Disclosing Data Enabled Research:
Intellectual Property Policy for Publishers, Technology Transfer Offices, and Librarians

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

Intellectual Property Scholars Conference
Stanford University

August 11, 2016

Agenda

1. Data-enabled Research is Pervasive
2. Data-enabled Research is Disruptive
3. IP for Dissemination of Novel Research Artifacts:
    i. Institutions and Libraries
    ii. Publishers
    iii. Funding Agencies
    iv. Researchers
4. Discussion: Practical Considerations and Transitioning

Data-enabled Research is Pervasive

wordhoard.northwestern.edu/userman/index.html

WORDHOARD

An application for the close reading and scholarly analysis of deeply tagged texts.

Version 1.4.4
March 1, 2011

Download and Run WordHoard (You probably want to read at least the Getting Started chapter before you download and run the program the first time.)

SDSS

Data

This is Data Release 13.

Search www.sd

Datasets | Imaging Data | Optical Spectra | APOGEE IR Spectra | MaNGA IFU Spectra | MARVELS Spectra | Algorithms

Data Volume Table

The table below lists the sizes of the various data products in DR13. Note that the total data volume is greater than 125 TB. A substantial fraction (~50%) of this is raw or intermediate data that is primarily of interest to experts. If your institution requires most or all of this data you may email us at the helpdesk to contact a data transfer expert.

There are additional, small value-added catalogs that may not be listed here, due to the timing of their release.

The Data Volume of Data Release 13

Directory	Description	Size	Dir Count	File Count
apo/logs	APO observing logs	85.3 GB	433	42,515
apo/spectro	All raw APO (BOSS) spectroscopy	4.14 TB	1,616	355,706
apo/ecam	Engineering Camera data	5.29 GB	52	9,205
apo/gcam	Guide Camera data	906 GB	1,582	3,067,981
apo/ircam	Cloud Camera data	288 GB	1,814	2,203,945
apo/mapper	Plate Mapper data	45.9 GB	1,088	55,051
apogee/spectro/data	Raw APOGEE spectroscopy	21.5 TB	2,836	177,697
apogee/spectro/data1m	Raw APOGEE spectroscopy (1-m telescope)	961 GB	80	23,178
apogee/spectro/redux/r6	APOGEE-2 spectro reductions	12.2 TB	64,633	5,668,055

YouTube

CSHL Keynote; Dr. Lior Pachter, UC Berkeley

The software contains “ideas that enable biology...”

Stories from the Supplement, 2013

The Impact of Technology

1. **Massive data**,

2. **Massive increases in compute power**,

3. **Software** as a first class research object,

4. **Communication**: research artifacts becoming *digitized* and *accessible* due to the Internet (e.g. the Open Access movement),

5. **Intellectual Property Law**: digitally shared objects bring existing and new IP considerations to the fore.

Disruption

Data-enabled research produces more than the traditional pdf: data, software, workflows, meta-data, survey instruments, …

Access to these additional research artifacts is often necessary for reproducibility,

So, the traditional pdf is insufficient as a dissemination mechanism. Dissemination becomes a *compendium* of digital artifacts necessary for verification of the findings.

The IP Implication of this Disruption

Institutions: Are these digital objects potentially patentable? Bayh-Dole provides a legal basis for research institutions to claim ownership, seek patents.

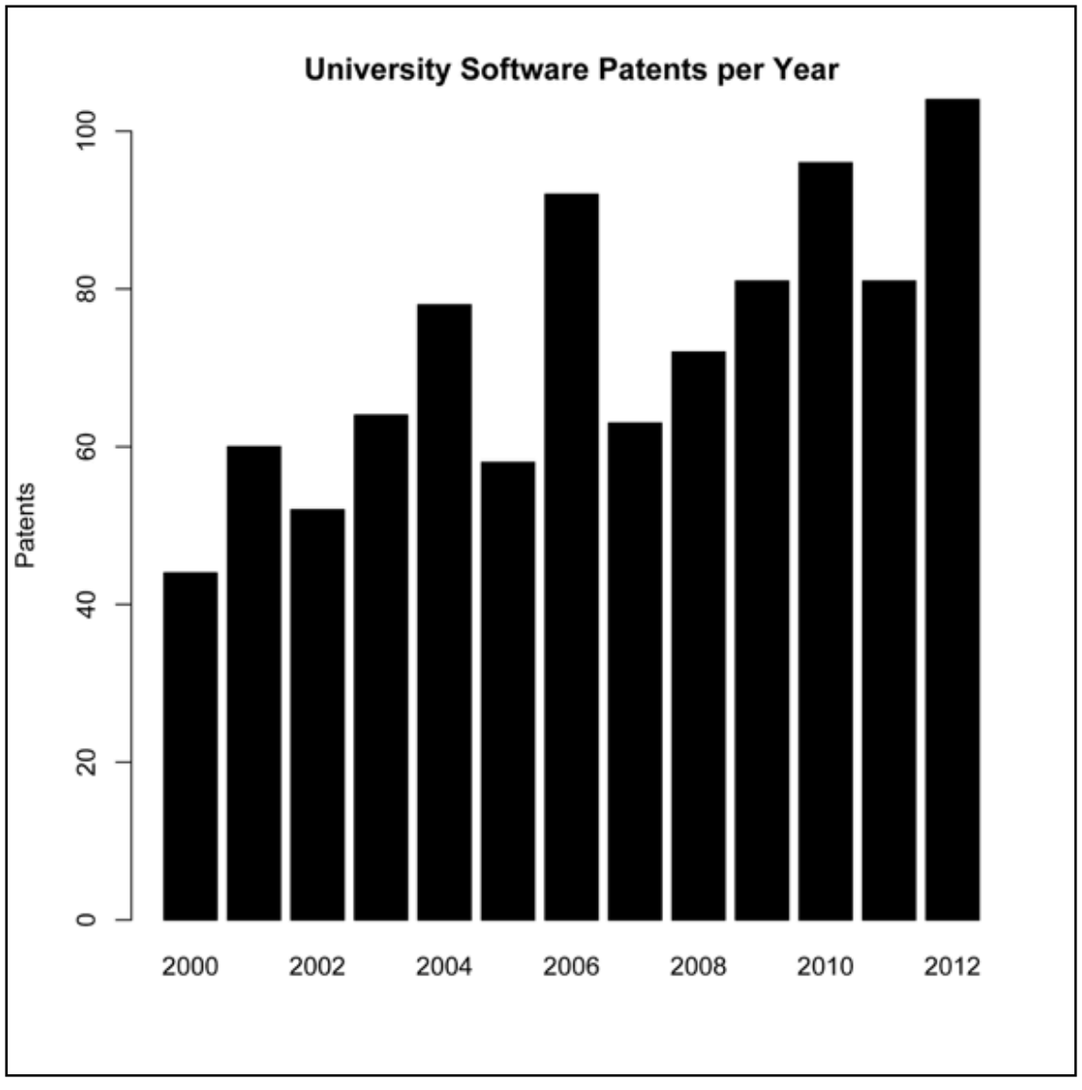
Libraries: How are these digital artifacts to be made available? Data/code repositories, access policies, linking to published findings,…

Publishers and Scientific Societies: association of artifacts with the publication, updating the traditional publication, data/code repositories and access,…

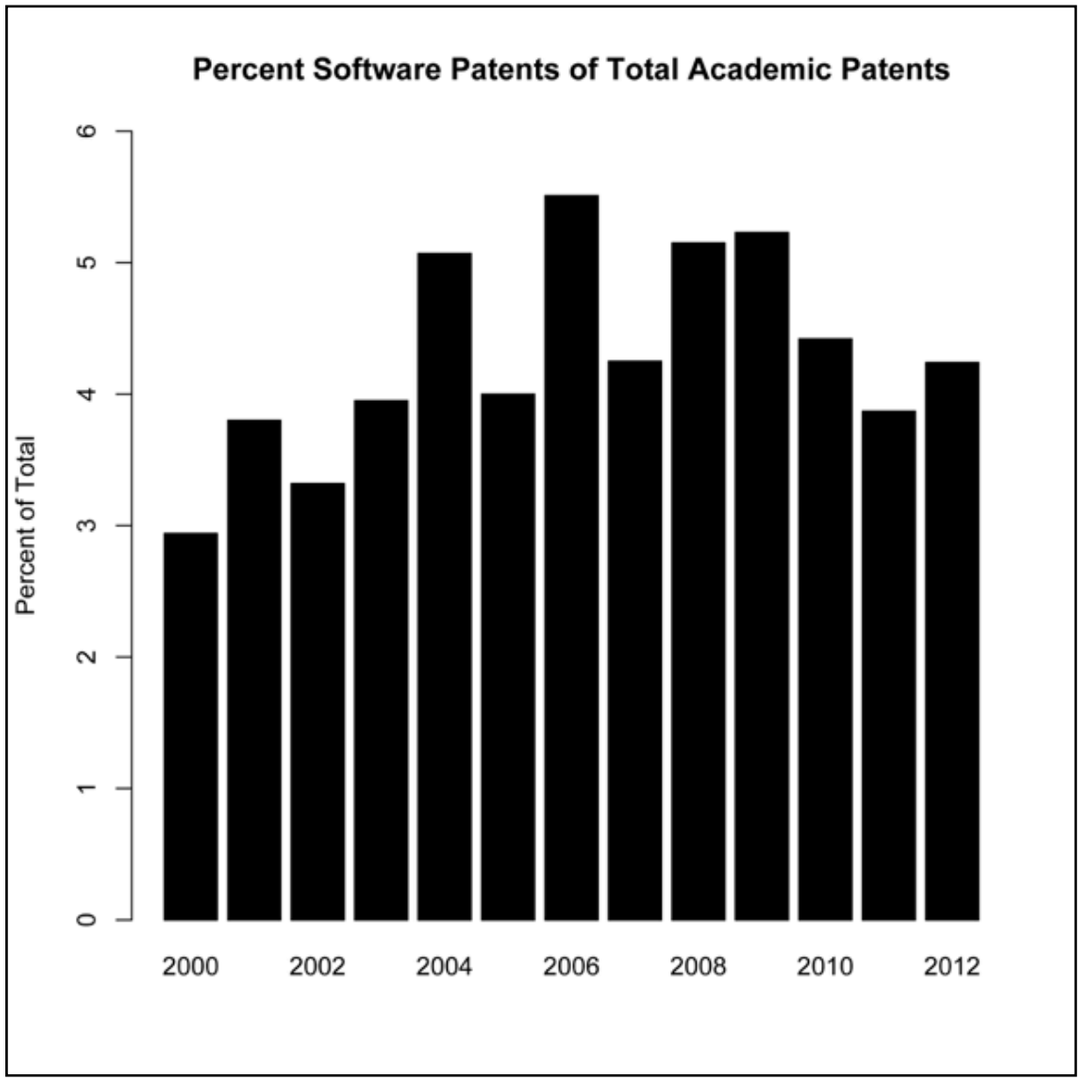
Funding Agencies: Responsibilities over artifacts produced by federally funded research? federal repositories, access policies, international and inter-agency policy and infrastructure coordination, …

Researchers: software, workflows, and perhaps data may be copyrightable. Researchers need to manage rights and enable appropriate access.

Institutions



Total Number of Software Patents filed by the top 23 University Patent Filers, 2000-2012; Classifications 341, 345, 370, 706-708, 715-716



Percent of University's Total Patent Portfolio Comprised of Software Patents, 2000-2012

Institutions

The role of the Technology Transfer Office:

- created largely as a result of the Bayh-Dole Act,
- Mission typically to provide a revenue stream to the University,
- *Proposal*: commercial vs public technology transfer.

Libraries

- Includes institutional libraries and trusted repositories,
- discovery, meta-data, interface with researchers and publishers,
- persistent link to publications,
- provenance, corrections, identifiers, re-use tracking,
- citation..

Publishers and Societies

- What does it mean to associate data and code with a publication?
- How much does it cost?
- What's our business model? Can we make money?
- Will be incur liability? How can we indemnify ourselves?
- What should our contract with the author look like when they hand artifacts off to us?
- Restricting access via tolls and gateways to data / other artifacts.

Publishers and Societies

- Science: data and code sharing since 2011.
- Nature: data sharing.
- 700+ TOP Standard signatories

See also Stodden V, Guo P, Ma Z (2013) "T*oward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals*." PLoS ONE 8(6)

Federal Funding Agencies

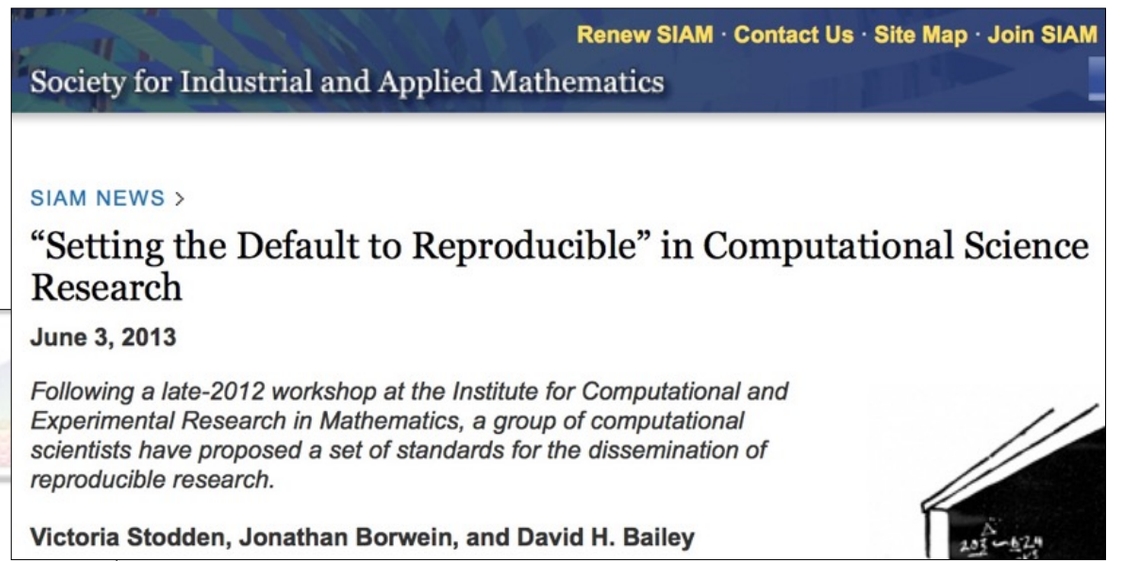
- OSTP 2013 Open Data and Open Access Executive Memorandum; Executive Order.
- "Public Access to Results of NSF-Funded Research"
- NOAA Data Management Plan, Data Sharing Plan
- NIST "Common Access Platform"
- …

Researcher Responses

Declarations and Documents:

- Yale Declaration 2009

- ICERM 2012



- XSEDE 2014

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

The Reproducible Research Standard

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.

Conclusion

- Sharing of research related digital artifacts involves addressing issues in copyright, patents and licensing, and perhaps trademark.

- IP policy cannot be implemented by a single entity but many entities are involved in the research dissemination process.

- (I assert) Policy should be driven by scientific norms, yet must be viable.

- Other considerations outside IP: privacy issues, proprietary data, legacy codes, culture and incentives, …

PTO Classification	Definition
341	Coded Data Generation or Conversion
345	Computer Graphics Processing
370	Multiplex Communications
706	Data Processing: Artificial Intelligence
707	Data Processing: Database and File Management or Data Structures
708	Electrical Computers: Arithmetic Processing and Calculating
716	Computer-aided Design and Analysis of Circuits and Semiconductor Masks
717	Data Processing: Software Development, Installation, and Management

Infrastructure Responses

Tools and software to enhance reproducibility and disseminate the scholarly record:

Dissemination Platforms

ResearchCompendia.org
MLOSS.org
Open Science Framework
IPOL
thedatahub.org
Madagascar
nanoHUB.org
RunMyCode.org

Workflow Tracking and Research Environments

Vistrails
Galaxy
Pegasus
Kepler
GenePattern
Kurator
CDE
Sumatra
Jupyter
Taverna

Embedded Publishing

Verifiable Computational Research
Collage Authoring Environment
SOLE
SHARE
knitR
Sweave