Victoria Stodden
Department of Statistics
Columbia University
1255 Amsterdam Ave, 10th fl.
New York, NY 10027

October 15, 2010

Committee on the Impact of Copyright Policy on Innovation in the Digital Era
National Academies
500 Fifth Street, NW
Washington, DC 20001

Dear Committee members,

Thank you for the opportunity to address this committee at the National Academy of Science. You are uniquely positioned to contend with the barriers to innovation that arise through the impact of copyright law on scientific integrity. In my remarks I hope to convince you of the urgent need for the Committee to redress these barriers directly by recommending open licensing for scientific works, in particular code and data. Copyright law works counter to scientific progress, with enormous impact on innovation both inside and outside the scientific enterprise.

It is widely recognized that scientific discovery is undergoing a deep and pervasive transformation. Over the past two decades, computation has become indispensable to scientific research, and will eventually emerge as central to the scientific method. Our stock of scientific knowledge is now accumulating in digital form. For example, our DNA is encoded as genome sequence data, scans of brain activity exist in image datasets, and records of our climate are stored in myriad time series datasets. Equally as importantly, our reasoning about these data are recorded in software; in the scripts and code that analyze the digitally recorded world. With the parallel development of the Internet as a pervasive digital communication mechanism, an unprecedented opportunity for access to society's scientific understanding is at hand. In this digitized world copyright law acts counter to scientific integrity and stands as a barrier to access and innovation. Let me explain.

Scientific research today has changed, in such a way that bring it in direct conflict with copyright law.

In a modern scientific research experiment, data are typically gathered and stored electronically as a database on a computer, perhaps a laptop or server. The data are typically scrutinized by the researchers for mistakes, labeling inconsistencies, or other errors, and then corrections are made. After the data have been cleaned and prepared, statistical modeling decisions are made and analysis is carried out. This entire process is geared to produce new scientific results. These are written up in a paper which typically follows the style established in the pre-digital age - usually containing only a short methods section. Compressing all the digital manipulation undertaken in the course of the research into this traditional format is simply impossible, thereby rendering the scientific findings essentially non-reproducible from the paper alone.

Now imagine a scientist developing an algorithm for the analysis of a novel type of data, perhaps generated by a new medical scanning device. The algorithm might permit a more confident identification of meaningful patterns in the scans than previous methods, perhaps allowing for improved identification of disease. In this case the researcher likely implemented the idea in code and tested it on some of the new data from the medical device. A typical publication of this type of work might give the mathematics explaining the approach, a description of the algorithm, and the figures and results from testing the algorithm. Again, the publication itself simply cannot communicate sufficient detail to permit others in the field to reproduce the results.

I cannot express strongly enough how typical these stories are of the scientific enterprise in the digital age, across disciplines ranging from astronomy to physics to the social sciences to bioinformatics.

In response to the inadequacy of today's scientific publication mechanisms to facilitate the verification of published computational results, there is a growing movement across the computational science community to restore scientific integrity by re-establishing reproducibility as a cornerstone of scientific research. In short: the **full release of the code and data** that generated the published findings, such that the results can be reproduced. Funding agencies are working to facilitate code and data release: the National Science Foundation (NSF) now requires data release plans to be submitted with research grant applications, and a recent NSF Task Force on Grand Challenge Communities and Virtual Organizations Report recommen1ded reproducibility as a fundamental component of computational research. A roundtable gathered prominent computational scientists and other stakeholders at Yale Law School last November to address issues of data and code sharing. Journals are beginning to incorporate the publication of data and code alongside their traditional paper publications, and domain-specific websites and repositories are cropping up to house scientific code and data.

This poses copyright issues. Scientific code and some aspects of data are subject to copyright by default, and this is a core barrier to innovation as code and data increasingly appear online. The scientific community as a whole has a long established ethos of sharing, and traditionally rejects property rights in scientific discoveries and contributions. The intent behind these norms is to further scientific progress in three ways: *to encourage full disclosure of the knowledge to facilitate* ***both*** *the transfer of scientific innovation and the reproducibility of published findings; to encourage the re-use of research output; and to increase our public stock of scientific knowledge*. Copyright as it applies today is a barrier to all these goals. Copyright vests authors with exclusive rights that prevent copy and re-use, and copyright adheres to written articles, figures and tables, software and code, and original selection and arrangement of data – nearly all the output of the modern scientific enterprise. The intellectual property framework scientists are subject to is at odds with their longstanding norms of openness, and this has becomes an acute problem in the digital era. As our stock of scientific knowledge is increasingly in digital form, copyright becomes a key issue blocking innovation in the digital sphere.

Scientific knowledge is regarded a public good, but scientific contributions cannot be freely applied, used, re-used, or built upon as intended if they remain subject to copyright restrictions. Discoveries that could spur innovation in areas such as the commercial sphere, among non-scientists, or even by other scientists, remain broadly inaccessible. Scientists tend not to develop their discoveries into commercial products and the open availability of scientific research output would encourage this practice. Without shared data and code, scientific results are unverifiable, opaque, and the rate of innovation slowed.

To develop innovation and foster scientific progress exceptions must be made to copyright to enable the free sharing of the code and data used in scientific publication, thereby aligning the legal framework with the established normative, and innovation producing, structure. Scientists should be able to make their scientific innovations freely and openly available online as a routine part of their work as a scientist, and as part of making their work transparent and reproducible. Scientists are not lawyers and cannot be expected to untangle the myriad open licensing options for their work. I encourage the Committee to recommend open licensing for scientific works to conform with scientific norms and to facilitate the innovation intended to derive from scientific knowledge.

Open licensing, by which I mean attaching terms of use to shared scientific code and data that free it for use with attribution as the only restriction, or commits it to the public domain, is a simple solution to the barriers placed by copyright law on innovation. For example, attaching the Creative Commons attribution license to media, such as figures. I have developed a series of licensing recommendations for scientific output called the *Reproducible Research Standard*. It is important for scientists, journals, and funding agencies to be aware of and use these licensing options, perhaps as a default, and this Committee has the opportunity to make a difference – in both awareness and in practice guidelines. Such leadership is required to solve this crisis.

As public policy becomes increasingly evidence-based, copyright acts a barrier to understanding the reasoning behind scientific results, to independent verification of the science, and to an open discussion of scientific issues. Without open sharing of code and data, it is difficult, even impossible, to resolve scientific debates.

Recommending attribution-only open licensing on scientific works would also be a step toward the untangling of the confusion regarding ownership at the university level. University ownership stakes vary according to the nature of the research output, the traditions, and the potential for profitability if commercialized. This Committee is positioned to encourage the open release of code and data, grounded in the principle of reproducible research and the enormous potential for future innovation.

The Committee should also recommend a broad fair use exception be made at the legislative level for all scientific works, including code and data. Because of the nature of scientific knowledge as a public good, this exception should apply to all uses, commercial and non-commercial. Scientific knowledge shouldn't be subject to the barriers induced by copyright.

Thank you for your time and attention. I would be happy to provide references to empirical research and legal analysis on this issue.

Victoria Stodden

Assistant Professor
Department of Statistics
Columbia University
http://www.stanford.edu/~vcs
vcs@stanford.edu

See accompanying APPENDIX for answers to committee questions.