open and closed access to author data

Thomas Krichel

LIU & НГУ

why this weirdo topic?

- Well at first I planned a talk comparing AuthorClaim as a an open access system to ORCID as a closed-access system.

- Since then there has been so much confusion about ORCID that I am not sure I can make an adequate representation of something that does not actually yet exists.

- So I prefer to talk about my work first since it is more concrete.

open access

- It means in this circumstances that is possible to get a bulk copy of the records
  - to build user services
  - to build merged datasets
- Access that is only via an API is not open access.

author identification

- It is a well-known issue that author names don't identify authors.

- Author identification can be done with the help of some records that say who has written what.

- I am not going into much more technical detail here.

implementation

- Legal detail is not important to me.
    - In applied work I don't see a problem with it.
    - Licensing issue raise the cost of open-access system by adding levels of bureaucracy.
- Technically, we can use
    - ftp
    - rysnc
    - OAI-PMH.

author claiming

- An author claiming service allows an author to claim what documents

    - they have written

    - they have not written

- This is a service type that is related to author identification.

- I have been the pioneer in setting up the first such services, the RePEc Author Service.

RePEc author service RAS

- It uses the document data from the RePEc database.

- RAS is an enabling service for other services.

- It has been a key component to the growth of RePEc.

- RePEc has grown because it is widely used in performance assessment.

some stats about RAS

- It now holds records for over 25,000 authors (meaning registrants who have claimed at least one paper).

- From an independent list of most famous economists, over 80% in the top 1000 have registered.

registration and performance

- RePEc really grows because it's data is used in performance assessment.

- The units assessed are the authors.

- The authors are assessed using performance measures based on document.

- An author is as good as all her claimed documents are.

- Authors have a keen urge to maintain their profiles current especially at ranking renovation time.

author registration and repositories

- RePEc is an aggregation of institutional repositories. Over 1300 repositories take part.

- The repositories fill because RePEc services

- Performance assessment is something that repositories will have to get into.

additional services

- RAS implements services that go beyond author claiming.

- One important service is the processing of references into citations.

- Another is the ability to claim to be working for an institution. RAS uses RePEc's large list of economics departments called EDIRC.

ACIS

- ACIS is a software system, written by Ivan V. Kurmanov, that supports the maintenance of author registration services.

- The development of ACIS was financially supported by the Open Society Institute.

- There are two running instances

  - RePEc Author Service

  - AuthorClaim

AuthorClaim

- I founded this in early 2008 to provide an interdisciplinary author registration service to support a wide variety of open access (or similar) repositories and bibliographic datasets.

- It uses a collection of institution handles called ARIW that I founded at the same time.

- In fact there is a trio of service all run by the Open Library Society, Inc.

AuthorClaim http and ftp

- http://authorclaim.org has the interface where authors can create profiles.
  - This should only be used for real records.
  - There is an test system available at http://test.authorclaim.org
- The profiles themselves can be mirrored from ftp://authorclaim.org as XML files.
- Sample profile ftp://authorclaim.org/d/u/pdu1.amf.xml

ARIW http and ftp

- ARIW contains name data, URL and identifiers for institutions.

- ARIW is an open <u>service</u>. The entire (!) system, including its maintenance scripts is available as one single tarball.

- There is also an ftp://ariw.org with the data files only.

3lib.org

- The main problem in author claiming is to get the document data.
- Good metadata about academic publications is difficult to get.
- I have to collect it for AuthorClaim.
- I use 3lib.org to redistribute it.
- I love sharing resources.

basic metadata

- One of the big advantages of a service like AuthorClaim is that it can be based an factual document data.

- For any document, only four metadata elements are required
  - Title
  - Author name expressions
  - URL for further information about the data

AuthorClaim stats

- About 100,000,000 authorships can be claimed or disclaimed. Sources (from 3lib.org) include
  - PubMed
  - DBLP
  - Driver/DMF
  - RePEc
- About 100 authors. Not bad since no advertisement has been made.

major issue

- The major issue is that to get authors to use AuthorClaim, the author records have to be widely used.

- But before there is a substantial take-up by authors of the AuthorClaim, other services will not take up the author records.

- The system is constructed as being useless in isolation.

current use of AuthorClaim

- ARIW uses AuthorClaim records to build profiles of authors affiliated with a particular institutions.

- There is an internal service to 3lib, that searches CrossRef (sigg engine) for data by registered authors. Those data will be fed back into 3lib.

- All AuthorClaim records have also been made available to the ORCID alpha.

feedback to 3lib resources

- Currently, bibliographic database have problems to cope with the growth of material.
- At the same time, there are substantial overlaps between in them, and no linkage between records.
- This can all be helped with author identification.
- For example, RePEc user service use author identification records to group versions of papers.

closed access

- "ORCID, Inc. aims to solve the author/contributor name ambiguity problem in scholarly communications by creating a central registry of unique identifiers for individual researchers and an open and transparent linking mechanism between ORCID and other current author ID schemes."
- In no scenario that I have seen (as a member of the technical committee) there has been open access to the ORCID records.

ORCID, Inc.

- ORCID Inc, has just been incorporated as a not-for-profit organization.
- The comparison with CrossRef is probably still most fitting.
  - CrossRef records are not freely available.
  - CrossRef membership is expensive.
  - CrossRef is technically simpler and probably lighter to run than ORCID.

ORCID alpha

- The alpha is a clone of the researcherID system of Thompson-Reuters, which itself is modeled after RAS.

- You need to sign a memorandum of understanding (MOU) to use it.

- There appear to be tons of MOUs when dealing with the publishing industry.

author claiming in ORCID

- Supposedly, there will be some author claiming in ORCID.

- At this time, it can only be done via sigg.

- In the future possibly through other APIs but that will be difficult to scale to more systems.

expansion

- There are ideas for
  - bulk matching of names
  - integration into publishers' workflow
  - ways for universities to bring in staff records
- This raises a massive concurrent curation of records problem.

contrast the foundation

- The idea of AuthorClaim is to build a system that does one thing, claiming and disclaiming document authorships.

- Author claiming does not provide author identification.

- It intends to export as much as it can, and as quickly as it can to other systems.

- The other systems then do their jobs.

- (it's just the way RePEc runs vs CrossRef runs).

other jobs

- Providing author records to publishers. This has been implemented in ACIS for EPrints but there is no running implementation.

- Report to metadata providers errors about the metadata that is uncovered by authors.

- Biggest job: build author ranking by combining author registration data with document usage data.

adding to workflow of others

- There are discussions between AuthorClaim and GESIS for the integration into the SOLIS data.
- SOLIS is the latest acquisition of 3lib.

building a user service

- Now since there is not a user interface constructed by others, I am thinking about building "author profile" a combination of 3lib and AuthorClaim data.

- I will talk about this some other time, possibly when I have something to show.

see you again?

- I will be at the "DINI-/Helmholtz-Workshop: Repositorien – Praxis und Vision"
- http://www.dini.de/veranstaltungen/workshops/dinihelmholtz-workshop-repositorien-praxis-und-vision/
- You can hear more from me and some my controversial work.

Thank you for your attention.

http://openlib.org/home/krichel