Open Metadata for Expertise in Economics, and beyond

Thomas Krichel$^{1,2,3}$

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

Paris 2012–03–28
thanks

- Emma Bester
- Open Society Institute and JISC
- the RePEc gang
- Bert Wendland
unclear purpose

- I am not sure what this seminar is about.
- I simply picked up on discipline and expertise talk.
- But I am a walker rather than a talker.
what gets me walking?

- I’m interested in scholarly communication.
- I am interested 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.
the open library

- Basically it’s a set of data about documents and related aspects of reality related to documents.
- I don’t want to pursue the theory here.
- I’ll introduce an example.
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.
so I am a disciplinarian

- 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.
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.
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 lion’s share of visibility.
- I did not think his decentralized system would work.
- My ideas won, but
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
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 runs 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.
RePEc in document number

- over 1400 archive from 75 countries
- 1,2 million items documented, from
- 1400 journals and 3300 working paper series
outlook on the subject of “expertise”

- going down: NEP
- going up: AuthorClaim
NEP: New Economics Papers

- A system I created in 1998.
- Name coined by Sune Karlsson.
- Initial coding by José Manuel Barruco Cruz
the task I

• We want to report new papers that come into RePEc.
• We need to exclude journal articles.
• Only working papers will be used.
the task II

- We want to split by subjects.
- We use editors who will split the papers into subjects.
- Each editor works on one or more topic only.
early

- We collected new additions in into something like an a subject specific email. That email contained all papers.
- Ask editor to edit out non-pertaining paper and forward to email list.
problem

• A study 2002 study of mine with Jeremiah Cochise Trinidad Christensen revealed inparsability of the NEP record.

• A new system needs building with a view on observability.
ernad

- stands for "editing new reports on academic documents"
- I wrote the "Altai paper" that specs it.
- Roman Davidovich Shapiro wrote first version in Perl.
who volunteers for NEP?

- Most of the volunteers are junior academics.
- They have good incentives.
  - need to be aware of latest literature;
  - are absent from the informal circulation channels of top level academics;
  - need to get their name around among researchers in the field.
NEP in numbers

- over 70000 subscriptions
- over 30000 subscribers
- over 40000 reports issued
- over 90 reports
huge dataset

- subscriber data
- editing behavior data
- learning evaluation data
- downloads from reports data
cross penetration

- reports space is flat but
- reports intersect with papers
- reports intersect with subscribers
research work to be done

- improving learning
- reports intersect by papers
- reports interest by subscribers
research work to be done II

- evaluating editor performance: relate target values to observable behavior pattern
- continue study on coverage of NEP as a whole
RePEc and expertise

- NEP can classify papers.
- But the expert is a human.
- So we need data on authors.
remember this?

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

- NEP data can be combined with personal data.
- This can lead to real-time identification of experts in areas.
NEP and institutional data

- NEP data can be combined with personal data and institutional data.
- This can lead to real-time identification of specialization of institutions.
RePEc for all disciplines

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

- RePEc bibliographic data → 3lib
- RePEc Author Service → AuthorClaim
- EDIRC → 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 40 million records from the likes of PubMed, OpenLibrary, DBLP, RePEc and institutional repositories
3lib problem I

- The sources have heterogeneous formats.
- We can try to unify the formats but it's a lot of work.
- We concentrate on what we need for the other steps.
3lib problem II

- There is unequal coverage of different disciplines.
- We try to cover more, but it’s not easy.
3lib problem III

- There is a multiple coverage of a single work.
- It’s better to leave this as we need to furnish data to the providers.
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 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
- It is useless on its own.
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

- CU0
- more than 100 profiles, growing slowly.
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
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.
author pages

- At the simplest, repositories can implement author pages.
- These would assemble the works of the authors.
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 identification has to be carried out at the publishers’ level.

Since we are adopters of the green approach let us think IR.
Let us again look at some theory. This concerns levels of interoperability between...
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.
EPrints makes bibliographic data available.

This is currently in the process of being fully realized more on this later.
EPrints document data contains identifiers for some authors. That identifier data will have to be provided by depositors or admin staff.
EPrints has a facility to help metadata curators to discover identifiers known to AuthorClaim.

This can be done in a centralized facility.
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 wrote 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 implement 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.
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.
DC:title → title

problem: no problem
author

- DC:creator $\rightarrow$ 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 if Eprints:generic around.
- We need OpenDOAR or ROAR.
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

<element name="dctitle"><value>INCISO: Automatic Elaboration of a Citation Index in Social Science Spanish Journals</value></element> <element name="dccreator"><value>Barrueco</value></element>
sample record

Cruz, José Manuel ; Osca-Lluch, Julia ; Krichel, Thomas ; Blesa, Pedro ; Velasco Arroyo, Elena ; Salom, Leonardo

<element><element name="dcyear"><value>2005</value></element><element name="dclink"><value>http://hdl.handle.net/10261/9348</value></element></element></document>
state of play

These records are being read into AuthorClaim.
the end

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