

Author identification: theory and current state of play

Thomas Krichel^{1,2,3}

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

Southampton 2012-03-19

thanks

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

structure

- background
- history: the RePEc Author Service
- author identification theory
- IRs and claiming systems

why are we here?

- We are interested in scholarly communication.
- We believe in open access.
- I believe in self-sustaining systems.

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.

my reason for complacency

- RePEc is a system that builds a free bibliography and free full text for Economics.
- It took a long time to build as a self-sustaining entity.
- It can be seen as a prototype.

institutional repository system

- RePEc archives are institutional.
- They are light-weight and old fashioned.
- They are better integratable than conventional IRs.

IRs are similar to RePEc

- IRs as a system are unfunded.
- They took a long time to build and their academic contents is growing slowly.
- What can they learn?

RePEc's added value

- 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.

an uneven problem

- 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 computation.
- We refer to author identification when the identity is confirmed by a trustworthy 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 and
 - using this form for all references to the author.

the problem

- 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.

and the corpus

- 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 to scale this up?

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 40 million records from the likes of 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 of 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 +3
- 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"

Klink's consequences

- Author identification can only be achieved if identifiers are deployed in bibliographical data.
- Problem is that most bibliographical data formats don't have a field for author name identifiers.

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.

slow growth

- Whenever self-claiming is involved, growth must be expected to be slow.
- It's like the green road to open access.
- We have to have a conviction we are on the right path.

AuthorClaim record rough 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

- Actual records contain the 3lib data for papers and ARIW-base affiliation data.
- The refused papers data can be used for learning about author names.
- Learning is import to the internals of AuthorClaim.

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.
- Usually, only a contributor is identified.'

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.

for IR design

- Ideally an IR should be able to be working with a bunch of author claiming system.
- A generic protocol does not need to be written, but say for EPrints, you want to have a general spec.
- I have time in the Summer to work on this.

author pages

- At the simplest, repositories can implement author pages.
- These would assemble the works of the authors.

author page items

- We can have links to local items.
- We can have links to remote items.
- We can have search items.

implementation

- Author records have to be harvested.
- This can be done by mirror from AuthorClaim.
- For other systems the processes may be more complicated.

local vs remote split

- Metadata is supposed to exhibit document records for accepted documents.
- There needs to be per author record collection way to split to local identifiers.

benefits

- Author gets more comprehensive list of works. This improve a sense of “my archive” .
- IRs get inbound links. This improves search engine rankings.
- Better aggregate information about paper versions.

author claiming to author id

- Author identification has to be carried out at the publishers' level.
- Since we are adopters of the green approach let us think IR.

theory

- Let us again look at some theory.
- This concerns levels of interoperability between

level 0

- There is an author claiming system, say AuthorClaim.
- There is an institutional repository, say EPrints.
- There could be more IRs and many author claiming systems, but that's not a problem.

level 1

- EPrints makes bibliographic data available.
- This is currently in the process of being fully realized more on this later.

level 2

- EPrints document data contains identifiers for some authors.
- That identifier data will have to be provided by depositors or admin staff.

level 3

- EPrints has a facility to help metadata curators to discover identifiers known to AuthorClaim.
- This can be done in a centralized facility.

level 1x

- In level 1x EPrints can push metadata to AuthorClaim.
- AuthorClaim processes this data immediately.
- The profile of the author is updated.

ACIS

- All levels of interoperability have been implemented in the past for EPrints 2.x and ACIS.
- Ivan Kurmanov write a patch for EPrints 2.x at the time.
- Doing it for EPrints 3 would be easier.

issues with level 2

- A key problem is that IR OAI DC metadata has definitely has no space for identifiers.
- EPrints would need to implements AMF or something similar.

issues with level 3

- Level 3 is useless if level 2 has not been implemented.
- A useful service could be formed by providing a query interface for claiming data from many systems. It could provide revenue.

implementation of level 1

- I discuss steps to implement level 1 for institutional repositories.

IRs and 3lib

- DC to 3lib is not all that hard.
- Let's look at it by element.

title

- DC:title \longrightarrow title
- problem: no problem

author

- DC:creator \longrightarrow author
- problem: separation where multiple authors in one value.

handle

- DC:identifier can not be used, it is overloaded.
- OAI identifier is better, but there are a lot of Eprints:generic around.
- We need OpenDOAR or ROAR.

displaypage

- DC has no field for this.
- IRs often have this in their records but they tend to put it into different places.
- It's a huge job to fiddle this out.

4th November movement

- This is an informal association around the BASE, the Bielefeld Academic Search Engine.
- BASE has a lot of IR data, and they maintain it.
- Aim is to make it more widely available.

public data

- BASE make metadata about the repositories available.
- `http://basestore.ub.uni-bielefeld.de:9999/unibi-base-repository-index-service`

informal agreement

- The Open Library Society has an informal agreement with the BASE gang for the delivery of item level data.
- Done with rsync special key.

sample record

```
<document id="info:lib/base:CSIC:oai:digital.csic.es:10261/9348">  
<element name="dctitle"><value> INCISO:  
Automatic Elaboration of a Citation Index in  
Social Science Spanish  
Journals</value></element> <element  
name="dccreator"><value>Barrueco
```

sample record

Cruz, José Manuel ; Osca-Lluch, Julia ;

Krichel, Thomas ; Blesa, Pedro ;

Velasco Arroyo, Elena ; Salom,

Leonardo</value>

</element><element name="dcyear"><value>2005

</value></element><element

name="dclink"><value>

<http://hdl.handle.net/10261/9348></value>

</element></document>

state of play

- These records are being read into AuthorClaim.

other initiative

- I am surveying other initiatives.
- A main concern is how open they are.

researcherID

- A RAS clone by Thompson Reuters.
- Uses ISI dataset.
- Neither bibliographic nor profile data available for free in bulk.

researcherID development

- Thompson Reuters realized they would not be able to do it on their own.
- They now participate in ORCID.

VIAF

- A merger of national authority files
- started as an OCLC research project
- 24 million source records.

VIAF

- 25 million links to OCLC bib record
- bulk download possible, apparently
- ODC BY license (?)

ISNI

- grew out of VIAF, now an ISO standard.
- registration through agencies, one ready
- initial database built

INSI membership

- membership from national libraries
- membership from contents industries
- membership from rights organizations drives

INSI records

- Some data is useful for disambiguation,
- Much data is given in confidence and stays confidential.
- The public view is small.

ORCID

- created in 2009
- US non-profit formed in 2010
- 2012Q2 to see start based on researcherID code.
- I am a member of the technical architecture group.

version 1 and 2

- initial version based on self-claim
- if an organization is close, the institution will be able to manage the profile
- in a version 2, claims are supposed allow registered parties to make claims about each other.

initial use case

- initial use case around manuscript tracking
- service limited to CrossRef data via sigg (?)
- integration with Scopus Author IDs

ORCID and open

- ORCID is open in the sense that anybody with an interested
- Committed to open source their software,
- Annual CC0 dump of (individual?) user contributed data.

the end

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