IR repository data in AuthorClaim

Wolfram Horstmann[1]
Thomas Krichel[2,3,4]

[1]Bielefeld University [2]Long Island University [3]Novosibirsk State University [4]Open Library Society

Repository Fringe 2011–08–03

thanks

- to the organizers
- to the BASE contributors
  - Bernd Fehling, Marek Imialek, Mathias Loesch, Renata Mitrenga, Dirk Pieper, Jochen Schirrwagen, Friedrich Summann, Sebastian Wolf

structure

- background
- 3lib
- Author claiming IRs
- IRs and BASE
- BASE and AuthorClaim

background

- Wolfram is chief information officer (scholarly Information) at Bielefeld University.
- They run the BASE search engine since 2004
- They are doing long-run work.

background

- Thomas the founder of RePEc.
- Thomas started this in the early 90s.
- Thomas is doing long-run work.

motivation

- Make (economics) papers freely available.
- Make information about the papers freely available.
- Have a self-sustaining infrastructure of this, don't rely on external sources.

RePEc

- RePEc is misunderstood as a repository.
- In fact it is a collection of 1300+ institutional (subject) repositories.
  - pre-date OAI
  - reduced business model
  - more tightly interoperable

RePEc sources of success

- There are a lot of sources of success.
- The reason can be classified
  - business case
  - technical matter
- both are linked

RePEc business case

- RePEc tries to decentralize as much as we can.
- RePEc run essentially on volunteer power.
- RePEc encourage reuse of RePEc data.

RePEc technical case

- RePEc registers authors with the RePEc Author Service (RAS).
- RePEc registers institutions (EDIRC).
- RePEc provides evaluative data for authors and institutions.

RePEc and IRs

- RePEc is not a repository.
- RePEc is a bibliographic layer over repositories.
- IRs can/will benefit from a similar bibliographic layer.

requirement for such a layer

- Not dependent on external funding.
- Freely reusable instantaneously.
- Must be there for the long-run.

a RePEc for all disciplines

- RePEc bibliographic data $\rightarrow$ 3lib
- RePEc Author Service $\rightarrow$ AuthorClaim
- EDIRC $\rightarrow$ ARIW

3lib

- 3lib is an initial attempt at building an aggregate of *freely* available bibliographic data.
- It's a project by OLS sponsored by OKFN.
- About 35 million records from the usual suspects: PubMed, OpenLibrary, DBLP, RePEc.

3lib elements

- The data elements in 3lib are very simple
  - title
  - author name expressions
  - link to item page on provider site
  - identifier
- 3lib is meant to serve AuthorClaim.

AuthorClaim

- AuthorClaim is an authorship claiming service for 3lib data.
- It lives at http://authorclaim.org.
- It uses the same software as the RePEc Author Service, called ACIS.
- It is running since early 2008.

author claiming history I

- Thomas started the first author claiming system, the RePEc author service in 1999.
- The system was written by Markus J.R. Klink.

author claiming history II

- ISI created researcherID in 2006 (?)
- arXiv have an author claiming system since 2009.
- NIH and Google Scholar are working on it.
- The ORCID initiative is looking into author identification since 2009.

claiming vs identification

- Author claiming records are NOT author identification records.
- The difference is called “Klink’s problem”.
- An person can claim to be an author of a paper. If there are several author, we don’t know what author (s)he is.

Klink's problem example

- Jane and John Smith write a paper.
- Author list say "J. Smith and J. Smith"

AuthorClaim data

- ftp://ftp.authorclaim.org
- CC0
- more than 100 profiles, growing slowly.

an example

- id: pbi1
- name variations: Geoffrey Bilder — G. Bilder — Bilder, G.
- isauthorof: info:lib/elis:856
- hasnoconnectionto: info:lib/pubmed:11127885 — info:lib/pubmed:7482633

more on the example

- The refused papers are there for services to build learning models for author names. Actually learning is an integral part of the way AuthorClaim works.
- Actually records also contain the 3lib data for papers.
- and they have ARIW-base affiliation data.

IRs and author identification

- IRs are generally too large to author identification by IR staff.
- Only registration of contributors is usually required.

IRs and author claiming

- IRs are too small to make it meaningful for authors to claim papers in them directly.

benefits of author claiming to IR

- All papers by an author can be put together.
- The task can be completely automated once an AuthorClaim record claims a paper in the IR.

to get it done

- From the four elements in a 3lib record only the link to a web page describing the item is problematic.
- But is cumbersome to customize to close to 2k IRs.

partnership with BASE

- We need a centralized collection.
- BASE is already doing this job.
- BASE can deliver all data to AuthorClaim regularly.

BASE aggregation services

- constant monitoring of OAI-PMH world
- configuration of harvesting for each new and erroneous repository
- metadata stores (raw and normalized)

BASE normalization

- highly heterogeneous use of OAI-DC requires cleaning and enrichment, e.g.
  - dc:type
  - dc:date
  - dc:language
- also enrichment with (missing) subject classifications

BASE search services

- builds index (solr)
- end user interfaces (vufind)
- iPhone-App
- API for index usage by third parties (http or SOAP)

BASE data services

- repository profile service (REST)
- raw metadata store access (http or SOAP)
- rsync for AuthorClaim

BASE data in AuthorClaim

- selection of records that have required data
  - author
  - title
  - link
  - identifier
- incremental updates

repository exclusion

- From BASE profile AuthorClaim discards some IRs that contain
  - student work
  - digitized old material
  - link collections
  - primary research data
- There are some minor manual exclusions.

results so far

- 1930 repositories, 12740116 records.
- 534 records claimed.
- The documentation at http://wotan.liu.edu/base/ needs some debugging.
- The collection is not yet announced because it is being read.

the end

- Contact
  - whorstma@uni-bielefeld.de
  - krichel@openlib.org

  for more information.