ENABLING REPRODUCIBLE RESEARCH: OPEN LICENSING FOR SCIENTIFIC INNOVATION

Victoria Stodden[*]

ABSTRACT

There is a gap in the current licensing and copyright structure for the growing number of scientists releasing their research publicly, particularly on the internet. Scientific research produces more 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 for scientific researchers to use for all components of their scholarship. It is intended to encourage reproducible scientific investigation, facilitate greater collaboration, and promote engagement of the larger community in scientific learning and discovery.

There is an analogy between the development of culture postulated by the Creative Commons licenses and fundamental scientific methodology: both envision advances through building on work that has come before. The Creative Commons licenses are designed to facilitate the creation of culture through the modification of existing media, whereas scientific understanding grows through the reproduction and extension of current scientific research. Providing an Open Research License in the spirit of the Creative Commons licenses serves to allay fears that prevent a scientist from publicly releasing all the scholarship by including an attribution component, as well as a provision that derivative works carry the same license. I argue using the ORL can only increase our scientific understanding, at very minimal cost.

[*] Research Fellow, Berkman Center for Internet and Society, Harvard Law School; M.L.S. Stanford Law School; Ph.D., M.S. Stanford University (statistics); M.A. University of British Columbia (economics). I am very grateful for invaluable discussion with David Donoho, Danny Hillis, Larry Lessig, and John Wilbanks. Of course, any errors are mine alone.

CONTENTS

Introduction 3

Reproducible Research is an Established and Desirable Goal 6
A. The Scientific Research Product... 6

B. What is Reproducible Research? 9

C. The Groundswell 12

The Rationale for the Open Research License: The Alignment of Incentives........ 15

A Compilation License is Required 16
A. Selection and Arrangement of Data 16

The Open Research License 18

Potential Problems in Delineating Components of the Compendium 20

The Costs and Benefits of the Open Science Research License 21

Conclusion 27

INTRODUCTION

While researchers often publish papers in academic journals describing their work and summarizing their findings, it is rare they publish the entire research product. Most of the components necessary for reproduction of the results and for building upon the research – the code and parameters used, the dataset and its acquisition system, documentation, and any meta-knowledge used in the experiment – almost always remain unpublished. This may be due to strict space limitations in journals and conference proceedings, or a lag in academic expectations behind technological changes, but the problem is serious since this practice is counter to fundamental scientific principles which ensure that any finding be reproducible before it becomes accepted as a genuine contribution to human knowledge.[1] In addition, as computational research becomes more pervasive and details of the computations remain unpublished, 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

[1] Jon Claerbout, Green Professor of Geophysics at Stanford, goes further and calls for research to be "*really* reproducible." He advocates that "[a]n article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures."

to repeat the experiment," computational research, a key to advancement of modern science, could end up undermining the scientific process and becoming "the last refuge of the scientific scoundrel."[2]

Research based on computerized analysis is an increasingly important component of a growing number of fields, including computer science, statistics, many areas of engineering and the social sciences, as well as the traditional sciences such as biology, physics, and geophysics. For example, in the June 1996 issue of the flagship Journal of the American Statistical Association nine of twenty articles were computational, and in the June 2006 issue 33 of 35 were.

There is another reason to promote reproducibility: It is often required. In 2004 National Science Foundation (NSF) grants comprised 64% of total academic research and development support, and that proportion is increasing.[3] The NSF requires data and other supporting materials for any research it funds to be made available to other researchers at no more than

http://sepwww.stanford.edu/research/redoc/ (last accessed Sep 6, 2007). See also Section X infra.

[2] R. J. LeVeque, "Wave propagation software, computational science, and reproducible research," in Proc. International Congress of Mathematicians, Madrid, Spain, 2006. See also, P. Vandewalle, G. Barrenetxea, I. Jovanovic, A. Ridol, and M. Vetterli, Experiences With Reproducible Research in Various Facets of Signal Processing Research (last accessed Sep 20, 2007). http://infoscience.epfl.ch/getfile.py?recid=97195&mode=best

[3] Rhonda Britt, "Industrial Funding of Academic R&D Continues to Decline in FY 2004," National Science Foundation InfoBrief, NSF 06-315, April 2006. Available at http://www.nsf.gov/statistics/infbrief/nsf06315/nsf06315.pdf (last accessed Oct 5, 2007).

incremental cost.[4] Publishing the complete research product will accelerate the pace of research in the field, and the benefits to the scientist are clear: open research is built upon and cited more frequently that work published in closed journals.[5]

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. I propose such a license, called the Open Research License or ORL.

Part I of this article establishes the current scientific landscape: Defining reproducible research and making clear precisely which research components are in need of protection. Part II discusses the rationale for such a license as the ORL: Why reproducible research is something we want to encourage and

[4] **38. Sharing of Findings, Data, and Other Research Products**

a. NSF expects significant findings from research and education activities it supports to be promptly submitted for publication, with authorship that accurately reflects the contributions of those involved. It expects investigators to share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work. It also encourages grantees to share software and inventions or otherwise act to make the innovations they embody widely useful and usable.

National Science Foundation (NSF) Grant General Conditions (GC-1), June 1, 2007. Available at http://www.nsf.gov/pubs/policydocs/gc1_607.pdf (last accessed Sept. 4, 2007).

[5] See e.g. Hajjem, C. and Harnad, S. "The Open Access Citation Advantage: Quality Advantage or Quality Bias?" available at http://eprints.ecs.soton.ac.uk/13328/ (last

at what expense. Part III evaluates the costs and benefits of the ORL.

REPRODUCIBLE RESEARCH IS AN ESTABLISHED AND DESIRABLE GOAL

There is a groundswell of support for reproducible research and a discussion follows about how existing regulatory bodies support and adopt this notion, following a description of the concepts of scientific research product an reproducible research.

A. The Scientific Research Product

With ever cheaper computing power and disk space, and the increasingly ease at which we collect data, many research fields are turning to computational research to advance understanding of their subject. Computational research can be as simple as standard statistical analysis of a well understood dataset, or as detailed as the testing of complex new algorithms on comprehensive and standardized testbeds. Gentleman and Lang introduced the term compendium to describe all components of the research that are necessary for others to understand and replicate the research.[6] Computational research is widely varied but these research components remain the same. They are:

accessed July 17, 2008).

[6] Robert Gentleman and D. T. Lang, "Statistical Analyses and Reproducible Research," Bioconductor Project Working Papers, paper 2, 2004. Available at http://www.bepress.com/bioconductor/paper2 (last accessed Oct 5, 2007).

a. The Research Paper.

a.1) If included in a compiled format, such as pdf, then include the source files (TeX, Word, or WordPerfect files for example).

b. The Data:

b.1) The data itself.

b.2) Documentation completely describing the data: Sources, components, and possibly interpretation.

b.3) A description of how the data was brought into the form used in the research.

b.4) The code and instructions used to bring the data into the form used in the research.

b.5) Documentation of any code used in this process.

c. The Experiment:

c.1) The code and instructions used in the experiment, including all source code.

c.2) Documentation of any code used, including pseudocode.

c.3) A clear listing of the parameters, settings, and conditions under which the code was used to achieve the results described in the paper.

c.4) A clear description of the experimental methodology.

d. Results of the Experiment:

d.1) Any figures, data, or the like produced by the code from the experiment. These 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.

d.2) Documentation and explanation of the experimental results.

e. Auxiliary material:

e.1) Code used for presentation on the web or an interface to the data or results.

e.2) Documentation of auxiliary code.

Typically the compiled paper alone is all that is released. This makes it unnecessarily difficult for other researchers to fully reproduce and understand the published results, and thus build on scientific discoveries.

Releasing the data itself is vital to scientific progress but is typically not useful without a clear understanding of how the dataset was built and what methodologies were employed in the construction of the dataset (ie. points **2.2 – 2.5** above). I will label these components <u>meta-data</u>: 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. Although the data itself is not copyrightable, the meta-data and the selection and arrangement of the data are,[7] and I argue its protection is vital for the success of reproducible research.

B. What is Reproducible Research?

Jon Claerbout, a Stanford geophysics professor, advocates that "[a]n article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures."[8] This encapsulates the idea of reproducible research: the notion that independent people will be able to reproduce the results claimed, given sufficient computer resources. It does not assume access to the infrastructure that created the data, for example, but does assume access to the data that was analyzed.

There are many reasons reproducible research is desirable. It

[7] See Section X.

[8] http://sepwww.stanford.edu/research/redoc/ (last accessed Sep 6, 2007). See also, Jonathan Buckheit and D. Donoho, "WaveLab and Reproducible Research," 1995. Available at http://www-stat.stanford.edu/~donoho/Reports/1995/wavelab.pdf (last accessed Oct. 6, 2007).

supports fundamental scientific principles, which provide that any finding be reproducible before it becomes accepted as a genuine contribution to human knowledge. Indeed, this is what qualifies work as science. Intuitively, we expect that researchers who provide reproducibility should have more impact than those who don't, although the community of scientists who engage in fully reproducible research is very small, one study has shown that papers available online are cited at three times the rate those not available online.[9]

Knowing your work will be fully open to inspection in the future creates an incentive for researchers to do better, more careful, science now. For example, it will prevent any temptation, even unconscious, to substitute the more impressive looking figures into the paper that may be a slight mismatch with the accompanying methodological description.[10] Scientists operating under the principle of reproducible research will be able to reproduce their own work as they go and ensure the accuracy of the descriptions of their work. This might even have the effect of preserving

[9] S. Lawrence, "Free online availability substantially increases a paper's impact," *Nature*, vol. 411, no. 6837, pp. 521, 2001, http://www.nature.com/nature/debates/e-access/Articles/lawrence.html.

[10] See e.g. J. Young, "Journals Find Fakery in Many Images Submitted to Support Research" The Chronicle of Higher Education, May 29, 2008. Available at http://chronicle.com/free/2008/05/3028n.htm (last accessed July 18, 2008).

valuable work. One researcher tells the story of losing unreproducible figures before publication and, because of time constraints and expense, being forced to abandon publication of compelling results.[11]

Generation of results often requires a detailed knowledge of parameters and software invocation sequences. Without a clear description it can be next to impossible, even for the original scientist, to try the published methodology in a new setting or on a new dataset. Every publishing author hopes his or her new method will be of broader use than just that single published paper, and reproducible research helps ensure that possibility.

Building on research becomes very difficult without a full understanding of what has been done previously. Reproducible research makes it possible for researchers to communicate to others the difficulties they might be having in their work and for others to help and contribute to solutions.

By making the entire research compendium publicly available, scientists not in the immediate field of research can download, modify, and apply the work, thereby facilitating interdisciplinary research and collaboration. This

[11] Buckheit and Donoho, "WaveLab and Reproducible Research," at 2.

access to complete information may satisfy a basic need, or even "spiritual desire," among independent scientists to understand scientific regions "as a whole, and to lend one another strength of that understanding."[12]

C. The Groundswell

The Internet is becoming the dominant way for researchers to communicate and publicize their research, and in light of the increasing pervasiveness of Internet publishing, the standards for scientific research are changing.

Demands for openness of data and research are growing. In June 2007, the OECD announced the Istanbul Declaration, calling for governments to make their data freely available online as a "public good." There is now an archive site for scientific research papers.[13] The Open Archives Initiative and Science Commons are proposing universal standards for data repositories to facilitate reproducibility and novel scientific research.[14] Companies such as Metaweb and Google are creating new web structures specifically to unify the housing of complex data.[15] There are some research labs that carry out reproducible research as a policy and this number is

[12] Norbert Weiner, Cybernetics, at 8.

[13] http://www.arxiv.org/ Open access to 439,703 e-prints in Physics, Mathematics, Computer Science, Quantitative Biology and Statistics. But this is just the papers (including source files).

[14] http://www.openarchives.org/

growing.[16] Similarly a growing number of papers have been published recently calling for reproducible research.[17] In July of 2007, Microsoft held a Research Faculty Summit discussing reproducible research.[18] If passed, the Federal Research Public Access Act will require that 11 U.S. government agencies with annual extramural research expenditures over $100 million make manuscripts of journal articles stemming from research funded by that agency publicly available via the Internet. [Add Harvard and Stanford open Initiatives]

On September 20, 2007, the NSF released a major new initiative on Cyber-enabled Discovery and Innovation (CDI).[19] The initiative is meant to foster American competitiveness through research contributing to "a new generation of computationally based discovery concepts and tools to deal with complex, data-rich, and interacting systems." The goals the NSF states encourage all of: Data mining of large sets; Interacting complex systems; High-performance computational experimentation; Virtual environments;

[15] See http://www.freebase.com/ and http://www.google.com/base (both last accessed Sep 23, 2007).

[16] Although it is still very small: http://sepwww.stanford.edu/, the Donoho group (http://www-stat.stanford.edu/~donoho), http://lcavwww.epfl.ch/ for a few examples.

[17] See Gentleman, R., & Lang, D. T. Statistical analyses and reproducible research. Bioconductor Project Working Papers, May 2004; and Giovanni Baiocchi, Reproducible research in computational economics: guidelines, integrated approaches, and open source software, Computational Economics, Volume 30, Issue 1, August 2007.

[18] http://research.microsoft.com/workshops/FS2007/agenda_mon.aspx (last accessed Sep 23, 2007).

[19] See http://mathinstitutes.org/cdi/

and Educating researchers and students in computational discovery.

The National Institutes for Health have mandated that research it funds becomes “available in a timely fashion to other scientists, health care providers, students, teachers, and the many millions of Americans searching the web to obtain credible health-related information.”[20] The NIH envisions a searchable database of NIH funded publications.

Paul Huber has been advancing open access to research articles and their preprints, free of copyright and licensing restrictions.[21] He advocates the explicit use of Creative Commons licenses for the research papers or a similar licensing structure that allows the copyright holder to “consent in advance to let users "copy, use, distribute, transmit and display the work publicly and to make and distribute derivative works, in any digital medium for any responsible purpose, subject to proper attribution of authorship...."”[22] The Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities has been signed by 242 organizations including universities and advocacy groups such as the Open Society Institute.[23]

Science Commons suggests that “the legal questions – how can an

[20] http://grants1.nih.gov/grants/guide/notice-files/NOT-OD-04-064.html (last accessed Oct 21, 2007).

[21] http://www.earlham.edu/%7Epeters/fos/overview.htm (last accessed Oct 21, 2007).

[22] The Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities, Oct 20-22, 2003. http://oa.mpg.de/openaccess-berlin/berlindeclaration.html (last accessed Oct 21, 2007).

author make her work available to the public, while taking comfort that she retains some rights to it - have yet to be answered."[24]

THE RATIONALE FOR THE OPEN RESEARCH LICENSE: THE ALIGNMENT OF INCENTIVES

Open standards and open access are insufficient to promote the free discovery and development of science since the success of a scientist is measured by citations and the amount of subsequent work he or she engenders. This reward system can create short-sighted incentives to both move quickly to working on the next scientific publication and not release the full research compendium in the belief that other scientists will "steal" work by building upon it without attribution. I suggest an attribution license is required that will perpetuate virally through all derivative works, thereby ensuring attribution for all parts of the research compendium. Secondly, scientists need to have a guide to make the release of their complete research product as easy and as useful to others as possible. An appropriate license will do this both by making it possible for researchers to release everything under one umbrella license and publicizing the concept of doing so. A tailored license would bring the discussion beyond mere open source to

[23] http://oa.mpg.de/openaccess-berlin/signatories.html (last accessed Oct 21, 2007).
[24] Id.

Richard Stallman's concept of free software and free research.[25] Thirdly, the current copyright system closes scientific research in such a way that is counter to the scientific ethos of reproducibility.

A COMPILATION LICENSE IS REQUIRED

Copyright law in the U.S. does not permit the copyright of "raw facts" but original products derived from those facts can be and are, in fact, assigned automatically whenever a creative work is produced. In this automatic assignment, comes the prevention of copying and using the work in another creative or scientific endeavor. In the case of scientific research a tension is created since the scientific ethos is to reproduce previous results and build on them to generate further scientific understanding. The default copyright can be limited if the authors take steps to limit those rights by using an alternative license for their work such as the GNU General Public License ("Copyleft") or the Creative Commons license.[26]

A. Selection and Arrangement of Data

In Feist Publications, Inc. v. Rural Telephone Service, the Court found that white pages telephone directories are not copyrightable;

[25] "Free as in speech, not free as in beer."

[26] See http://www.gnu.org/licenses/ and http://creativecommons.org/ (last accessed Oct 21, 2007).

copyrightable works must have creative originality:[27]

> . . . the copyright in a factual compilation is thin. Notwithstanding a valid copyright, a subsequent compiler remains free to use the facts contained in another's publication to aid in preparing a competing work, so long as the competing work does not feature the same selection and arrangement.[28]

Currently the Court holds databases protectable.[29] 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 copyrightability. This permits the application of a license to foster reproducible research.

Most computational research work takes place in a university setting and many universities claim some ownership rights over the research product. In a November 1, 2007 discussion with Katharine Ku, Director of the Office of Technology Licensing (OTL) at Stanford University, the concern was not in copyright and focused on primarily on patents. The OTL did not perceive any conflict between the Open Research License I am proposing and their interests as a university.

[27] Feist Publ'ns Inc. v. Rural Tel. Serv. Co., 499 U.S. 340 (1991) at 363-364.

[28] Id. at 349. See also Bitton, Miriam, "A New Outlook on the Economic Dimension of the Database Protection Debate."

[29] Bitton, Miriam, "A New Outlook on the Economic Dimension of the Database Protection Debate" at 4.

THE OPEN RESEARCH LICENSE

The Open Research License is a compilation of existing licenses: the **Creative Commons BY** attaches license to the media components of the compendium, the **BSD** license[30] to 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.

The CC BY license is designed for media: to "share your creations with others and use music, movies, images, and text online that's been marked with a Creative Commons license." If used alone, it is misapplied to the academic research compendium since it does not adequately cover code and, in fact, using the CC BY license for code is actively discouraged by Creative Commons.[31] The BSD license evolved from the development of Berkeley Unix code and is a standard license for open code. Using the BSD license alone for scientific compendia leaves the documentation, figures, final paper and other forms of scholarship, the experimental design, GUIs interfacing with the algorithms, pseudocode, and dataset build methodologies for example, without an adequate license. But all of these works could be released appropriately under the CC BY license that ensures

[30] Since the release of the "Simplified BSD License" in January 2008 the BSD license is roughly equivalent to the MIT License.

[31] "[W]e do not recommend that you apply a Creative Commons license to software code." http://wiki.creativecommons.org/FAQ (last accessed Sep 5, 2007).

consistent viral attribution for the entire compendia.

This selection of licenses allows for viral attribution and, by avoiding the Share Alike aspect common to many licenses, ensures each scientist is attributed for only the work he or she has created. If Share Alike were not excluded from this license, each the entire derivative work (or new scientific discovery) would carry the ORL license, including the upstream work's attribution. In order 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 work (although the ORL remains attached to any upstream work he or she may have incorporated). Specifically, there must be no bar to building upon previous scientific research. A corollary benefit to the ORL's relaxation of the Share Alike component is that it becomes easier for startups to employ the research as part of their technology without having all their (possibly) proprietary work come under the ORL.

As a simple umbrella license the ORL is easier to use than the alternative. Without the ORL, each time he or she releases scholarship, the scientist would have to fashion together a combination of licenses from an entire spectrum of choices. Since the ORL uses common existing licenses, there are no compatibility or interoperability issues with existing licensing schemes.

POTENTIAL PROBLEMS IN DELINEATING COMPONENTS OF THE COMPENDIUM

Making a distinction as to which components of the research compendium belong under which license might be blurry: for example algorithm descriptions and pseudocode are frequently included in computational research. Arguably, each could be considered either code (there is no requirement that code must be functional to be covered by the BSD License for example, just that it be "source") or media (pseudocode is also text that traditionally could be covered under a CC license). Finally, there is no adequate licensing structure that intentionally applies to the structures that house the data used. The data itself is not copyrightable but often a phenomenal amount of work goes into preparation of the dataset for research and there is no reason why this should not 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. These aspects could be emphasized as important and captured by the ORL. This dovetails neatly with the aspirations of Claerbout's really Reproducible Research.

THE COSTS AND BENEFITS OF THE OPEN SCIENCE RESEARCH LICENSE

The NSF goal that publicly funded research be publicly available achieves important objectives: accountability and oversight in the use of government funds; promotion of scientific knowledge through both 1) direct conveyance and 2) facilitation of the opportunity to verify and improve answers to scientific questions; and the "sunshine principle" (knowledge of future public release creates incentives for better work). A license that can protect and promote these goals by aligning the scientific researcher's interests by providing for attribution, could not only forward our scientific knowledge but dramatically improve participation by scientists in collaborative research, encourage citizen-scientists to actively engage in research, and institutionalize the web as the mode for release of scientific discovery.

Attribution of work is a cornerstone of scientific discovery and currently a tension exists for scientists between the public release of research, thereby risking loss of attribution, and limited but attributed journal publication. This can be resolved by releasing scientific research under an appropriately tailored license.[32] The ORL would encourage

[32] "The WIPO Copyright Treaty (WCT), concluded in 1996, recognizes "the need to maintain a balance between the rights of authors and the larger public interest, particularly

academic researchers to release their work completely, permitting verification of the current findings, facilitating further scientific results in the particular area of research, and preserving attribution for research work. Such a license would also have the corollary effect of producing better science: a researcher who anticipates release of all his or her work to the public is apt to do a much more careful job.[33]

The ORL will provide a mechanism for scientists to license the meta-knowledge associated with the creation and perfecting of their data. Prior to the ORL, this would not fall under any license. The ORL will also provide metadata that can be used to associate the entire research product license status in a machine-readable way as a single product, which would be inherently more difficult if different components were under different licenses.

The ORL holds the promise of encouraging better tools for research dissemination and investigation. The license will provide cultural pressure that encourages reproducible research, and perhaps encourages journals to

education, research and access to information" in updating international copyright norms to respond to challenges arising from advances in information and communications technologies, including global digital networks.1" WIPO Copyright Treaty, CRNR/DC/96 (Dec. 20, 1996) (quoting preamble). http://people.ischool.berkeley.edu/~pam/papers/Reverse%20Notice%20June%2007%20_0628_.pdf (last accessed Sep 21, 2007).

[33] This is acknowledged by Richard Stallman when he suggests that if you develop code not under a free license, you "work on it only enough to write a paper about it, and

publish papers fully compliant with the ORL and principles of reproducible research.[34]

When the entire research compendium is released to the public, this can obviate the ability of the researchers to covertly begin a commercial venture based on the research results. This concern is contrary to scientific principles and the funding mandate of the NSF in the sense that science is a public good - work licensed under the ORL can be commercially used, it just cannot be built upon secretly. As one researcher has pointed out, an advantage to open code and clarity of experimental method is publicity of the new work.[35]

Another concern is the inherent confidentiality of some data. Some data, for example personal medical records, sensitive national security data, or proprietary industry data should not be publicly released. This can be counteracted by sanitizing the data as much as possible so that any personal or sensitive information is not released. In fact the National Academy of Sciences advocates the release of as much data as possible, even if there is a

never make a version good enough to release." http://www.gnu.org/philosophy/university.html (last accessed Sep 6, 2007).

[34] See http://lcav.epfl.ch/reproducible_research/ICASSP07/Kovacevic07.pdf

[35] See http://infoscience.epfl.ch/getfile.py?recid=97195&mode=best (last accessed Sep 20, 2007)

risk terrorist organizations may use it to damage United States interests.[36] Their evaluation is that the value of the scientific output outweighs the risk of information falling into dangerous hands. The NAS also would like to promote international scientific cooperation and is concerned undue restrictions on data would hamper this process. It may also be the case that some data may require built-in security and integrity checks that must be kept confidential for the experiment to operate. This creates the corollary concern that not all the data methodology can be released. This may not be a true cost of this license since it is clear such data would not have been released in any event. The ORL may encourage innovative ways to allow some reproducibility, such as providing an online system for other researchers to choose algorithm parameters or specific sections of data and simply be returned processed results.[37]

Algorithms may rely on proprietary libraries. Hopefully these libraries will be brought under the rubric of the ORL and opened to the wider research community. If not, the ORL may discourage the use of potentially

[36] National Academies of Sciences Press Release, "To Maintain National Security, U.S. Policies Should Continue to Promote Open Exchange of Research" Oct 18, 2007. http://www8.nationalacademies.org/onpinews/newsitem.aspx?RecordID=12013 (last accessed Oct 21, 2007).

[37] Id.

fruitful proprietary libraries.[38] Use of the ORL may involve a rethinking of university copyright and patenting policies. There may be conflicting third party obligations or conflicts with previously patented work used in the current research.

The ORL may encourage a change in the valuation of scientific work away from pure research results toward algorithm modification for useful purposes.[39] For example, industrial applications may become a vital part of research on the web and non-researchers may be able to use the scientific research more readily than under traditional publication methods.

Opening scientific research to the public has the benefit of providing the opportunity will exist for anyone with a web connection to get involved, even releasing their own derivative works under the ORL. This throws open the peer review process to anyone so motivated.[40]

Since the ORL facilitates the communication of research and ensures attribution, it avoids two of the stumbling blocks to very large scale collaboration. The internet naturally suggests such collaboration and the ORL, by making entire research product coherently and consistently

[38] See http://lcav.epfl.ch/reproducible_research/ICASSP07/BarniPCB07.pdf (last accessed Sep 20, 2007)

[39] See http://lcav.epfl.ch/reproducible_research/ICASSP07/Kovacevic07.pdf (last accessed Sep 20, 2007)

[40] like Peer-to-Patent. Mention the analogy and possible expansion of the scientific

available and ensuring attribution, encourages this use of the internet's potential. The ORL may facilitate internet-based data sharing research models. Such a machine readable license will enable researchers to search for ORL licensed work more easily.

A researching scientist may have done more experimentation than is practical for a traditional research paper. Releasing the full research product allows for the reporting and attribution of more results and experimental configurations than would ordinarily be publishable.

As alluded to in the introduction, by ensuring open easy access to others' research, the ORL will stand as a bulwark against plagiarism and falsification of scientific results. If even the potential exists for peers to verify all your methodologies, the incentive to cheat is greatly reduced. For exactly the same reason that attribution is an important feature of the ORL, a scientist's reputation is his or her career and the threat of being known as scientifically dishonest is exceedingly strong.

The role of third parties will be clear and consistent under the ORL, and this may not be if scientists do not have a clear licensing structure for computational work. This is especially important as the university is a common setting for computational research, and universities nearly always

peer review process in a similar fashion as the patent review process.

claim rights to work developed using university facilities, although are often amenable to open release of software.[41] The ORL releases the compendium to the public sphere and is not incompatible with university ownership rights as discussed previously.

CONCLUSION

The Open Research License blends the viral attribution aspect of the BSD license for the code component of the research product, Creative Commons viral attribution protection for text, documentation, figures and other media, including dataset creation methodologies, to create a new license for the computational research in all fields and manifestations. The ORL ensures viral attribution for all components of the research compendium thus supporting and promoting scientific ideal of reproducibility and encouraging the extension of research results.

[41] "… if a creator/inventor wants to put her software in the public domain so that no one has any intellectual property rights in the software, or if a creator/inventor wants to make the IP freely available, Stanford will be agreeable, so long as such an action does not conflict with any existing contractual obligations and does not create a conflict-of-interest issue." January 2002 issue (PDF) of *Computing Research News*, pp. 3, 8. available at http://www.cra.org/CRN/articles/ku.html (last accessed Oct 21, 2007).