Two Ideas for Open Science (forget Open Data!)

Victoria Stodden
Postdoctoral Associate in Law and
Kauffman Fellow in Law and Innovation
Yale Law School

Open Science Summit
UC Berkeley, California
July 29, 2010

Open Science as a Movement

Reproducibility as a Principle
- The Credibility Crisis in Computational Science

Knowledge Sharing in Science: Code and Data
- A day in the life of...
- Code must be open too
- Implementing Reproducibility

Conclusions

Open Science as a Movement

What comprises a movement in scientific methodology? Unified changes across fields and disciplines occurring at the same time.

We have changing communication modalities:

- computational pervasiveness across fields, and changing the nature of knowledge discovery,
- changing communication modalities.

not just technological... but cultural components:

- journal requirements for code and data release,
- funder data release plans (NSF),
- expectations of digital sharing and acknowledgment among scientists.

Thesis: adoption and adaptation not fast enough.. we have a credibility crisis in computational science.

A Credibility Crisis in Computational Science..

- Climategate,
- Potti & Nevins and the Duke clinical trials,
- Geoffrey Chang retractions 2006,
- fMRI correlation analysis 2005,
- "Editorial Expression of Concern" from *Science* in January 2010,
- more...

Solution?

To address these concerns we must ensure *reproducibility* of computational scientific results.

- Sharing of the code and data that underly published results, at the time of publication, such that a knowledgeable person can replicate the findings.

Subthesis 1: Reproducibility is a key framing issue for open science.

A day in the life of...

A not atypical computational project workflow:

1. Experimental design,
2. Data collection,
3. Data filtering, cleaning, sorting, preparation for analysis,
4. Data analysis, modeling,
5. Results, conclusions,
6. Distillation of findings into publication.

Each step embodies deep intellectual contributions to science, and often myriad decisions necessary for replication of the results.

A day in the life of...

- Data filtering can be complex and highly impactful on outcomes,
- Data analysis typically encodes statistical methodology and algorithms (often new and deep intellectual contributions to science).

Both are embedded in software and necessary for the verification of findings.

Subthesis 2: Open code is as much a part of open science as open data and must be included in the open science movement with equal prominence.

Implementing Reproducible Research

Intellectual frameworks:

- Reproducible Research Standard (Stodden, 2009),
  1. Release media components (text, figures) under CC BY,
  2. Release code components under Modified BSD or similar,
  3. Release data to public domain (CC0) or attribution license.
- Notion of *Research Compendium* (Gentleman & Lang, 2004).

Tools to assist in code and data sharing:

- Publication software: Sweave, GenePattern..
- Sharing software and platforms: mloss.org, DANSE, Madagascar..
- Workflow tracking and provenance software: Taverna, Pegasus, Trident Workbench, Galaxy, Sumatra..

Conclusion

- Open code and data is a unifying principle across all computational fields, solidifying the open science movement.
- Open code and data can be grounded as foundational to computational science through the reproducibility requirement of the scientific method.
- The Open Science Movement isn't an update to the social contract, but more fundamentally a return to the scientific method.

References:

"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,"

"The Scientific Method in Practice: Reproducibility in the Computational Sciences"

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

Data and Code Sharing Roundtable, Nov 2009:
`http://www.stanford.edu/~vcs/Conferences/RoundtableNov212009/`