Author Identity in the Bibliographic Knowledge Network

Jim Pitman

Dept. Statistics, University of California, Berkeley

CERN workshop on
Innovations in Scholarly Communication
June 19, 2009

Bibliographic Environment

Bibliographic Knowledge Network

General Issues

Linked Data

Author Identification

MathPeople

Conclusion

Bibliographic Environment

Traditional Subscription Services

- MathSciNet, Zentralblatt MATH
- Web of Science (Thomson Reuters)

Emerging Free Services

- Google Scholar, Scirus (Elsevier), WorldCat, OAIster,
- CiteSeer,
- PubMed, ...
- Social/Semantic Web Services (Web 2.0/3.0):
- del.icio.us
- CiteULike,Connotea,BibSonomy, Freebase, Zotero, Rexa, kReef, ...

Bibliographic Knowledge Network

- NSF Sponsored Project
- Principal Partners:
  - American Institute of Mathematics (Brian Conrey, David Farmer)
  - Berkeley Statistics and Mathematics (Jim Pitman)
  - Harvard IQSS (Gary King, Micah Altman)
  - Rice Statistics (Hadley Wickham)
  - Stanford Education (Jon Willinsky)
- Shared Goals:
  - to create and maintain an open network of locally controlled bibliographic data stores and associated services
  - to engage a large number of individuals and virtual organizations in bibliographic data enhancment
  - to create an open semantic network of machine-readable bibliographic information
  - to demonstrate the value of open bibliographic data by application of machine learning and graphical visualization tools for knowledge discovery

General Issues

Technical/Legal/Economic/Political

Technical

- Architecture (Central/Distributed)
- Data Format/Structure (Objects, Types, BibTex, XML, RDF, JSON)
- Software (Drupal, Virtuoso, CouchDB)
- Navigation (Compartments)

Legal

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

Economic/Political

- Organization
- Business Models
- Software Development
- Maintenance

Linked Data

- Strongly typed (people, bibitems, organizations, ...)
- Approaches to data enhancement: combination of
  - bottom up: software support for individuals and small orgs to curate and publish linked bib data (BibServer)
  - top down:manual/machine processing of large stores, citation parsing, ...
  - itermediate: RePEc-like aggregations using machine methods to automate tedious tasks: deduplication, classification, ...
  - use data from large and intermediate aggregations and publisher feeds to reduce the maintenance burden for individuals and small organizations
  - provide incentives to various agents to contribute to a pool of linked bib data

Author Identification

Strategy

- Do it first locally with personal bibs, easily name disambiguated
- Each local bib publisher assigns arbitrary local identifiers for its authors
- Each local bib publisher can assert identities between local authors and otherwise identified authors
- Bib data aggregators can combine identity assertions from different sources
- Machine learning methods and suitable user interfaces (e.g. Krichel's AuthorClaim) can be used to assist this process

MathPeople

- 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.
  - 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 http://www-history.mcs.st-and.ac.uk/ history/MacTutor History of Mathematics Archive
  - hundreds of other aggregations
  - tens of thousands of homepages

Conclusion

- Diverse sources of author name data are available on the web.
- Systematic local identification, publication and matching of author name lists should lead by a network effect to
  - largely reliable global identification of authors
  - increased quality of the bibliographic web of data,
  - increased incentives to individuals and organizations to maintain their bibliographic data in ways useful to other researchers.