Best Integrated Bibliographic Services

Jim Pitman

Dept. Statistics, University of California, Berkeley

August 5, 2008

Traditional Subscription Services

Emerging Free Services

General Issues

Statistical Aspects

Some Problems

Bibliographic Knowledge Network

Development Program

Partners

Ongoing projects

BibServer

MathPeople

Statistics Topics

Traditional Subscription Services

- Current Index to Statistics
- MathSciNet
- Zentralblatt MATH (ZMATH, STMA-Z)
- Web of Science (Thomson Reuters, ISI Web of Knowledge)

Emerging Free Services

- Google Scholar
- Scirus (Elsevier)
- WorldCat (Library meta search engine)
- arXiv.org (Stat since April 2007), U.C. Davis, IOP
- CiteSeerX (Penn. State, Lee Giles)
- PubMed (NIH free archive: biomed and bio)
- Open Access Journals, e.g. Statistics Surveys
- Wikipedia,Google Knol,MedPedia, ...
- Social Bookmarking (Web 2.0): del.icio.us (Yahoo), CiteULike,Connotea (Nature)
- Semantic Web Services (Web 3.0): (API) BibSonomy, Freebase, Google Docs, Zotero (Firefox extension for browsing/organizing)

General Issues

Technical/Legal/Economic/Political/Statistical

Technical

- Architecture (Central/Distributed)
- Data Format/Structure (Objects, Types, BibTex, XML, ...)
- Software (LAMP, P = perl/php/python ..., also RoR)
- Navigation (Compartments)

Legal

- Ownership/Control/Licensing
- Privacy/Identity/Security

Economic/Political

- Organization
- Business Model
- Software Development
- Maintenance

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)

Some 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?

Bibliographic Knowledge Network

Proposal developed 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).

Goal: To create

- openly navigable network of websites
- each node a bib 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 (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 (S. Iacus, K. Hornig, M. Hahsler)
- Journal of Statistical Software (ASA, Jan de Lieuw)
- Springer (John Kimmel)
- CrossRef (maintainer of the DOI System)

Ongoing projects

- **BibServer**
- **MathPeople**
- **StatTopics**

BibServer

(maintained in part by IMS/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
- MathPeople 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.
  - 507,716 mathematicians in Math. Reviews Authors Database
  - 124,687 mathematicians in The Mathematics Genealogy Project
  - 160,000 name strings in Current Index to Statistics
  - 2,000 mathematicians in http://www-history.mcs.st-and.ac.uk/ history/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 for statisticians to push towards more open services
- cf. R Project, BioConductor, Dataverse Network
- potential for improvement in scholarly communication is very great
- special potential for making statistical knowledge more accessible to researchers in other fields

Conclusion

What is most needed is human resources:

- 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 to attract the human ones are also needed.

Want to get involved? Please get in touch!