).

#### 2.1.1 Licenses for Code: The GNU Public License (“Copyleft”) and the Berkeley Software Distribution License

Since the copyright default 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

---


5[^5]: Stanford’s School of Education has 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 July 26, 2008. See [http://www.earlham.edu/~peters/fos/2008/06/oa-mandate-at-stanford-school-of-ed.html](http://www.earlham.edu/~peters/fos/2008/06/oa-mandate-at-stanford-school-of-ed.html)

6[^6]: See 17 U.S.C 102 and [http://www.copyright.gov/title17/](http://www.copyright.gov/title17/)

7[^7]: E. von Hippel, Democratizing Innovation at 85.

8[^8]: USC Title 35 §154(a)(1)
compiled for end users. His license, the GPL, has two main components:

1. if publicly distributed, all software subject to the license must also have its source code released,

2. once the license is attached to code, it also attached to any body of code that uses the original code.

In brief, Stallman’s license has a “viral” effect designed 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. The key for a computational researcher is that this license applies to work “as long as it is clear what constitutes the ‘source code’ for the work.” The GPL is intended for code.

The Berkeley Software Distribution (BSD) License is an attribution license and does not contain the Share Alike provision. Software under a BSD License retains the license when used in a derivative work, but the entire derivative work does not necessarily become BSD licensed unless the downstream author chooses to do so.

Typically the computational researcher will release code consisting 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 huge increase in the use of such quantitative programming environments as Matlab, SPSS, SAS, R, and Stata, for experimentation and data display. Not all the research of a computational scientist can be considered “source code,” for example, figures, papers, data structures, even pseudocode descriptions would not fall under the GPL. Stallman has also created the GNU Free Documentation License but this is designed for the documentation accompanying code and also doesn’t extend to media beyond text, such as figures.

2.1.2 Creative Commons

In 2001 Larry Lessig founded Creative Commons to lower licensing transaction costs to enable a greater sharing of works and information. The Creative Commons non-commercial attribution license, for example, is designed for non-software works: to “share your creations with others and use music, movies, images, and text online that’s been marked with a Creative Commons license.” Creative Commons states explicitly that its licenses are not intended to cover code so as not to create further license incompatibilities, and because of the lack of patent provisions and the lack of reference to “the preferred form of the work for making modifications to it” (ie. source code). “[W]e do not recommend that you apply a Creative Commons license to software code.” Many of the Creative Commons licenses also incorporate Stallman’s notion of viral attachment through their Share Alike concept: “If you alter, transform, or build upon this work, you may distribute the resulting work only under a license identical to this one.”

---

10http://www.gnu.org/licenses/gpl-faq.html#GPLOtherThanSoftware
11I refer to the New BSD License - without the UC Berkeley advertising clause
12Other licenses exist, such as the Open Software License or the Academic Free License, but the former retains the reciprocal disclosure Share Alike component of the GPL and the latter is redundant with version 2 of the Apache License. See http://www.opensource.org/proliferation-report
13GPLv3 http://www.fsf.org/licensing/licenses/agpl-3.0.html
14http://wiki.creativecommons.org/FAQ
The Creative Commons CC BY license ensures attribution but does not contain the Share Alike component.

2.1.3 The Inappropriateness of the Share Alike Provision for Scientific Research

I argue the Creative Commons licenses alone are not well-suited to the needs of computational scientists. Not only are they poorly suited to code, but the Share Alike concept is inappropriate in the scientific context. It could serve to discourage scientific research and confuse attempts at attribution. By requiring that any scientific research that builds on research covered by a license also carry the license, as Share Alike does, the bar to scientific investigation and discovery is raised. Research becomes unavailable to researchers who do not wish to license their compendia with the same license as the upstream work was under. The Share Alike condition may deter researchers from following on. Under Share Alike, it is also no longer clear how to attribute upstream work in a derivative product since work by different authors could be subsumed and conflated under a single attribution scheme. To be consistent with current scientific attribution norms, the viral attribution component of the license should attach only to the component of the derivative work actually carried out by the original author, which is not possible under a Share Alike provision. Scientists should be free to license the original components of their compendia as they see fit, and not be restricted in their licensing by the fact of having built on another’s licensed work by, for example, using or modifying another researcher’s code.

3 The Gaping Hole: Computational Scientific Compendia are not Covered by a License

The National Science Foundation (NSF) funds the majority of scientific academic work in the United States: in 2006 federally funded science and engineering R&D comprised 63% of total academic research and development support. 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. 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 demands of validation and verification as part of the scientific method? As described in another paper in this edition (15 Years of Reproducible Research in Harmonic Analysis) computational research is ideally suited to reproducible research and there is a movement underway to establish it as the paradigmatic scientific model of conducting computational research. The default option, copyright, works in the opposite direction as it prevents the active sharing of research and the creation of derivative scientific works. The established work-arounds for this, the software li-

---


16 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.

licenses and the Creative Commons licenses discussed previously, are trying to fix problems for code and for media distribution respectively, and are not suitable for scientific research which entails a unique hybrid of code and media and other products, and a unique set of goals.

Many prominent researchers are following the lead established by Jon Claerbout, a Stanford geophysics professor, to whom Claerbout’s principle has been attributed: “[a]n article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.”[2] This encapsulates the idea of reproducible research. David Donoho and his research group have been practicing this form of reproducible research for at least fifteen years with the release of the Matlab toolboxes WaveLab, BeamLab, SymmLab, SparseLab, to name the most prominent ones.17 These software packages allow anyone with access to Matlab to reproduce figures from their papers, inspect the source code, change parameters, and access their datasets. This is not yet a common phenomenon but increasing numbers of researchers are proceeding this way, such as Jalal Fadili at ENSI Caen18 and the Audio Visual Communications Lab (LCAV) at Ecole Polytechnique Fédérale de Lausanne19, for example.20

Reproducible research is an ideal way of encouraging quality computational research: Any research finding should be reproducible before it becomes accepted as a genuine contribution to human knowledge. But how to encourage this?

We need a license designed with the needs of computational researchers in mind. It needs to cover the entire research product and ensure attribution for researchers whose work is built upon. It also should encourage scientific discovery and reproducibility and support the efforts of scientists in these areas. As computational research is becoming more pervasive, details of the work often remain unpublished and the opportunity to hide poor scholarship increases. Without full publication of “a careful description of the methods used, in sufficient detail that others can attempt to repeat the experiment,” computational research, a key to progress in modern science, could end up undermining the scientific process and becoming “the last refuge of the scientific scoundrel.” [7] There are two blocks to really reproducible research: the lack of reward for producing reproducible work (norms in the scientific community) and a legal obstacle to the full sharing of methodologies, writing, code, papers, and data. This paper proposes a solution to the latter problem.

4 A Solution: The Open Research License

I propose the Open Research License (ORL) to address problems of copyright and reproducible research in scientific research. [8] Computational research produces an entire Research Compendium [4], which comprises:

   (a) Including all the source files from which the manuscript was built: e.g. TeX, Word, or WordPerfect files.

18http://www.greyc.ensicaen.fr/~jfadili/software.html
19http://lcavwww.epfl.ch/reproducible_research/
20For further details on these package and the rationale for Reproducible Research in the computational sciences see the article “15 Years of Reproducible Research in Harmonic Analysis” in this issue.
2. The Data:
   (a) The data itself.
   (b) Documentation completely describing the data: Sources, components, and interpretation.
   (c) A description of how the data was brought into the form used in the research.
   (d) The code and instructions used to bring the data into the form used in the research.
   (e) Documentation of any code used in this process.

3. The Experiment:
   (a) The code and instructions used in the experiment, including all source code.
   (b) Documentation of any code used, including pseudocode.
   (c) A clear listing of the parameters, settings, and operating system dependencies under which the code was used to achieve the results described in the paper.
   (d) A clear description of the experimental methodology.

4. Results of the Experiment:
   (a) Any figures, data, or the like produced from the experiment. These appear in full, as produced by the experiment and described in the research paper, (ie. high resolution figures) since it is usually not possible to include them in the research paper directly. Any illustration source files should be included: e.g. Abode Illustrator.
   (b) Documentation and explanation of the experimental results.

5. Auxiliary material:
   (a) Code used for presentation on the web or an interface to the data or results.
   (b) Documentation of auxiliary code.

Typically the research paper alone is all that is released, and this is usually all that is published by traditional journals. By contrast, to encourage scientists to release their entire compendium, the ORL applies to every aspect, except the data (discussed in the next section) and ensures attribution for any elements of the compendium used in downstream scientific research. This means that any papers published using components of the research compendium must cite the original author. Since citations are important evidence of impact for scientists, and often play a role in hiring and promotion decisions, the ORL can offset researchers’ fears that parts of their released compendium will be “stolen” and publications made without citation. There is also evidence that papers that are fully reproducible are cited more frequently.

The ORL requires attribution for any part of the compendium used in derivative research and requires that component of the downstream research to carry the license. This is less restrictive than the Share Alike component of the GPL or some Creative Commons licenses in that the ORL does not require all co-mingled works to carry the same license. The goal is to ensure attribution for research in any derivative compendium. The reason for not requiring the entire derivative compendium to carry the ORL, as would be required by the Share Alike component, is to encourage scientific research that builds upon previous research without bar and for consistency with the scientific ethos of attribution solely for work done. Ideally, the downstream researcher will choose to license the original components of his or her compendium under the ORL, but encouragement of scientific research must be primary over encouragement of license use. Specifically, there must be no bar to building upon previous scientific research.
4.1 Data Under the ORL

Raw data is not copyrightable and thus it is meaningless to apply the ORL to it. But original meta-data associated with the production of the dataset (such as documentation, explanations of arrangement and data cleaning) as used in the current research is subject to copyright and will fall under the purview of the ORL. [1] The meta-data comprises 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 of the data, whether the data is generated synthetically for this paper or obtained from a data collection process, and enumerating any changes made to the dataset. Specifically, original “selection and arrangement” of the data is copyrightable can thus protected by the ORL. This means that use of a particular incarnation of the data must be attributed to the creator of that form in downstream research.

A license that applies to the “selection and arrangement” of a database, in a virally attributive way, can encourage scientists to release the datasets they have compiled by providing a legal framework for attribution. Because of researchers’ reluctance to “lose” the work they have done creating the dataset, a license protecting their claim to authorship is an essential tool to assuage concerns about loss of attribution and provide for both greater transparency in dataset construction and to encourage release of the dataset.

There is no adequate licensing structure that intentionally applies to the structures that house the data used. Although the data are not copyrightable, often a phenomenal amount of work goes into preparation of the dataset for research this could be attributed to the scientist and explained openly to future researchers who would like to use these data. Precisely how the data were generated or gathered, any processing done to the data to clean or verify it, and the current layout of the data are all vital pieces of information for a scientist to reproduce or understand the final result. This aspect of the ORL emphasizes the importance of transmitting the dataset construction details and this dovetails neatly with Claerbout’s aspirations of Really Reproducible Research. See [8] for further details.

4.2 The Open Research License Defined

I propose the Open Research License as a compilation: attaching CC BY (the Creative Commons attribution license that does not have a Share Alike provision) to the media components of the compendium, the BSD license to the code components, and if the scientist chooses to release his or her data to the public domain, attaching the Science Commons Database Protocol to the data. This allows for viral attribution and, by avoiding the Share Alike aspect explained above, ensures each scientist is attributed for only the components of the compendium he or she has created. To promote scientific research, it is sensible to allow the downstream researcher the choice of whether he or she would like to attach the ORL to his or her compendium (although the ORL remains attached to any elements of the upstream compendium he or she may have incorporated).

As an umbrella license the ORL is easier to use than the alternative: Without the ORL, each time a scientists releases scholarship, he or she would have to fashion together a combination of licenses from an spectrum of choices.

Incentives in the scientific community focus on citation: receiving recognition for research work and contributions made to scientific discovery. Within the citation ethos are embedded incentives to produce research that leads to further citations i.e. uncover deep problems and create results useful to other scientists. The ORL supports this ethos and the notion of reproducible

---

research through viral attribution.

A corollary benefit to the ORLs relaxation of the Share Alike component is that it becomes easier for industry to employ the research as part of their technology without having all their (possibly) proprietary work come under the ORL.

5 What Does This Mean for a Scientific Researcher?

Rescinding copyright requires the scientist to take the active step of securing the appropriate license for his or her compendium – I argue the ORL – and this can be easily done by notification on the webpage that the compendium is under the ORL. A machine-readable tag can be added to the html that hosts the compendium.22 Embedding tags in html code facilitates machine readability of attribution and other facets of the research, which simplifies search and attribution of elements of compendia.23 A webpage will be offered that steps scientists through a series of questions about their compendium and provides the appropriate tags for each component.

Since there is no copyright attached to data, it does not make sense to attach a license rescinding copyright. Since the ORL is a compilation of other licenses, individual elements of the compendium can be subjected to the ORL separately. This means that, say, private medical data would not have to be released for the ORL to be used on the rest of the compendium.

Since the ORL is an amalgamation of commonly used existing licenses, there is no license proliferation with its introduction and it is as compatible as its component licenses.

6 Conclusion

Computational research is at a turning point. In such a young field, there is an immediate opportunity to set standards for quality and verifiability in our work, but at the same time we have the problem of working within a copyright structure that was not designed with our research in mind. The Open Research License is in part an explanatory tool for the meaning of reproducible research in the computational sciences, and thereby helps to communicate scientific standards for acceptable practice. The ORL is also an important tool for fostering good: encouraging scientific research by rescinding the aspects of copyright that prevent the sharing of important research information. By its adoption we can solve the dual problems of standards for computational science and ensuring the continuation of the scientific ethos in the communication and dissemination of our work. A step forward for the field of computational science would be the requirement of release of any research compendia funded by grant giving agencies such as the NSF. This would satisfy the requirement for publicly funded work to be made publicly available, as well establishing important structure and standards for the nature of computational research.

References


23This point was made to me by Peter Suber at the Yale Law School CyberScholar Series on April 22, 2008.


