Legal Issues in Reproducible Research

Victoria Stodden
School of Information Sciences
University of Illinois at Urbana-Champaign

2nd ACM Workshop on Data, Software, and Reproducibility in Publication
New York City, NY

May 4, 2016

Querying the Scholarly Record

- Show a table of effect sizes and p-values in all phase-3 clinical trials for Melanoma published after 1994;
- Name all of the image denoising algorithms ever used to remove white noise from the famous "Barbara" image, with citations;
- List all of the classifiers applied to the famous acute lymphoblastic leukemia dataset, along with their type-1 and type-2 error rates;
- Create a unified dataset containing all published whole-genome sequences identified with mutation in the gene BRCA1;
- Randomly reassign treatment and control labels to cases in published clinical trial X and calculate effect size. Repeat many times and create a histogram of the effect sizes. Perform this for every clinical trial published in the year 2003 and list the trial name and histogram side by side.

Courtesy of Donoho and Gavish 2012

The Issues

1. Software: Intellectual property is associated with software (and all digital scholarly objects) via:

    a) Copyright (always),

    b) Patents (possibly).

2. Data: possible intellectual property, other legal restrictions:

    a) Copyright (perhaps residual),

    b) Privacy protection laws.

1. Legal Issues in Software

Intellectual property is associated with software (and all digital scholarly objects) via the Constitution and subsequent Acts:

> "*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*." (U.S. Const. art. I, §8, cl. 8)

Argument: both types of intellectual property are an imperfect fit with scholarly norms, and require affirmative action from the research community to enable re-use, verification, reproducibility, and support the acceleration of scientific discovery.

1a. Copyright

- Original expression of ideas falls under copyright by default (papers, code, figures, tables..)

- Copyright secures exclusive rights vested in the author to:

    - reproduce the work

    - prepare derivative works based upon the original

- limited time: generally life of the author +70 years

- Exceptions and Limitations: e.g. Fair Use.

1b. Patents

Patentable subject matter: “*new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof*” (35 U.S.C. §101) that is

1. *Novel*, in at least one aspect,
2. *Non-obvious*,
3. *Useful*.

USPTO Final Computer Related Examination Guidelines (1996) “A practical application of a computer-related invention is statutory subject matter. This requirement can be discerned from the variously phrased prohibitions against the patenting of abstract ideas, laws of nature or natural phenomena” (see e.g. Bilski v. Kappos, 561 U.S. 593 (2010)).

Bayh-Dole Act (1980)

- Promote the transfer of academic discoveries for commercial development, via licensing of patents (ie. Technology Transfer Offices), and harmonize federal funding agency grant intellectual property regs.

- Bayh-Dole gave federal agency grantees and contractors title to government-funded inventions and charged them with using the patent system to aid disclosure and commercialization of the inventions.

- Hence, institutions such as universities charged with utilizing the patent system for technology transfer.

2. Legal Issues in Data

- In the US raw facts are not copyrightable, but the original “selection and arrangement” of these facts is copyrightable. (Feist Publns Inc. v. Rural Tel. Serv. Co., 499 U.S. 340 (1991)).
- Copyright adheres to raw facts in Europe.
- the possibility of a residual copyright in data (attribution licensing or public domain certification).
- Law doesn’t match reality on the ground: What constitutes a “raw” fact anyway?

Privacy and Data

- HIPAA, FERPA, IRB mandates create legally binding restrictions on the sharing human subjects data (see e.g. http://www.dataprivacybook.org/)

- Potential privacy implications for industry generated data.

- Solutions: access restrictions, technological e.g. encryption, restricted querying, simulation..

Ownership: What Defines Contribution?

- Issue for producers: credit and citation.
- What is the role of peer-review?
- Repositories adding meta-data and discoverability make a contribution.
- Data repositories may be inadequate: velocity of contributions
- Future coders may contribute in part to new software, others parts may already be in the scholarly record. Attribution vs sharealike.
    - ➡ (at least) 2 aspects: legal ownership vs scholarly credit.
- Redefining plagiarism for software contributions.

Three Goals

Assertion: Software in some form underlies a preponderance of published findings today and should be subject to standards of transparency that conform to historical standards to permit:

- reproducibility, verifiability of published findings,
- knowledge transfer, re-use,
- credit.

3. Solutions

Background: Open Source Software

- Innovation: Open Licensing
    - ➡ Software with licenses that communicate alternative terms of use to code developers, rather than the copyright default.
- Hundreds of open source software licenses:
    - GNU Public License (GPL)
    - (Modified) BSD License
    - MIT License
    - Apache 2.0 License
    - ... see http://www.opensource.org/licenses/alphabetical

Solutions: Copyright

The *Reproducible Research Standard* (*RRS*) (Stodden, 2009)

- A suite of license recommendations for computational science:
- Release media components (text, figures) under CC BY,
- Release code components under Modified BSD or similar,
- Release data to public domain or attach attribution license.

➡ Remove copyright’s barrier to reproducible research and,

➡ Realign the IP framework with longstanding scientific norms.

4. Strawman Recommendations

1. Use of the Reproducible Research Standard as a default licensing strategy.

2. Relinquish patentability through open availability of software, or use dual licensing strategy (open availability for research purposes, license for industry applications).

3. Use of persistent repositories, citation standards, discovery via the publication.