Intellectual Property and Computational Science

Victoria Stodden

Abstract This chapter outlines some of the principal ways United States Intellectual Property Law affects the sharing of digital scholarly objects, particularly for those who wish to practice reproducible computational science or Open Science. The sharing of the research manuscript, and the data and code that are associated with the manuscript, can be subject to copyright and software is also potentially subject to patenting. Both of these aspects of Intellectual Property must be confronted by researchers for each of the these digital scholarly objects: the research article; the data; and the code. Recommendations are made to maximize the downstream reuse utility of each of these objects. Finally, this chapter proposes new structures to manage Intellectual Property to accelerate scientific discovery.

Introduction

A deep digitization of the scientific enterprise is taking place across the research landscape and generating new ways of understanding our surroundings. As a result, our stock of scientific knowledge is now accumulating in digital form. Our DNA is encoded as genome sequence data, scans of brain activity exist in functional magnetic resonance image databases, and records of our climate are stored in myriad time series datasets—to name but a few examples. Equally as importantly, our reasoning about these data is recorded in software, in the scripts and code that analyze and make sense of our digitally recorded world. Sharing the code and data that underlie scientific findings is a necessary step to permit the transfer of knowledge embodied in the results, so that they can be independently verified, re-used, re-purposed, understood, and applied in new areas to solve new problems.

V. Stodden (✉)
Columbia University, Manhattan, USA
e-mail: victoria@stodden.net

S. Bartling and S. Friesike (eds.), *Opening Science*,
DOI: 10.1007/978-3-319-00026-8_15,

The inability to access scientific data and code stands as a barrier to the verification of scientific findings, and as a barrier to the knowledge transfer needed to both facilitate scientific advancement and spurn innovation and entrepreneurship around scientific findings (Stodden 2011).

These computational advances have taken place in parallel with the development of the Internet as a pervasive digital communication mechanism, creating an unprecedented opportunity to broaden access to scientific understanding. In this chapter I describe Intellectual Property barriers to the open sharing of scientific knowledge, and motivate solutions that coincide with longstanding scientific norms. In "Research Dissemination: The Narrative", I frame scientific communication as a narrative with a twofold purpose: to communicate the importance of the results within the larger scientific context and to provide sufficient information such that the findings may be verified by others in the field. With the advent of digitization, replication typically means supplying the data, software, and scripts, including all parameter settings and other relevant metadata, that produced the results (King 1995; Donoho et al. 2009). Included in this discussion is the importance of access to the primary research narrative, the publication of the results. "Research Dissemination: Data and Raw Facts" and "Research Dissemination: Methods/Code/Tools" then discuss Intellectual Property barriers and solutions that enable data and code sharing respectively. Each of these three research outputs, the research article, the data, and the code, require different legal analyses in the scientific context.

Research Dissemination: The Narrative

A typical empirical scientific workflow goes something like this: a research experiment is designed to answer a question; data are collected, filtered, and readied for analysis; models are fit, hypotheses tested, and results interpreted; findings are written up in a manuscript which is submitted for publication. Although highly simplified, this vignette illustrates the integral nature of narrative, data, and code in modern scientific research. What it does not show is the limited nature of the research paper in communicating the many details of a computational experiment and the need for data and code disclosure. This is the subject of the sections "Research Dissemination: Data and Raw Facts" and "Research Dissemination: Methods/Code/Tools." This section motivates the sharing of the research paper, and discusses the conflict that has arisen between the need for scientific diss00emination and modern intellectual property law in the United States.

A widely accepted scientific norm, labeled by Robert K. Merton, is *Communism* or *Communalism* (Merton 1973). With this Merton described an ideal in scientific research, that property rights extend only to the naming of scientific discoveries (Arrow's Impossibility Theorem for example, named for its originator Kenneth Arrow), and all other intellectual property rights are given up in exchange for

recognition and esteem. This idea underpins the current system of publication and citation that forms the basis for academic rewards and promotions. Results are described in the research manuscript which is then published, typically in established academic journals, and authors derive credit through their publications and other contributions to the research community. They do not receive financial or other material rewards beyond recognition by peers of the value of their contributions. There are many reasons for the relinquishment of property rights over discoveries in science, but two stand out. It is of primary importance to the integrity of our body of scientific knowledge that what is recognized as scientific knowledge has as little error as possible. Access not just to new discoveries, but also to the methods and derivations of candidates for new knowledge, is imperative for verification of these results and for determining their potential admission as a scientific fact. The recognition that the scientific research process is error prone—error can creep in at any time and in any aspect of research, regardless of who is doing the work—is central to the scientific method. Wide availability increases the chances that errors are caught - "many eyes make all bugs shallow." The second reason Intellectual Property rights have been eschewed in scientific research is the historical understanding that scientific knowledge about our world, such as physical laws, mathematical theorems, or the nature of biological functions, is not subject to property rights but something belonging to all of humanity. The U.S. federal government granted more than $50 billion dollars for scientific research last year in part because of the vision that fundamental knowledge about our world isn't subject to ownership but is a public good to be shared across all members of society.[1] This vision is also reflected both in the widespread understanding of scientific facts as "discoveries" and not "inventions," denoting their preexisting nature. Further, current intellectual property law does not recognize a scientific discovery as rising to the level of individual ownership, unlike an invention or other contribution. Here, we focus on the interaction of intellectual property law and scientific research article dissemination.

Copyright law in the United States originates in the Constitution, when it states that "The Congress shall have Power … To promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries".[2] Through a series of laws and interpretations since then, copyright has come to automatically assign a specific set of rights to original expressions of ideas. In the context of scientific research, this means that the written description of a finding is copyright to the author(s) whether or not they wish it to be, and similarly for code and data (discussed in the following two sections). Copyright secures exclusive rights vested in the author to both reproduce the work and prepare derivative works based upon the original. There are exceptions and limitations to this power, such as Fair Use, but none of these provides an intellectual property framework for scientific knowledge that is

[1] The Science Insider: http://news.sciencemag.org/scienceinsider/budget_2012/

[2] U.S. Const. art. I, §8, cl. 8.

concordant with current scientific practice and the scientific norms described above. In fact far from it.

Intellectual property law, and how this law is interpreted by academic and research institutions, means that scientific authors generally have copyright over their research manuscripts. Copyright can be transferred, and in a system established many decades ago journals that publish the research manuscripts typically request that copyright be assigned to the publisher for free as a condition of publication. With some notable exceptions, this is how academic publication continues today. Access to the published articles requires asking permission of the publisher who owns the copyright owner, and usually involves paying a fee. Typically scientific journal articles are available only to the privileged few affiliated with a university library that pays subscription fees, and articles are otherwise offered for a surcharge of about $30 each.

A transformation is underway that has the potential to make scientific knowledge openly and freely available, to everyone. The debate over access to scientific publications breaks roughly into two camps. On one side are those who believe tax-payers should have access to the fruits of the research they've funded, and on the other side are those who believe that journal publishing is a business like any other, and the free market should therefore be left unfettered.[3] The transformation started in 1991 when Paul Ginsparg, Professor of Physics at Cornell University, set up an open repository called arXiv.org (pronounced "archive") for physics articles awaiting journal publication. In the biosciences, a new publishing model was brought to life in 2000—Open Access publishing—through the establishment of the Public Library of Science, PLoS.[4] PLoS publishes scientific articles by charging the authors the costs upfront, typically about $1300 per article, and making the published papers available on the web for free.[5] The PLoS model has been extraordinarily successful, gaining in prestige and publishing more articles today than any other scientific journal.[6]

The U.S. government has joined in this movement toward openness in scientific literature. In 2009 the National Institutes for Health (NIH) began requiring all published articles arising from research it funds to be placed in the publicly accessible repository PubMed Central[7] within 12 months of publication. In January of 2011, President Obama signed the America COMPETES Reauthorization Act of 2010.[8] This bill included two key sections that step toward the broad implementation of Open Access mandates for scientific research. The Act both

[3] Association of American Publishers Press Release: http://www.publishers.org/press/56/

[4] See: http://blogs.plos.org/plos/2011/11/plos-open-access-collection-%E2%80%93-resources-to-educate-and-advocate/ for a collection of articles on Open Access.

[5] See http://www.plos.org/publish/pricing-policy/publication-fees/ for pricing information.

[6] See http://scholarlykitchen.sspnet.org/2011/06/28/plos-ones-2010-impact-factor/ for recent impact factor information.

[7] PubMed Central: http://www.ncbi.nlm.nih.gov/pmc/

[8] America COMPETES Reauthorization Act of 2010: http://www.gpo.gov/fdsys/pkg/BILLS-111hr5116enr/html/BILLS-111hr5116enr.htm

required the establishment of an Interagency Public Access Committee to coordinate dissemination of peer-reviewed scholarly publications from research supported by Federal science agencies, and it directed the Office of Science and Technology Policy in the Whitehouse to develop policies facilitating online access to unclassified Federal scientific collections. As a result, on November 3, 2011 the Whitehouse announced two public requests for information on, "Public Access to Peer-Reviewed Scholarly Publications Resulting From Federally Funded Research" and "Public Access to Digital Data Resulting From Federally Funded Scientific Research," As this chapter goes to press, the Office of Science and Technology Policy at the Whitehouse is gathering plans to enable Open Access to publications and to data from federal funding agencies.[9]

These events indicate increasing support for the public availability of scientific publications on both the part of regulators and the scientists who create the content.[10] The paradoxical publishing situation of sustained high charges for content generated (and subsidized) for the public good came about in part through the scientific norm of transparency. As mentioned earlier, establishing a scientific fact is difficult, error-prone work. The researcher must convince skeptics that he or she has done everything possible to root out error, and as such expose their methods to community scrutiny in order to flush out any possible mistakes. Scientific publication is not an exercise in *informing* others of new findings, it is an active dialog designed to identify errors and maximize the integrity of the knowledge. Scientific findings and their methodologies that are communicated as widely as possible have the best chance of minimizing error.

Scientific knowledge could be spread more widely, more mistakes caught, and the rate of scientific progress improved. Scientists should be able to share their published articles freely, rather than remitting ownership to publishers. Many journals have a second copyright agreement that permits the journals to publish the article, but leaves copyright in the hands of the authors.[11] We are in need of a streamlined and uniform way of managing copyright over scientific publications, and also copyright on data and code, as elaborated in the next section.

---

[9] See http://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research

[10] Unsurprisingly, the journal publishers are not so supportive. Just before the 2011 winter recess, House representatives Issa and Maloney introduced a bill that would do enormous harm to the availability of scientific knowledge and to scientific progress itself. Although no longer being considered by Congress (support was dropped the same day that publishing giant Reed-Elsevier claimed it no longer supported the bill), the "Research Works Act" would have prohibited federal agencies and the courts from using their regulatory powers to make scientific articles arising from federally funded research publicly available.

[11] See for example Science Magazine's alternative license at http://www.sciencemag.org/site/feature/contribinfo/prep/lic_info.pdf (last accessed January 29, 2013).

Research Dissemination: Data and "Raw Facts"

Computational science today is facing a credibility crisis: without access to the data and computer code that underlies scientific discoveries, published findings are all but impossible to verify (Donoho et al. 2009). This chapter discusses how Intellectual Property Law applies to data in the context of communicating scientific research. Drawing on our vignette introduced in the beginning of "Research Dissemination: The Narrative," data is understood as integral in the communication of scientific findings. Data can refer to an input into scientific analysis, such as a publicly available dataset like those at Data.gov[12] or one gathered by researchers in the course of the research, or it can refer to the output of computational research. In short, it is typically inference, and array of numbers or descriptions, to which analysis interpretation is applied.

In 2004, Gentleman and Temple Lang (Gentleman and Temple Lang 2004). introduced the concept of the *compendium*: a novel way of disseminating research results that expands the notion of the scientific publication to include the data and software tools required to reproduce the findings. At core, the research compendium envisions computational results not in isolation, but as components in a description of a meaningful scientific discovery.

Reproducible computational science has attracted attention since Stanford Professor Jon Claerbout wrote some of the first really reproducible manuscripts in 1992.[13] Since then a number of researchers have adopted reproducible methods (Donoho and Buckheit 1995; Donoho et al. 2007; Stodden et al. 2012) or introduced them in their role as journal editors[14] (Trivers 2012). Mature responses to the ubiquity of error in research have evolved for both branches of the scientific method: the deductive branch relies on formal logic and mathematical proof while the empirical branch has standards of statistical hypothesis testing and standardized communication of reproducibility information in the methods section. Unifying the scholarly record to include digital objects such as code and data with the published article facilitates the new types of information flows necessary to establish verifiability and reproducibility in computational science.

As we saw in the previous section, copyright attaches to the original expression of ideas and not to the ideas themselves. In the case of data, U.S. copyright does not attach to raw facts.[15] In 1991 the U.S. held that raw facts are not copyrightable although the original "selection and arrangement" of these raw facts may be.[16] The Supreme Court has not ruled on Intellectual Property in data since and it

[12] See https://explore.data.gov/

[13] See http://sepwww.stanford.edu/doku.php?id=sep:research:reproducible

[14] Journal of Experimental Linguistics: http://elanguage.net/journals/jel

[15] Although copyright does attach to raw facts under European Intellectual Property Law. This is a key distinction between European and U.S. Intellectual Property systems in the context of scientific research.

[16] Feist Publications v. Rural Telephone Service Co., 499 U.S. 360 (1991).

seems plausible that in modern scientific research the original selection and arrangement of facts may create a residual copyright in a particular dataset, if there was "original selection and arrangement" of these raw facts. Collecting, cleaning, and readying data for analysis is often a significant part of scientific research.

The *Reproducible Research Standard* (Stodden 2009a, b) recommends releasing data under a Creative Commons public domain certification (CC0) in part because of the possibility of such a residual copyright existing in the dataset.[17] Public domain certification means that as the dataset author you will not exercise any rights you may have in the dataset that drive from copyright (or any other ownership rights). A public domain certification also means that as the author you are relying on downstream users' ethics, rather than legal devices, to cite and attribute your work appropriately.

Datasets may, of course, have barriers to re-use and sharing that do not stem from Intellectual Property Law, such as confidentiality of records, privacy concerns, and proprietary interests from industry or other external collaborators that may assert ownership over the data. Good practice suggests planning for data release before beginning a research collaboration, whether it might be with industrial partners who may foresee different uses for the data than really reproducible research, or with scientists subject to a different Intellectual Property framework for data, such as those in Europe (Stodden 2010, 2011).

Research Dissemination: Methods/Code/Tools

Computational results are often of a complexity that makes communicating the steps taken to arrive at a finding prohibitive in a typical scientific publication, giving a key reason for releasing the code that contains the steps and instructions that generated the published findings. Of the three digital scholarly objects discussed in this chapter, code has the most complex interactions with Intellectual Property Law as it is both subject to copyright and patent law.

Software is subject to copyright, as it is an original expression of an underlying idea. The algorithm or method that the code implements is not subject to copyright, but copyright adheres to the actual sequence of letters and numbers that is the code. Copyright prohibits others from reproducing or modifying the code—for scientific applications this would prohibit running the code on a different system (reproducing) or adapting the code to a new problem (re-using). These action are openly encouraged in scientific research and again, scientific norms are at odds with Intellectual Property Law. An open license that permits others to re-use scientific code is essential.

[17] Creative Commons was founded in 2001 by Larry Lessig, Hal Abelson, and Eric Eldred to give creators of digital artistic works the ability to set terms of use on their creation that differ that those arising from copyright. Creative Commons provides a set of licenses with terms of use for work that differ from, and are usually more permissive than, the default copyright.

The Creative Commons licenses discussed in the previous section were created for digital artistic works and they are not suitable for code. There are, however, a great number of open licenses written for software. Each of these licenses sets some specific terms of use for the software (none of them rescind the underlying copyright). Software can exist in two forms, source and compiled, and for modification transmission of the compiled form alone is not sufficient. In the context of scientific research, source code is often in the form of scripts, python or R for example, that execute in association with an installed package and are not compiled. Communication of the source code, whether intended to be compiled or not, is essential to understanding and re-using scientific code.

There are several open licenses for code that place few restrictions on use beyond attribution, creating the closest Intellectual Property framework to conventional scientific norms. 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 any modified downstream software. The license 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 followed by a disclaimer releasing the author from liability for use of the code. The above copyright notice and list of conditions, including 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. The Apache 2.0 license is also commonly used to specify terms of use on software. Like the Modified BSD and MIT licenses, the Apache license requires attribution. It differs from the previously discussed licenses in that it permits users the exercise of patent rights that would otherwise only extend to the original author, so that a patent license is granted for any patents needed for use of the code. The license further stipulates that the right to use the software 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 any modified code carries a copy of the license, with notice of any modified files and all copyright, trademark, and patent notices that pertain to the work must be included. Attribution can also be done in the notice file.

The *Reproducible Research Standard* (Stodden 2009a, b) recommends using one of these three licenses, Modified BSD, MIT, or Apache, for scripts and software released as part of a scientific research compendium, or a similar open license whose only restriction on reuse is attribution.

Patents are a second form of intellectual property that can be a barrier to the open sharing of scientific codes. For example, as noted of the University of British Columbia's website,

> Members of faculty or staff, students and anyone connected with the University are encouraged to discuss and publish the results of research as soon and as fully as may be reasonable and possible. However, publication of the details of an invention may make it impossible to seek patent protection.[18]

Publication is, of course, the primary way research findings are made available, and authors who seek patents may be less likely to openly release their software, as software is a patentable entity (Stodden 2010, 2011). As university technology transfer offices often encourage startups based about patentable technology and software, the incentive to release code that permits others to replicate published findings is reduced. These two systems, technology transfer through patents and scientific integrity through openly available software, can co-exist. A dual-licensing system, for example, can be introduced that enables patent revenues for commercial downstream use, while permitting Open Access for research use such as verification of findings and re-use of code for research application (Stodden and Reich 2011).

It should be made clear that the code and scripts alone are not generally sufficient to ensure reproducible research, nor to understand the scientific findings in question. The accompanying narrative, documentation, and meta-data are an essential part of understanding the research findings and for their verification and replication.

Conclusion

The current set of scientific norms evolved to maximize the integrity of our stock of scientific knowledge. Hence they espouse independent verification and transparency, and historically this has been part of the rationale for the publication of research findings. The complexity of modern computational science means that in order to make reproducibility possible new types of scholarly objects, data and code, must be communicated. In this chapter I have traced how Intellectual Property Law creates barriers to scholarly communication, through both the copyright and patent systems and suggested solutions and workarounds.

[18] University of British Columbia Policy on Patents and Licensing, March 1993, http://www.universitycounsel.ubc.ca/files/2010/08/policy88.pdf

For broad reuse, sharing, and archiving of code to be a possibility, it is important that open licenses be used that minimize encumbrances to access and reuse, such as attribution only licenses like the MIT license or the Modified BSD license. A collection of code with an attribution only licensing structure, or public domain certification, permits archiving, persistence of the code, and research on the code base itself. Similarly for collections of research articles. The current system of distributed permission-based ownership makes archiving, research extensions, and scholarly research on publications next to impossible. For these reasons, as well as the integrity of our body of scholarly knowledge, it is imperative to address the barriers created by current Intellectual Property Law in such a way that access, reuse, and future research are promoted and preserved.

References

Donoho, D., & Buckheit, J. (1995). *WaveLab and reproducible research*. Stanford Department of Statistics Technical Report.

Donoho, D., Stodden, V., & Tsaig, Y. (2007). About sparseLab. Available at: http://www.stanford.edu/~vcs/papers/AboutSparseLab.pdf.

Donoho, D. L., et al. (2009). Reproducible research in computational harmonic analysis. *Computing in Science & Engineering, 11*(1), 8–18. doi:10.1109/MCSE.2009.15.

Gentleman, R., & Temple Lang, D. (2004). Statistical analyses and reproducible research. Available at: http://biostats.bepress.com/bioconductor/paper2/.

King, G. (1995). Replication, replication. *Political Science and Politics, 28*, 443–499.

Merton, R. K. (1973). The normative structure of science. In R. K. Merton (Ed.), *The sociology of science: Theoretical and empirical investigations*. Chicago: University of Chicago Press.

Stodden, V. (2009a). *Enabling reproducible research: Licensing for scientific innovation* (pp. 1–25). Law and Policy: International Journal of Communications.

Stodden, V. (2009b). The legal framework for reproducible research in the sciences: Licensing and copyright. *IEEE Computing in Science and Engineering, 11*(1), 35–40.

Stodden, V. (2010). The scientific method in practice: Reproducibility in the computational sciences. MIT Sloan Research Paper No. 4773-10. Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1550193.

Stodden, V., et al. (2011). Rules for Growth: Promoting Innovation and Growth Through Legal Reform. Yale Law and Economics Research Paper No. 426, Stanford Law and Economics Olin Working Paper No. 410, UC Berkeley Public Law Research Paper No. 1757982. Available at SSRN: http://ssrn.com/abstract=1757982 or http://dx.doi.org/10.2139/ssrn.1757982..

Stodden, V., & Reich, I. (2011). Software patents as a barrier to scientific transparency: An unexpected consequence of bayh-dole. *SSRN Working Paper*. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2149717.

Stodden, V., Hurlin, C., & Perignon, C. (2012). RunMyCode.Org: A novel dissemination and collaboration platform for executing published computational results. *SSRN Electronic Journal*. Available at: http://www.ssrn.com/abstract=2147710.

Trivers, R. (2012). Fraud, disclosure, and degrees of freedom in science. *Psychology Today*. Available at: http://www.psychologytoday.com/blog/the-folly-fools/201205/fraud-disclosure-and-degrees-freedom-in-science.