“Science” in the Computational Sciences

Victoria Stodden

MIT Free Culture

October 3, 2008

The “Third Branch” of the Scientific Method

- Scientific study is no longer only *deductive* or *empirical,* it is also *computational*:
  - Simulations
  - Massive data research, data mining
- Across many fields:
  - Genetics, Biostatistics
  - Machine Learning / Network Analysis
  - Computer Science, Statistics
  - Geophysics/Earth Sciences
  - Law …

Computational Research and the Ubiquity of Error

- Standards are established in deductive and empirical fields (proofs, error detection)
- Need standards for computational research -> reproducibility; peer review could be reformed

- Publicly funded work is not made publicly available, as often mandated
- Scientific research remains closed and difficult to access; fields fragment

Solution: Standards for Computational Research

- Need clear explanations of the process to find and eliminate errors
- Need standardized mechanisms for verifying work and validating conclusions

Definition: Research *Compendium*

(Gentleman & Lang 2004)

The entire set of materials required to reproduce the results:

- Research paper and source files
- Data and its documentation, methodology, code (meta-data)
- Code, parameters, instructions from the experiment
- Results and documentation
- Auxiliary materials

Definition: Research *Compendium*

(Gentleman & Lang 2004)

The entire set of materials required to reproduce the results:

- Research paper and source files
- Data and its documentation, methodology, code (meta-data)
- Code, parameters, instructions from the experiment
- Results and documentation
- Auxiliary materials

The Solution: Reproducible Research

- Align incentives to promote reproducible research:
  - Allay attribution concerns
  - Educate about copyright and compendium release
  - Peer review of open research, verification of results

The Reproducible Research Standard

- Ensures attribution
  - Viral attribution attached to all derivative works that use the original research product, limited to the researcher's contribution in derivative products (not Share Alike)
  - machine readable license for attribution
- Encourages release of entire compendium
  - includes meta-data but excepting the data itself (original "selection and arrangement")

RR Standard Specifications

- Media components of the compendium fall under CC-BY
- Code components fall under (new) BSD

- Data encouraged into the public domain
- Original “selection and arrangement” of the data falls under CC-BY

Why Reproducible Research?

- Computational research is becoming more pervasive
- Reproducible papers cited more frequently
- Publicity creates an incentive for better quality work (“sunshine principle”)
- Accountability and oversight
- Knowledge extends outside the immediate field to other fields, regular citizens

Why? As the Scientist:

- Incentive for the scientist to release work with attribution
- Easier to use ORL than a GPL/CC hybrid
- Publicity of the licensed compendium release concept

Why This Standard?

- GPL: designed for code
- CC: designed for media
- GPL and many CC licenses have Share Alike provisions

- Attention to reproducible research and computational standards
- Promotion and diffusion of knowledge, science
- Vehicle to tie funding to reproducibility

- More information available at:

- http://www.stodden.net
- http://www.stanford.edu/~vcs