The *Reproducible Research Standard*:

Reducing Legal Barriers to Scientific Knowledge and Innovation

Victoria Stodden

Science Commons Fellow

MIT Sloan School

vcs@stanford.edu

Communia: Global Science & Economics of Knowledge-Sharing Institutions

Torino, Italy

June 30, 2009

Agenda

1. The Scientific Method is being transformed by massive computation
   - New modes of knowledge discovery? Code, data intensive.
   - New methods of communication of digitized science - irregularly used.

2. Facilitating reproducibility: the *Reproducible Research Standard*

Transformation of Scientific Enterprise

Massive Computation: emblems of our age include:

- data mining for subtle patterns in vast databases,
- massive simulations of a physical system's complete evolution repeated numerous times, as simulation parameters vary systematically.

Raises new questions about science..

Examples of Digitized Science

- Mathematical proof via simulation, not deduction,
- Large collaborative data-intensive simulations and research:
  - High Energy Physics: DANSE, LHC,
  - Community Climate Models,
  - Astrophysics Simulations Collaboratory.
- Individual researchers with code and data.

Computation is Increasingly Pervasive

- JASA June 1996: **9** of **20** articles computational,
- JASA June 2006: **33** of **35** articles computational,

Increase to all but 2 articles.

Emerging Credibility Crisis in Computational Science

- Error control forgotten? Typical scientific communication doesn't include code, data.
- Published computational science near impossible to replicate.
- JASA June 1996: none of the 9 made code or data available,
- JASA June 2006: 3 of those 33 articles claimed to have code publicly available.

Changes in Scientific Communication

- Internet: communication of all details/data of computational research possible,
- Scientists often post papers but not their complete body of research.

- Changes coming: individual efforts, journal requirements, repositories…

Copyright vs Science

Scientific norms:

1. Copy/replicate results before accepting them as knowledge,
2. Build on these results for new discoveries,
3. Give up rights over works, with the exception of attribution.

Implies a communality of research, in exchange for citation.

Legal Barriers to Sharing

- Original expression of ideas falls under copyright by default.

- Copyright creates exclusive right of the author to:
  - reproduce the work,
  - prepare derivative works based upon the original.

Potential Solution: Really Reproducible Research

Pioneered by Jon Claerbout

"An 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."

(quote from David Donoho, "Wavelab and Reproducible Research," 1995)

Reproducibility

- *Definition*: A result is reproducible if a member of the field can independently verify the result.
- Typically, provide code and data,
- Does not imply access to proprietary software or specialized equipment or computing power.

Open Source Software Licensing

- Creative Commons open licensing for creative works,
- Code licenses:
  - MIT license: attribution
  - GNU GPL: attribution and share alike
  - Hundreds of software licenses..

Apply to Scientific Work?

- Remove copyright's block to fully reproducible research,
- Attach a license with an attribution component to *all* elements of the research compendium (including code, data), encouraging full release.

Releasing Data?

- Raw facts not copyrightable.

- Original “selection and arrangement” of these facts is copyrightable. (Feist Publ’ns Inc. v. Rural Tel. Serv. Co., 499 U.S. 340 (1991))
- Appropriate license on description of selection and arrangement.

Reproducible Research Standard

Realignment of legal rights with scientific norms:

1. Release media components (text, figures) under CC BY.
2. Release code components under MIT license or similar.
3. Attribution license on selection and arrangement.
4. Data released under CC0.

Frees the scientific work from copying and reuse restrictions, with attribution.

Benefits of RRS

- Focus becomes release of the *entire* research compendium,
- IP framework supports scientific norms,
- Hook for funders, journals, universities,
- Standardization avoids license incompatibilities,
- Clarity of rights (beyond Fair Use),
- Restoration of scientific method,
- Facilitation of research, thus citation, discovery.
- Access by those outside the Ivory Tower.

Benefit for Scientists

- Openness means increased citation.
- Working reproducibly engenders better science.
- Easier for the scientists to build on his or her own work.
- Showcase of skillset for potential collaborators/funders/employers.

But…

"Behind one door is tenure - behind the other is flipping burgers at McDonald's."

Real and Potential Wrinkles

- Reproducibility neither necessary nor sufficient to prevent mistakes,
- Attribution in digital communication:
  - Legal attribution and academic citation not isomorphic
  - Contribution tracking (RDFa?)
- RRS: Need for individual scientist to act.

- "progress depends on artificial aids becoming so familiar they are regarded as natural" I.J. Good ("How Much Science Can You Have at Your Fingertips", 1958).

Publications (2009)

- "Enabling Reproducible Research: Open Licensing for Scientific Innovation,"
- "15 Years of Reproducible Research in Computational Harmonic Analysis"
- "The Legal Framework for Reproducible Research in the Sciences: Licensing and Copyright"

http://www.stanford.edu/~vcs