Tools for Facilitating Code and Data Release in Scientific Research

Victoria Stodden
Yale Law School
Science Commons

Foo Camp
Sebastopol, CA
August 29, 2009

Agenda

Code and Data Not Revealed for Published Results

Typically, the code and data that underly published computational scientific results are not available, implying:

- inability to directly verify the result,
- difficulty in building on results.

Computational science is young, needs to develop standards for what constitutes knowledge.

Reproducibility of published results is a necessary step in that direction.

Facilitating Code and Data Release

Biggest barrier to code and data release:

- The time it takes to clean up and document.

Possible solution: Tools that help track workflow as the research progresses.

- Provenance Tools
- Work Sharing / Collaboration Tools

Beyond Version Control: Tools under Development

Provenance and Workflow tracking

- Provenance Challenges: 2009, 2007, 2006 (Simon Miles, King's College London)
- descriptions of workflow: computations, their parameters, I/O data, control dependencies between them..
- Pegasus workflow compiler/mapper, Askalon, Taverna, Galaxy
- Issue: accessibility of such systems by non-CS specialists, and for less complex research tasks.

Work Sharing / Collaboration Tools

- Repositories and file tracking systems (NSF based?)
- Google Wave
- Using provenance tools for collaborative purposes and vice versa.

Private and Public Incentives to Share

Private incentives at work:

- Greater publicity
- Reuse of own work
- Encouraging others to work on the problem
- opportunity for feedback

Societal incentives:

- Encourage scientific advancement
- Encouraging sharing in others
- Set standards for the field
- Improve the calibre of research

(Survey Results, Stodden 2009)

Tools for Tagging Work

Idea: track code and data contributions through tagging

- Search by promotion/hiring committees
- Automatic satisfaction of open license attribution requirements
- Search for tools/code/collaborators

The Reproducible Research Standard

Goals:

- realign IP rights with scientific norms,
- release of all scientific research, including code and data.

To satisfy the Reproducible Research Standard:

1. CC BY on media such as text, figures,
2. Attribution license on code: such as Apache 2.0, MIT, LGPL,
3. Data under CC0 or Science Commons Open Access Data Protocol,
4. "original selection and arrangement" of the data, under CC BY or attribution open source license.

(Stodden, 2009)