

Open Access to Scholarly Metadata: Author Claiming and Institutional Identification

Thomas Krichel^{1,2,3}

¹Long Island University ²Novosibirsk State University ³Open
Library Society

thanks

- to the organizers
- to the Open Society Institute
- to the JISC
- to Robert James Griffin, III
- the RePEc gang

structure

- background
- the RePEc Author Service
- author identification
- author claiming

Valentine's day myth

- woman and man fall in love
- it is very emotional, irrational
- and there they stay

the love of my life

- the collection and care of academic data
- the free sharing of such data
- I am working on things that are long-run

historic context

- like lovers reproduce the behaviour their parents
- my passion of sharing does not come out of context

not devoid of historic context

- In the academic world people have long been sharing to build the human knowledge.
- Academics did not sell the access to their research papers.

as an economist

- Since 1990 I have been astounded that
 - academics gave their contents away,
 - scientific journals where expensive and inconvenient.

irrationality obscured by practice

- Stevan Harnad has been most vociferous
- he has been apt advocating some ideas I held before him.
- Today is the 10th anniversary of the Budapest Open Access Initiative.

my inspiration

- It's the open source software movement.
- Ideally, human knowledge should be like a set of open source software.
- That not being currently feasible, at least metadata about documents should be.

now in economics

- Economists have had a system of without non-commercially intermediation.
- This the working papers system.

working papers

- Recent research papers written by research staff in an institution,
- circulated on exchange base.
- stored in coffee rooms.

bringing this to the Internet age

- I created a project called NetEc.
- As a part of that project I published the first online economics working paper.

the economics working paper archive

- At that time, Paul Ginsparg's xxx.lanl.gov, later the arXiv.org was all the rage.
- Robert B. Parks adopted it to economics.

central vs decentral

- Bob and I quarreled a lot.
- He had the Lyon's share of visibility.
- I did not think his decentralized system would work.
- My ideas won, but

centralized and decentralized

- We created a system that was both centralized and decentralized,
- based on a set of institutional repositories,
- in 1997, way before that term was in common use.

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 encourages 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.

institutional registration

- It is done by a single individual, Christian Zimmermann, (CZ)
- He created a registry for all economics departments that have a web page.
- This data is reused.

personal registration

- I created the RePEc Author Service RAS in 1999.
- Initially called “HoPEc”.
- The first programmer was Markus J.R. Klink.

stage i

- At initial registration, the author gives personal information.
 - email address
 - name and name variations

name variation

- It is assumed that a paper may have been written by an author if it has matched one of her name variations.
- RAS also performs some fuzzy searches offline to spot spelling mistakes.

institutional affiliation

- Registrants can search the EDIRC database for names of institutions they work at.
- When a matching institution is found it can be added to the list of institutions a registrant is affiliated with.

new institutions

- RAS contains a proposal screen for new institutions a registrant can claim to be affiliated with.
- Items are entered as string data in the profile.

officializing

- When CZ adds an institution following an accepted proposal he replaces the string data in the registrants profile
- The registered institution handle is henceforth used.

stage ii: claiming papers

- This is the heart of RAS.
- Authors claim or disclaim papers that carry a name variation of their.
- There is an email alert service for new matching papers.

success of RAS

- over 30k authors registered
- from an old independent list of top 1000 economists over 80% are registered.

reason for success of RAS

- RePEc has collected a lot of data
 - download data
 - citations data
 - classification data
- we build rankings. You can only rise in ranking if you claim papers.

a RePEc for all disciplines

- RePEc bibliographic data → 3lib
- RePEc Author Service → AuthorClaim
- EDIRC → ARIW

author identification theory

- There are documents.
- There are authors who wrote the documents.
- Authors are identified when we know what person wrote what document.

limitation

- Note that my setting I deliberately ignore the fact that
 - There are other relationship type other than authors.
 - We may be interested in document collections.

currently

- Authors are referenced on documents by name expressions.
- There is no universal personal identification scheme to piggy back on.

ultimate solution

- Author identification is a temporary problem until a government-backed identification scheme becomes widely available.
- For example, something like the US Social Security number
 - generalized across countries,
 - without problems of id theft.

name references

- Name references are clearly insufficient.
- But the insufficiency is unevenly distributed.
- It affects people with common name expressions.
- It affect incidences of short name expressions.

name disambiguation

- We can try to extract context data from the documents and try to disambiguate authors by building sets of documents presumed to be from different authors.
- We can call this author name disambiguation.

disambiguation vs identification

- I (maybe others) say that there is disambiguation when there are sets of document written by a presumed same author using machine compilation.
- We refer to author identification when the identity is confirmed by a trustworth person.

a librarian

- Librarians have been operating a system of authority control.
- Authority control means
 - deciding for each person what variant of the name is authorized,
 - using this form for all

here comes the elephant

- When we talk about author identification, we refer to a collection of documents. I will call this the corpus.
- What is the corpus?

In a library...

- In the library, the corpus is what the library has collected.
- It is possible have cataloging staff to use authority control to solve the author identification for the *non-periodicals* in the collection.

periodicals

- Libraries don't catalog periodical contents.
- They relied on 3rd parties for this.
- Before RAS, none of these 3rd parties had author identification.

in 1999, comes in RAS

- This is the first time authors get involved in author identification.
- First author identification system for periodical contents.

back to the elephant

- The corpus of RAS here is the RePEc database.
- The incentives for authors are to create profiles so that they can appear in rankings.

how many claiming system

- Pitman's approach: create a bunch of claiming system. Create a system that federates them.
- Krichel's approach: create a vast bibliographic database. Have authors claiming for that dataset.

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 and institutional repositories

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.

advantages AuthorClaim

- Bulk data freely available
- Robust and simple design.
- Avoids authors to work with many claiming systems.

limitations to AuthorClaim

- It is limited to author claiming, rather than author identification +2
- It is useless on its own.

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”

isolated uselessness

- AuthorClaim only aims to produce a machine-readable set of 3lib data about the documents that the author wrote and did not write.
- We have to integrate these data into other systems.

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.

the end

- Thank you for your attention
- <http://openlib.org/home/krichel>