Open Bibliography

Speaker: Jim Pitman
Venue: Information Access Seminar, UC Berkeley School of Information
Friday, February 18, 2011, 3:00 pm - 5:00 pm, 107 South Hall

Abstract

The Open Bibliographic Data Working Group of the Open Knowledge Foundation has published a set of principles for open bibliographic data. These principles express a philosophy of openness for bibliographic data in support of research and knowledge enhancement:

> *For society to reap the full benefits from bibliographic endeavors, it is imperative that bibliographic data be made open — that is, available for anyone to use and re-use freely for any purpose.*

This presentation will address

- the social, technical, legal and economic issues involved in the management and dissemination of bibliographic data;
- the changes already taking place as the principles of open bibliography are being widely promoted;
- the likely nature of further developments if the principles become widely accepted.

Outline

Background

My basement:

OCCTIMES
GNEDIN - PITMAN 03-X
HERMITE
LIBRARIES
DIFFUSIONS
GNEDIN PITMAN
THE SOURCE COMPARE & SAVE
CIS
SIM & CO
MANAGEMENT

Inpirational Quotes

- The purpose of formal citation practice is to make life a little easier for your reader: any device that thwarts that purpose should be abandoned.
  - Daniel Gore, Bibliography for Beginners (1968)
- Books should only be catalogued once. Currently the public purse pays for having the same book catalogued over and over again. Librarians should act as they preach: data sets created through public funding should be made freely available to anyone interested. Open Access is natural for us, here at CERN we believe in openness and reuse… By getting academic libraries worldwide involved in this movement, it will lead to a natural atmosphere of sharing and reusing bibliographic data in a rich landscape of so-called mash-up services, where most of the actors who will be involved, both among the users and the providers, will not even be library users or librarians
- Jens Vigen, Head of the CERN Library (CERN Library publishes its book catalog as Open Data) (YouTube)
- It may be helpful to visualize my motivation as from a scientist who until recently had no interaction with mainstream library practice. The motivation springs from the fact that secondary publishers use metadata to control our actions and also charge us money for it. We live in occupied territory.
  - Peter Murray-Rust on the Open Bibliographic Mailing List, December 19, 2010. (petermr's blog)

Local Realities:

- your research proposal is an interesting one, but we cannot give you the Melvyl database records because they do not belong to us, and are not ours to give. They belong to the campuses, and we have the ability to use them for the support of system wide services, but we are not empowered to give them away. If this should change, I'd be happy to notify you and let you know about that. Some of the campuses feel very strongly that they own the intellectual property that they have invested in these records. I am sorry to disappoint you, but please know that you sparked some lively discussions around here. If you wanted a smaller set for only your own research, and you could assure us that they would not be redistributed beyond your lab, we could talk some more about that.
  - Patricia Martin and Laine Farley (CDL)
- I discussed your request with the Library Administrative Group and we all feel that there needs to be consultation with OCLC before the records are made openly available. Personally, I don't believe that we own these records. We certainly did not create the vast majority of them.
  - Bernie Hurley, Director for Library Technologies, U.C. Berkeley.

General Issues: Technical/Social

Technical:

- Architecture (Central/Distributed) (Dinosaurs/Mammals)
    - Here Comes Everybody: The Power of Organizing Without Organizations (Clay Shirky)
- Data Format/Structure (Datsets, Records, Objects, Types, BibTeX, XML, RDF, JSON, BibJSON, ...)
- Software (LAMP, L = Linux, A = Apache, M = MySQL, P = perl/php/python ..., also RoR)
    - Drupal, OpenScholar
    - NoSQL   CouchDB, MongoDB, ...
    - Google Docs, Google-Refine, Needlebase
- Navigation (Compartments/Silos)
- Layers (Library/Publisher, Homepages/ Open_Repositories)
- Data analysis and visualization (Bibliometrics, Scientometrics)
- Statistical (Deduplication, Identity uncertainty)

Social: Legal/Economic/Political

- Ownership/Control/Copright/Licensing
- Identity/Security/Privacy
- Organization and Business Model
- Software development and maintenance
- Data collection : maintenance : archiving

Fundamental Problems

- **Compartmentalization:** (silos, stovepipes)
  - Organizational structure of disciplines
  - Quality and presentation of info limited by providers
- **Navigation:** Students and scholars need guidance.
  - How to map the landscape of fields?
  - from the literature and from experts?
  - how to combine taxonomy/folksonomy?
  - how to connect researchers to literature they should know?
  - something like Google Earth to explore fields of knowledge?
- **Maintenance:** Incentive to maintain bib data reduced by free search services. Need to
  - create better maintenance tools
  - engage individuals and organizations to apply them
- **Types:** How to deal with the proliferation of types of structured documents?

Legal

U.S. Copyright Law §102. Subject matter of copyright: In general

(a) Copyright protection subsists ... in original works of authorship fixed in any tangible medium of expression, now known or later developed, from which they can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device. Works of authorship include the following categories:

(1) literary works;
(2) musical works, including any accompanying words;
(3) dramatic works, including any accompanying music;
(4) pantomimes and choreographic works;
(5) pictorial, graphic, and sculptural works;
(6) motion pictures and other audiovisual works;
(7) sound recordings;
(8) architectural works

(Note: no mention of scientific or research works)

(b) In no case does copyright protection for an original work of authorship extend to any idea, procedure, process, system, method of operation, concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such work.

U. S. Copyright Law §103. Compilations and derivative works

From §101 Definitions:

- A **derivative work** is a work based upon one or more preexisting works, such as a translation, musical arrangement, ... , abridgment, condensation, or any other form in which a work may be recast, transformed, or adapted. A work consisting of editorial revisions, annotations, elaborations, or other modifications, which, as a whole, represent an original work of authorship, is a "derivative work".
- A **collective work** is a work, such as a periodical issue, anthology, or encyclopedia, in which a number of contributions, constituting separate and independent works in themselves, are assembled into a collective whole.
- A **compilation** is a work formed by the collection and assembling of preexisting materials or of data that are selected, coordinated, or arranged in such a way that the resulting work as a whole constitutes an original work of authorship. The term "compilation" includes collective works(a) The subject matter of copyright as specified by section 102 includes compilations and derivative works ...

(a) The subject matter of copyright as specified by section 102 includes compilations and derivative works, ...

(b) The copyright in a compilation or derivative work extends only to the material contributed by the author of such work, as distinguished from the preexisting material employed in the work, and does not imply any exclusive right in the preexisting material.

Copyright in Compilations and Databases (*Feist Publications v. Rural Telephone Service Co., 6th Circuit 1996*)

In this case the U.S. Supreme Court affirmed:

- "information" is not copyrightable, but collections of information can be.
- Rural claimed a collection copyright in its telephone directory.
- The court clarified that the intent of copyright law was not to reward the efforts of persons collecting information, but rather "to promote the Progress of Science and useful Arts" (U.S. Const. 1.8.8), that is, to encourage creative expression.
- In regard to collections of facts, copyright can only apply to the creative aspects of collection: the creative choice of what data to include or exclude, the order and style in which the information is presented, etc., but not on the information itself
- The Court rejected the "sweat of the brow" doctrine (copyright protection for databases and compilations based on effort of compilation).
- Decided that compilations and databases are protected by copyright only when they are arranged and selected in an original manner
- White pages of a phone books are not protectable because the selection of the data (all customers in a geographic area) and the arrangement of the data (in alphabetical order) are not sufficiently original.

- Consequently, the competing telephone directory publisher (Feist) was allowed to extract all of the data from Rural's white pages without copyright infringement.

No copyright for facts (*Assessment Technologies v. Wiredata*, *7th Circuit, 2003*)

The 7th Circuit Court of Appeals ruled:

- that a copyright holder in a compilation of public domain data cannot use that copyright to prevent others from using the underlying public domain data, but may only restrict the specific format of the compilation, if that format is itself sufficiently creative.
- it is a fair use of a copyrighted work to reverse engineer that work in order to gain access to uncopyrightable facts.
- it is a copyright misuse and an abuse of process if one attempts to use a contract or license agreement based on one's copyright to protect uncopyrightable facts.

Implications for bibliographic metadata

Core attributes of bibliographic metadata associated with a document:

- author (string or list)
- title (string)
- when_published (date)
- how_published (string, or further structured)
    - publisher
    - editors
    - series title
    - page numbers
- links
    - urls, dois, ...

These are **facts** related to a publication event, hence **uncopyrightable** by U.S. Law. Repeating previous point:

- it is a copyright misuse and an abuse of process if one attempts to use a contract or license agreement based on one's copyright to protect uncopyrightable facts.

Encirclement of the Public Domain

Notwithstanding these cautions by U.S. Courts, publishers and Abstracting and Indexing (A&I) services typically impose copyright and licensing restrictions intended to limit the capability of users to copy or republish public domain bibliographic metadata. Some of the worst offenders:

- Web of Science (Thomson Reuters, ISI Web of Knowledge)
- Scopus (Elsevier)
- Chemical Abstracts Service
- MathSciNet

Typical license clauses (from MathSciNet License Agreement)

- Users may download search results to hard disk or diskette, provided that such data is not made available to anyone who is not an authorized user.
- Authorized users who do download and/or print search results must maintain all copyright and other notices.
- Downloading of substantial portions of the database is prohibited.
- Automated searching or downloading, by use of scripted searches, robots, spiders, crawlers, or otherwise, is also prohibited.
- However, all users may make use of free tools provided by the AMS for batch retrieval of information from the database provided such use is in compliance with copyright statements and terms of use posted from time to time on MathSciNet.

   - Conduct automated searching or downloading, by use of scripted searches, robots, spiders, crawlers, or otherwise;
   - make password-protect content of this site available to any third party, whether by telephone link, password sharing, permitting access through your computer, or by other similar or dissimilar means or arrangements.

3. You must maintain all copyright and other notices that appear on any material you download or print from this website.

Image of this attitude.

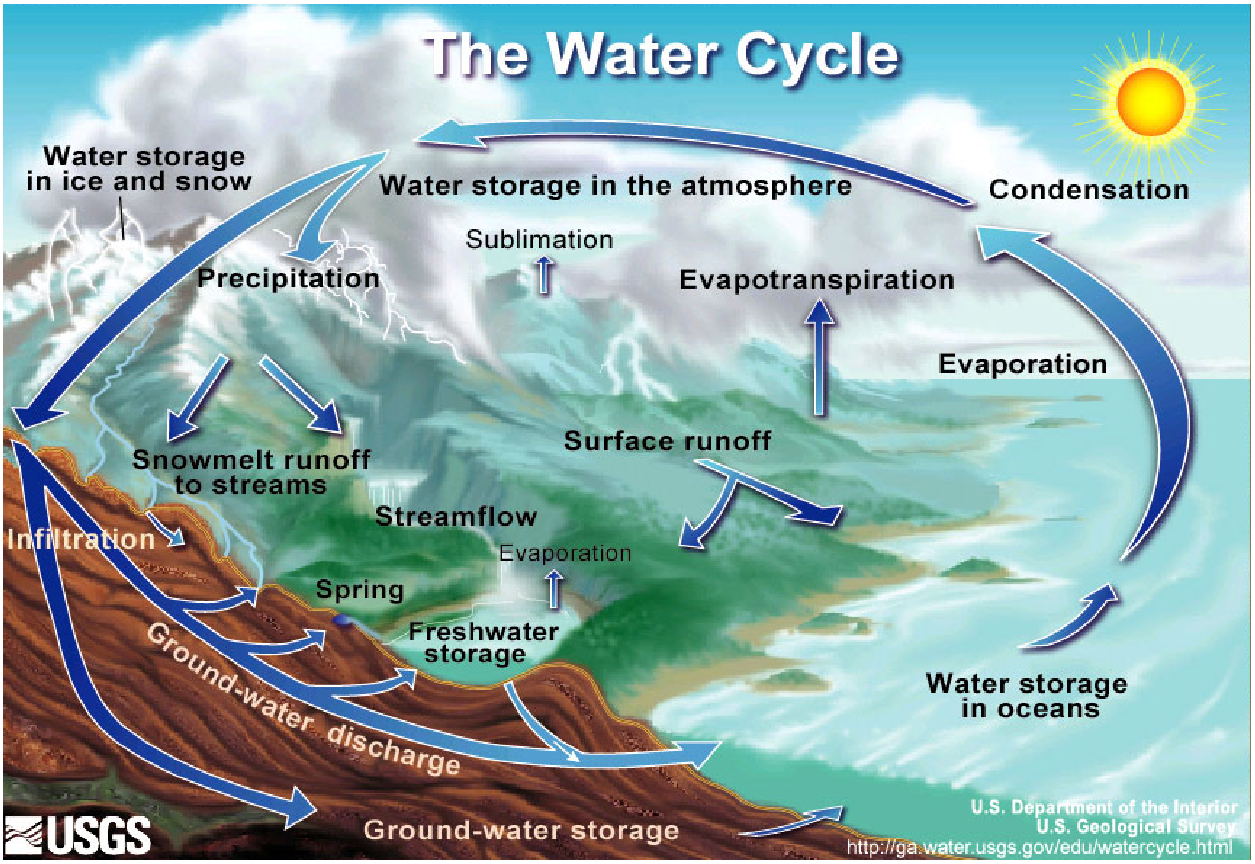
Less restrictive services

- http://arxiv.org/ Open access to 657,700 e-prints in Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance and Statistics
- http://front.math.ucdavis.edu/search?a=Pitman%2C+Jim&t=&q=&c=&n=50&s=Abstracts
- Zentralblatt MATH (ZMATH, STMA-Z) 2M + abstracts and metadata in mathematics.
    - http://www.zentralblatt-math.org/zbmath/authors/?q=Pitman,%20Jim
    - http://www.zentralblatt-math.org/zbmath/search/?q=ai:pitman.jim-w
    - http://www.zentralblatt-math.org/zbmath/authors/profile.xml?q=ai:pitman.jim-w&amp;title_=Author%20Profile
- many more
- http://upload.wikimedia.org/wikipedia/commons/9/94/Water_cycle.png

Analogies

Bibliographic data as a resource/commodity

- bibliodata subject to a cycle of use (publication/collection/abstracting/listing/citation/)
- can be commodified, put in containers, owned/controlled
- maintenance issues depend on the size of the container



Unlike water and energy:

- bibliodata not subject to conservation laws
- copying is technically trivial, inhibited only by social restrictions

- these social restrictions are unsustainable.

Strategy for Bibliographic Data Liberation

Given a raw bibliographic item, machine match it into a variety of information services:

- save copies of all records obtained
- publish a composite record back to the web under CC0 with links to all sources
- defy the sources to complain
- if they do, delete their record: they lose the link back, usually you lose nothing
- or appeal to: no copyright in facts/copyright misuse/ abuse of process

Social/Economic

Community Information Services

General problem:

- **How to organize all the data?**

Partial solution (developed with NSF sponsored Bibliographic Knowledge Network Project)

- **Empower communities of practice to take control of their bibliographic data**
- **Enable network effects in such communities by open biblio data principles**

Other components: Google, Microsoft Academic Search, ... Libraries??

Exemplars:

- **RePEc**: http://repec.org/
- **Probability Abstract Service** http://pas.imstat.org
- **Probability Web** http://probweb.berkeley.edu/
- **Departmental Services** (Berkeley, Toronto, Oxford, Sydney, ...)
- BibServer

Requirements: Adequate software, editorial commitment, and open biblio data.

Challenge to Abstracting and Indexing Services:

How to accomodate community information services?

Proposed solution: provide monthly metadata and identifier dumps at some URL, under CC0

Benefit to the community: Allows others to

- extract and develop whatever data they care about
- link back to the service for reviews, authoritative identifiers, library quality service
- mash-up the data in community-specific ways (ratings, rankings, ...)
- provide APIs and visualizations over the data
- connect to the web of Linked Data http://linkeddata.org/

Cost to service: Close to zero provided libraries continue to support it

Benefit to service: Engagement with community supported interest groups (cf. ASA, SIAM)

What if library service providers do not cooperate with community services?

- The tide of open biblio data will rise anyway (ORCID, CERN, British Library, ...)
- The value of closed sources will diminish.
- Community hostility to closed information services will rise.
- Libraries should refuse to sign license agreements which restrict scripted access or substantial downloads.
- Communities can work with Thomson WoS and Scopus which already have APIs allowing companies like Symplectic to provide university-wide bibliogaphic services.

Further discussion

Statistical Aspects

Bibliometry: Quantitative analysis of bibliographic data: selection/scoring/ranking/network stats
Citation Statistics Report: (IMS/IMU/ICIAM, 2008) [pdf]
scoring/ranking

- articles (citation counts) (Google Scholar)
- journals (impact factors) Eugene Garfield ISI 1960. [Ranked List]
- authors (h-index Jorge Hirsch 2005, ...)
- web pages (PageRank, Google)

Data Visualization, Machine Learning, Automated Classification, Collaborative Filtering (NetFlix Prize)

Bibliographic Knowledge Network

NSF Sponsored Research Project directed by Jim Pitman in collaboration with

- Brian Conrey: American Institute of Mathematics (AIM)
- Gary King: Institute for Quantitative Sciences (IQSS) at Harvard: Dataverse Network

and numerous other partners (listed later).

Goals: To create

- openly navigable network of websites
- each node a biblio guide to a specific topic or field
- each node maintained by a virtual organization
- incorporate/improve existing subject sites
- establish collective knowledge systems

Development Program

Create software and bibliographic workflows to

- select, brand, maintain, and annotate collections of structured scientific content.
- engage many small and distributed organizations in this activity
- expose bib data in machine-readable formats
- use machine learning to automate selection/cataloging/ranking
- develop statistical analysis of bib data
- establish collective knowledge systems on various scales
- promote connections between systems and disciplines

Partners

- U.C. Berkeley (Jim Pitman, M. Jordan, T. Griffiths, J. Regier)
- AIM (Brian Conrey, David Farmer)
- IQSS /Dataverse (Gary King, Micah Altman)
- IMS/CIS (Hadley Wickham, Stefano Iacus)
- ZMATH (FIZ Karlsruhe, Bernd Wegner)
- Stanford School of Ed./ Public Knowledge Project (John Willinsky)
- Metaweb / Freebase
- PlanetMath (Aaron Krowne)
- RePEc (Thomas Krichel)
- Creative Commons
- R Foundation (Stefano Iacus, Kurt Hornig, Michael Hahsler)
- Journal of Statistical Software (ASA, Jan de Lieuw)
- Springer (John Kimmel)
- CrossRef (maintainer of the DOI System)
- kReef

Ongoing projects

- BibServer
-

BibServer

(developed in collaboration with VTEX)

- Personal BibServer
- Oded Schramm
- Departmental BibServer Typical Faculty Listing
- IMS Biobibs: S.R.S. Varadhan
- IMS Fellows
- UCB Math Sci Memorial
- Portaits of Statisticians (Peter Lee, York)

MathPeople

- developed with Jaeyhun Paek (Dalhousie D-Drive) and Hadley Wickham
- supported by multiple organizations
- leverages multiple sources of name data to provide a distributed name authority system for people in the mathematical sciences.
- aggregates data about the same person from many different data sources e.g.
    - 500K mathematicians in Math. Reviews Authors Database
    - 125K mathematicians in The Mathematics Genealogy Project
    - 160K name strings in Current Index to Statistics
    - 2K mathematicians in MacTutor History of Mathematics Archive
    - hundreds of other aggregations
    - tens of thousands of homepages

Statistics Topics

- developed with Jeff Regier, supported by IMS.
- Stat Topics leverages multiple data sources to provide a comprehensive collection of topics in statistics
- provides scripted links to glossary and encyclopedia pages
- associates topics with people
- foundation for development of an open access Encyclopedia of Prob/Stat
- cf. Wikipedia, PlanetMath, Google Knol, MedPedia
- current initiative by Springer to engage editorial support from statistical societies (John Kimmel)

Summary

- Services for the managemen/analysis/delivery of bibliographic data are in rapid flux
- unique opportunities to push towards more open services
- cf. R Project, BioConductor, Dataverse Network
- potential for improvement in scholarly communication is great
- special potential for making statistical knowledge more accessible to researchers in other fields, and developing statistical anlysis of bibliographic networks.

Conclusion

What is most needed is for:

- individual researchers to make their bib data (including fulltext) available with open access
- individuals to persuade organizations of all sizes to make their aggregated bib data openly accessible
- software developers to provide data structures and workflows for large amounts of bib data
- editors and curators to improve the quality of bib data in their areas or expertize
- researchers to develop statistical analysis of bib data as a tool for advancement of knowledge
- senior statisticians to advise administrators about use of citation statistics in research assessment

Fiscal resources are also needed to attract the human ones.

Want to get involved? Please contact Jim Pitman