

# Authors, their names and documents. New thoughts on an old topic

Thomas Krichel<sup>1,2,3</sup>

<sup>1</sup>Long Island University <sup>2</sup>Novosibirsk State University <sup>3</sup>Open Library Society

Bielefeld University 2011-05-06

# themes

- socio-economic themes
- technological themes

# structure

- Open bibliographic data
- Author claiming on such data
- AuthorProfile

# open bibliographic data

- There is a growing (albeit, slowly growing) movement for open bibliographic data.
- Open bibliographic data is bibliographic data that comes with more liberal licensing conditions for its reuse.

# who is involved

- The main player here is the Bibliographic Data Working Group of the Open Knowledge Foundation (“the group”)
- The group works on two types of data.

# bibliographic data types

- There is basically two types
  - library catalog data
  - academic paper data
- Our interest is area 2.

# the Open Library Society

- The society complements the group.
- OLS aggregates bibliographic data without insisting on legal conditions.
- OLS builds useful services that make contributors happy.
- It's the road that RePEc took in 1993.

# a 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 35 million records from the usual suspects: PubMed, OpenLibrary, DBLP, RePEc.
- May contain some BASE data?

# 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 is modeled after the RePEc Author Service
- It uses the same software, called ACIS.
- It is running since early 2008, predates ORCID.

# author claiming

- 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 problem example

- Barak Obama claims a paper
- Author list says “Obama B., Laden et alii”

# AuthorClaim data

- <ftp://ftp.authorclaim.org>
- CC0
- only about 100 profiles

# 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



# AuthorProfile

- AuthorProfile is an information service that combines 3lib and AuthorClaim data.
- Joint work with Robert J. Griffin III, since late 2010.
- The site is running at <http://authorprofile.org>.

# vision

- We believe that scholarly communication is for authors more than for readers.
- Authors have to take priority.
- We want to invert bibliographic data from its conventional structure into a CV-like structure.

# problem

- Without author claiming, it is based on author name expressions.
- It is therefore error-prone.
- It gives us the chance to show how poor name data can be.
- It may also incentives people to register.

# first step: auversion

- We build data files that contain all papers written by authors with the same name expression.
- This leaves us with a large number of “auverted files” .
- Each contain the document by an author name expression, henceforth aunex.

# navigation

- We can navigate auverted data through co-authorship relationships.
- In addition we build
  - horizontal integration
  - vertical integration
  - search

# horizontal integration

- Horizontal integration links name expressions which may represent the same author.
- For example: “F. Lefèvre”  $\leftrightarrow$  “Fred Lefevre”
- This feature has yet to be implemented.

# vertical integration

- Vertical integration explores collaboration between authors and aunexes.
- From aunexes, we can link to authors.
- We want to choose one upstream author or aunex to link to.

# vertical integration

- Vertical integration explores collaboration between authors and aunexes.
- From aunexes, we can link to authors.
- We want to choose one upstream author or aunex to link to.



# closest author from aunex

- An aunex may have collaborated with several authors
- We use collaboration strength.
- At equal strength we use a ranking of authors with home-baked criteria.

# search

- A search feature gives priority to registered authors
- If more than one author is identified by the same name, links to both authors shall be returned by the query
- This feature has yet to be implemented.

# ranking

- We intend to compute a pageRank type algorithm to rank authors.
- At this time authors are the only entry points to the collection.

# self-documenting

- There is a special documentation interface at <http://authorprofile.org//>
- It shows the entire software written for the project both in Perl and XSLT.

# problems with AuthorProfile

- Scale, scale, scale
- The problem of representing negativity

# the end

- Thank you kindly for your time, and for your attention.
- AuthorProfile is joint work with James R. Griffin III.