Software and Reproducible Research

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

Data Management Maturity (DMM) Advisory Board Meeting
Washington, DC
May 26, 2016

Agenda

1. Recent Activity
2. Main Issues: Legal and Social
3. Strawman Recommendations

1. Recent Workshops

- February 16-17 2016. **AAAS / Arnold Foundation Workshops on Reproducibility. Workshop III: Code and Modeling**. Washington DC.

- May 4 2016. **2nd ACM Workshop on Data, Software, and Reproducibility in Publication**. New York City, NY.

- various journal requirements for availability of software, code, or scripts that support scientific findings.

AAAS Workshop Goals

- *This workshop will consider ways to make code and modeling information more readily available, and include a variety of stakeholders.*

- *The computational steps that produce scientific findings are increasingly considered a crucial part of the scholarly record, permitting transparency, reproducibility, and re-use. Important information about data preparation and model implementation, such as parameter settings or the treatment of outliers and missing values, is often expressed only in code. Such decisions can have substantial impacts on research outcomes, yet such details are rarely available with scientific findings.*

- published document

http://www.aaas.org/event/iii-arnold-workshop-modeling-and-code

AAAS Workshop Agenda

Panel 1: Integration with the scholarly record: Case Studies and Lessons Learned (Michela Taufer)

Panel 2: Interoperability standards, proprietary codes, and verification/testing (Michael Heroux)

Panel 3: Licensing and facilitating re-use (Victoria Stodden)

Panel 4: Credit and citation standards, persistence (i.e. DOIs, repositories, embargo periods) (Kate Keahey)

Panel 5: Edge cases, specialized hardware, large or exceptionally complex code bases (Lorena Barba)

Panel 6: Minimal sharing requirements, workflows (Ewa Deelman)

2.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.

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.

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.

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.

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.

3. 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.