RatSWD Working Paper Series

www.ratswd.de

RatSWD.
German Data Forum

222

Author Identification in Economics, ... and Beyond

Thomas Krichel
Christian Zimmermann

August 2013

Federal Ministry of Education and Research

Working Paper Series of the German Data Forum (RatSWD)

The *RatSWD Working Papers* series was launched at the end of 2007. Since 2009, the series has been publishing exclusively conceptual and historical works dealing with the organization of the German statistical infrastructure and research infrastructure in the social, behavioral, and economic sciences. Papers that have appeared in the series deal primarily with the organization of Germany's official statistical system, government agency research, and academic research infrastructure, as well as directly with the work of the RatSWD. Papers addressing the aforementioned topics in other countries as well as supranational aspects are particularly welcome.

RatSWD Working Papers are non-exclusive, which means that there is nothing to prevent you from publishing your work in another venue as well: all papers can and should also appear in professionally, institutionally, and locally specialized journals. The *RatSWD Working Papers* are not available in bookstores but can be ordered online through the RatSWD.

In order to make the series more accessible to readers not fluent in German, the English section of the *RatSWD Working Papers* website presents only those papers published in English, while the the German section lists the complete contents of all issues in the series in chronological order.

Starting in 2009, some of the empirical research papers that originally appeared in the *RatSWD Working Papers* series will be published in the series *RatSWD Research Notes*.

The views expressed in the *RatSWD Working Papers* are exclusively the opinions of their authors and not those of the RatSWD.

Contact: S C I V E R O Press | German Data Forum (RatSWD) | Mohrenstr. 58 | 10117 Berlin | office@ratswd.de

Author Identification in Economics, ... and Beyond

Thomas Krichel
Novosibirsk State University
Christian Zimmermann*
Federal Reserve Bank of St. Louis, IZA, CESifo and RCEA

July 30, 2013

Abstract

Identifying authorship correctly and efficiently is a difficult problem when the literature is abundant, but poorly recorded. Homonyms are tedious to differentiate. This paper describes how the field of economics has organized itself with respect to author identification. We describe the RePEc project with a special emphasis on the RePEc Author Service. We then discuss how the concept is currently being expanded to the entire scientific body with the AuthorClaim project.

JEL Classification: A14
Keywods: author identification, author service, AuthorClaim, RePEc

*The views expressed are those of individual authors and do not necessarily reflect official positions of the Federal Reserve Bank of St. Louis, the Federal Reserve System, or the Board of Governors.

1 Introduction

Librarians and others who assist researchers and organize information often need to categorize published articles. As an example, a librarian may need to categorize a series of papers, including one by Robert Lucas. The paper is part of the literature in economics, so it should not be that problematic to attribute it to the right person. Well, it turns out that there is a Robert E. B. Lucas, a Robert F. Lucas and a Robert E. Lucas, Jr. To make things worse, the last two both work in the same sub-field, macroeconomics.

Luckily, economics is rather well organized for this task thanks to the RePEc project. RePEc stands for Research Papers in Economics and was founded in 1997 on the foundations of NetEc, itself dating back to 1992. The goal of RePEc has been to improve the dissemination of research within the field of economics. It does so by letting publishers and research institutions index their works and then disseminates them through email and the web. This is done at no cost for all parties.

The need for such a service arose because publication lags are extremely long in the field of economics—the review process alone is measured in years—so that a pre-print culture has established itself. At least until 1992, pre-prints were, however, disseminated only among those within a close circle. The consequence was that the frontier of research became apparent to those outside this circle through journal publication many years after the actual research was conducted. The lags made it very difficult for outsiders to contribute to it.

RePEc has democratized the dissemination of research in economics, both for authors and readers, and made it possible for anyone to be aware of the research frontier. An important collateral effect of the project has been the RePEc Author Service, in which authors create personal accounts and identify their works among the papers listed in RePEc. Important complementarities between the bibliographic aspect of RePEc and the RePEc Author Service, along with properly aligned incentives for all participants, allowed RePEc to grow to encompass all major publishers and an impressive number of authors.

This paper describes RePEc, how it collects its bibliographic metadata and how the RePEc Author Service (RAS) played an important role in it. We then look a bit more into the detailed workings of RAS. Finally, we discuss how the concept can be expanded to the scientific literature in general, with proper modifications.

2 RePEc

Technically speaking, RePEc is just a set of metadata definitions and principles that drive the organization of metadata files on public servers (Barrueco Cruz and Krichel, 2000). RePEc was founded in May 1997 at a meeting at Thomas Krichel's apartment in Guildford, United Kingdom, of various people involved in the collection and cataloging of pre-prints in economics. There, the Guildford Protocol was adopted (Krichel, 1997a) on the foundations of the NetEc project

that was active since 1993 (Krichel, 1997c). It specifies that metadata files in RePEc need to be flat text files with an *.rdf* extension. There is one file describing the archive hosting the metadata. Another file describes the series contained in the archives, of which there are currently five types: working paper series, journals, books, book chapters, and software components. The latter are code and sometimes accompanying data used to replicate research. The metadata files with the contents of each series are then located in separate directories.[1]

All this is hosted on an http or anonymous ftp server of the publishing institution, which then maintains the metadata. The only central operation of RePEc is then to hold a list of the archive descriptors, thereby allowing anybody to gain access to the decentralized metadata collection.

RePEc does not have an official facility that uses the collected data and disseminates it to users. Instead, the data is put in the public domain and several so-called RePEc services have been developed that use the metadata in various ways, such as an email dissemination service (NEP, http://nep.repec.org/, see Chu and Krichel, 2003, and Barrueco Cruz and Krichel, 2005), websites with search and browsing functionalities (the most prominent ones being EconPapers, http://econpapers.repec.org/, and IDEAS, http://ideas.repec.org/), citation analysis (CitEc, http://citec.repec.org/), the RePEc Author Service to be discussed below, and more (see http://repec.org/). Other service may be created by people who want to use the RePEc data in new ways.

The metadata itself is following the Research Document Information Format (ReDIF), which was created specifically for RePEc (Krichel, 1997b). Indeed, there was no accepted format for bibliographic metadata at the time of its inception.[2] Given the decentralized nature of RePEc and the fact that many non-specialists are maintaining RePEc archives, the metadata format was purposely kept very simple. It also has relatively few mandatory fields in order to facilitate the inclusion of as many works as possible. A minimal template for a journal article would be

```
Template-Type: ReDIF-Article 1.0
Title: Another boring study about some economic question
Author-Name: Adalbert Pumpernickel
Handle: RePEc:aaa:boring:1234
```

Note that 1.0 stands for the metadata format version, and it is still at 1.0, a remarkable stability for over 15 years. Note also that the handle here predates the Handle System independently launched in 2003. Finally, information about the journal where the exciting study described above is published is implied from a similar series template. A more common representation of the metadata would be the following:

[1] For instructions on how to participate in RePEc, see Zimmermann (1997).

[2] Which explains the choice of the *.rdf* file extention, which unfortunately later collided with the RDF (Resource Description Framework) metadata file extension.

```
Template-Type: ReDIF-Article 1.0
Title: Another boring study about some economic question
Author-Name: Adalbert Pumpernickel
Author-Workplace-Name: Department of Pontification, University
 of Grandiose Thoughts
Abstract: This study fills scarce journal space with some
 random thought about important economic questions of the day.
Journal: Biannual Journal of Economic Writing
Pages: 23-46
Issue: 1
Month: January
Volume: 15
Year: 2012
Classification-JEL: A01; Z44
Keywords: economic issues; economic literacy; hysteresis
File-URL: htto://www.jeconwriting.org/pdfs/2012/1234.pdf
Handle: RePEc:aaa:boring:1234
```

Several other fields can be specified, including distinguishing first and last names of authors, links to other versions, contact details, or supplementary material. Unfortunately, few publishers provide this information. Either they do not have it or they do not think it is worth the effort.

It may seem impossible to impose this kind of indexing rigor on publishers, but many publishers have participated thanks to proper incentives. These incentives arose because the publishers wanted to make sure their material is present and specifically thanks to the RePEc Author Service. As of the writing of this paper, close to 1600 publishers have joined RePEc, from the largest commercial publishers and repositories to small research centers, located in 75 countries. The metadata cover over 3700 pre-print series, 1700 journals, among others, adding up to 1.4 million works, growing at over 10% per year despite the relative maturity of the project.

3 The RePEc Author Service

The origin of the RePEc Author Service was a collection of links to homepages of economists. It was started by Barry Schachter in 1996 under the name HoPEc. Soon, it became unsustainable to be managed by a single person. The RePEc Author Service (RAS), initially still under the name HoPEc, would allow people to create a record for their homepage in RePEc from 1998. It became functional as a paper claiming system—where authors would create a portfolio of their works listed in RePEc—in 1999. A first version was written by Markus Klink, supported by funds from the Joint Information Systems Committee to the WoPEc project. It allows users to create an account to control the information about them. This includes creating a portfolio of their works listed in RePEc. This can be called an author claiming service. RAS was the first such service ever

created. The current incarnation of RAS at http://authors.repec.org/ started in 2004 with a complete rewrite of the code by Ivan Kurmanov. The rewrite was funded by the Open Society Institute. Note that the software on which the RePEc Author Service has been developed, ACIS, is available in open source under a GNU general public licence. See http://acis.openlib.org/.

For this purpose, the economist is asked to provide a list of name variations (beyond what can be automatically inferred from the first, middle and last names) so that the service can search for matches in the bibliographic metadata. This has become invaluable for those whose name are commonly misspelled or who have changed name through marriage, divorce or religious conversion. It also allows us to distinguish between homonyms. Authors also provide their affiliation(s), picked from a large database of institutions within the scope of economics. Authors are provided with a unique and public identifier that can be used by other RePEc services. There are plans for it to be used for authentication in other RePEc services.

A typical metadata template for an author would look like this:

```
Template-Type: ReDIF-Person 1.0
Name-First: Adalbert
Name-Last: Pumpernickel
Name-Full: Adalbert Pumpernickel
Workplace-Name: University of Grandiose Thoughts
/ Department of Pontification
Workplace-Institution: RePEc:edi:dpugtuk
Email: a.pump@mail.grandiose.ac.uk
Homepage: http://fac.grandiose.ac.uk/~a.pump/
author-paper: repec:zzz:sleepy:wp45
author-paper: repec:zzz:sleepy:wp58
author-paper: repec:zzz:awaken:2008-11
editor-book: repec:aaa:bedtim:bk234
author-chapter: repec:aaa:asleep:bk234-ch1
author-article: repec:aaa:boring:1234
Short-Id: ppu1234
Handle: repec:per:1968-09-05:adalbert_pumpernickel
Last-Login-Date: 2011-01-05 06:16:45 -0500
```

These templates are disseminated in the same ways other bibliographic templates from RePEc are. RePEc services can thus use the person template and choose what information to extract from them. Note that authors can choose not to publicize their email address and can use the Short-Id as an identifier, for example in Wikipedia templates for economists.

Templates internal to the RePEc Author Service contain much more information, in particular items rejected by authors, name variations and recognized citations. Only the latter are exported, to the CitEc project that does citation analysis and uses the RePEc Author Service as a support tool for the fuzzy matching algorithm, where authors help approve marginal matches.

In 2007 ACIS added citation recognition. RePEc tries to identify citations using a fuzzy matching algorithm. Where there is little confidence in the matching, human help can prove useful and authors are happy to help, especially if it increases their citation counts.

This brings us to the incentives for participation. All services started with a small core of enthusiastic believers in the project. But this is not sufficient to sustain each of the projects. First, the services need to add value to the collected data. Second, there needs to be critical mass for services to become useful at all. Third, there need to be complementarities between services to allow one service to grow when another grows.

The purely bibliographic aspect of RePEc and the RePEc Author Service are natural complements (Krichel and Zimmermann, 2009). Authors want their works listed and encourage publishers to participate. This complementarity has been critically strengthened by the use that has been made of the collected data. The computation of rankings of authors and their institutions, and later of impact factors for publication series has been particularly helpful. Indeed, to be well ranked, authors have pestered their various research outlets to contribute metadata to RePEc. And they have also asked their colleagues to register to boost their institution's rankings, amplifying the complementarity. As RePEc grew, more people recognized that this was worth participating in.

As of the writing of this paper, over 36,000 authors are registered. The vast majority (order of magnitude: 85 to 90%) of the top economists, as recognized by meaningful studies[3], are registered. The total RePEc membership is more than double the membership of the largest society of economists, the American Economic Society. 860,000 claims to works listed in RePEc populate their research portfolios. We believe this remarkable success is largely due to the fact that the right incentives have been in place.

4 Beyond economics

What are the lessons of this beyond economics? What would a RePEc for all disciplines look like? There are two answers to this.

A system like RePEc that would encompass all disciplines would register every single published academic work. It would relate the work to its authors and the authors to their institutions. It could then be used to automate many quantitative measures of academic performance evaluation. This is the technical answer to the question.

A system like RePEc that would encompass all disciplines would be one where holders of academic information about papers, authors and institutions make such information instantly freely available, in such a way as to form a dataset. Records from the dataset can then be combined. This is the business case answer to the question.

Both answers do not contradict each other, they reinforce each other, as demonstrated by AuthorClaim, an author registration service across sciences

[3]For example, Coupé (2003).

available at http://authorclaim.org/.

Bibliographic databases

There are, of course, initiatives that have compiled bibliographic databases for certain subjects or certain types of material.

The granddaddy of all article-level bibliographic datasets is the PubMed dataset of the U.S. National Library of Medicine (NLM). This is funded by the government of the United States and contains over 22 million records. The collected metadata is extensive. It does not only cover basic bibliographic information but also contains subject classification data. The data is available at no cost, but it is not freely available. To get access to it, one has to enter a licensing agreement with the NLM. It is not clear if the license allows for the building of author claiming services based on it. The language of the license simply has not been written with such an application in mind. The ERIC database, commissioned by the US Department of education, is funded in a similar way. Its license appears to be less restrictive.

Most bibliographic datasets for disciplines are not available but zero payment access may be negotiated on a case-by-case base. For freely available datasets, we have two examples. One is the DBLP library. It is maintained by Michael Ley of Trier University (http://dblp.uni-trier.de/). It offers over 2.3 million bibliographic. The records are fairly minimal, but the quality is very high. The data is available in bulk, but when say bulk, we mean it. It is is one large XML file. This requires special processing as the decoded XML would take too much memory under most machines. The web interface to DBLP has identified authors, presumably manually maintained by DBLP contributors. The downloadable XML file does not make that data available.

The second example of a free data set is AGRIS (http://agris.fao.org/), by the United Nations Food and Agricultural Organization (FAO). It offers over 4.3 million citations from the agricultural literature. The data are collected by national providers. Its quality is variable. It is freely available from an ftp server based at the FAO.

For high energy physics, the SPIRES dataset (http://inspirehep.net/) furnishes about 500,000 references to papers that are not in the arXiv.org server. It is not freely available, but it may be obtained by negotiation. Similarly, the Solis database, compiled by GESIS Leibnitz Institut fuer Sozialwissenschaften, Germany, can be obtained only by negotiation. It mainly covers German language or German based social science publications. Among them is *Schmoller's Jahrbuch*.

For discipline-based collections, in the natural sciences alone, important gaps are apparent in mathematics, physics, chemistry and biology. Other sciences fare worse. There are vast swaths of the literature for which no open access bibliographic index is avaible. Even when one is available, it is organized in a different way from the others. To bring them together, Thomas Krichel has created the 3lib project, pronounced freelib (http://3lib.org/). The idea is to create a large bibliographic dataset that can be used to build an interdisciplinary author identification system. This system needs to build on some existing dataset, if only to create critical mass for starters.

In recent years, universities have started to set up institutional repositories for output produced by the university. While initially such repositories were conceived to contain scientific papers, their scope has grown over time to include student work as well as digitized cultural assets. No official central repository is available that would register all such repositories. Given the varied nature of the material in such repositories, and the overall size that they represent, a starting point that delivers some metadata about repositories is useful. 3lib uses the collection maintained by the Bielefeld Academic Search Engine (BASE, http://www.base-search.net/). This data are accessed by negotiation. At this point, AuthorClaim is the only service where the data gathered for BASE is used further. Currently, there are about 12 million items from BASE available.

There is another type-based collection that is freely available. It is the monograph data from the Open Library project (http://www.openlib.org/). It is an aggregate of monograph data from several sources and thus contains many apparent duplicates. It is freely available in bulk, though.

Institutional data

Compiling a large bibliographic dataset is the first challenge of for author claiming. The second challenge is to collect institutional data. This data are used to allow authors to state what institution they work for. If this would be entered as string data by registrants, aggregation of the data would be difficult to perform.

Each record is simple: it contains and identifier, an official name and then some name variation that allows the institutions to be found under alternative names. Some of these alternative names may, in fact, be abbreviations. The use of many names is convenient for registrants who search for their institution. Or the name can be expressed in several languages, or there may be a popular name differing from the official Thus, registrant may enter "MIT" to get to the Massachusett Institute of Technology.

The non-trivial problem lies in the recognition of an institution as being separate. For example, a university may have several campuses. Is each campus a separate institution or will they all be a part of the university? A slight variant of this issue is the desired level of granualatity. An other problem is the level of the institution. Say some folks work at the Foo Library of the University of Bar. They may search for foo. They find no match.

To set up author identifcation, it is useful to have a set of institutional data. Such data is avaiblable though project called ARIW, at http://ariw.org. The principal criterion used by ARIW to consider that an istitution is separate is the domain name. If an institution mentions its name in the domain, it is considered to be separate. Thus the University of California at Berkeley is considered separate from the University of California. But the Harvard-Smithsonian Center for Astrophysics is part of Harvard University.

AuthorClaim

AuthorClaim is the earliest open-access interdisciplinary author claiming service. By open access we mean that the records created by registrants are available for further use. AuthorClaim is built on the 3lib and ARIW dataset.

At registration, name variations are queried in much the same way as they

are collected at RAS. The registrats can search for institution to associate with. If a registrants don't find an istitution, they are invited to propose a new institutional record, which is placed as string data into the personal profile. But it can happen that users propose a new institution despite the fact that it is already registered in ARIW. This implies that while registrants can propose new institutional records, they cannot automatically be included into ARIW and they have to be manually inspected. If they muster the test for an independent and not yet included institution, the record can be added to ARIW. When AuthorClaim has updated ARIW data, the AuthorClaim administration can proceed to implement an official record. This implies removing a string record that a user has entered describing the institutions, and replacing it with the official record.[4]

The use of 3lib brings an important challenge. The number of homonyms is vastly amplified when the literature of several fields are joined, in particular when some dataset may capture only the initials of authors' first names. Just the raw numbers or records illustrate the challenge There are close to 60 Million bibliographic records, vs somewhat over one million that RAS has to deal with. One unpleasant consequence is that registered authors with common names would be overwhelmed by the number of potential matches to their works that the system would suggest to them for verification.

To alleviate this, AuthorClaim tries to distinguish an author from other authors with matching name variations. At the time when a registrant accepts certain documents and rejects others, a statistical learning engine works in the background. It is based on the popular libSVM (Chen and Lin, 2011) library that implements support vector machines. It sorts the remaining undecided items by the likelihood that they are close to the accepted items. The features that learning algorithm can use are co-author names and words in the title. Clearly, this algorithm does not perform well with the first proposed records since it has no . However, once there are many rejected documents, the algorithm performs quite well.

In principle, making data available to AuthorClaim should not prevent its commercial exploitation in other services. There are two reasons for this. First, from any bibliographic record, AuthorClaim needs only the title of the work, the list of author name expressions and a link to the site of the provider of the data documenting that item. It excludes, at this time, where it was published, and other information one would associate with a bibliographic record. That information can be gathered from the original bibliographic information, if one has access to that through a link or otherwise. If that information is not public then it is difficult for AuthorClaim to include it. Likewise, abstracts, classification data and full-text files are not used at all.

Second, AuthorClaim only allows an author name search to be performed, usually over a small number of search terms known as the registrant name variations. Other searches, such as by date, collection title, etc. are not supported.

[4]The same problem arises at the RePEc Author Service, where this can be managed without the use of too much human time thanks to the limited scope of the project. That would be quite different in AuthorClaim.

This additional information is not necessary to help authors determine in claiming what they wrote or not. Nevertheless, it is important for any service that would be using AuthorClaim data to display information and build links to publications by publishers.

AuthorClaim is an open system. The profiles are instantly available in bulk from the ftp.authorclaim.org server. They come with an explicit creative commons CC0 license. This means, essentially, that they are in the public domain. The profiles contain the basic metadata—title, author names, URL and handle—for each document that the registrant has accepted and for all that s/he has rejected. The documents rejected by authors are included because they allow users of profiles to build learning models for documents they don't know the status of. This could be used in research for training purposes to find what SVM parameters are best at guessing correct authorship. The inclusion of rejected papers makes the profiles rather large. It is probably best to look at a sample profile. Here is an extract of ftp://ftp.authorclaim.org/l/e/ple3.amf.xml

```
<amf xsi:schemaLocation="http://amf.openlib.org
   http://amf.openlib.org/2001/amf.xsd">
<person id="info:lib/am:1979-05-27:ralph_reese_levan">
 <name>Ralph Reese LeVan</name>
<givenname>Ralph</givenname>
<additionalname>Reese</additionalname>
<familyname>LeVan</familyname>
<homepage>http://staff.oclc.org/~levan</homepage>
<acis:shortid>ple3</acis:shortid>
<acis:last-change-date>2011-03-30 18:11:04 +0200</acis:last-change-date>
<ispartof><organization ref="info:lib/we:gbxho">
<name>OCLC Online Computer Library Center, Inc.</name>
<homepage>http://www.oclc.org/</homepage></organization>
</ispartof>
<acis:names>
<acis:variation>ralph reese levan</acis:variation>
<acis:variation>levan ralph reese</acis:variation>
<acis:variation>ralph r levan</acis:variation>
<acis:variation>levan ralph r</acis:variation>
<acis:variation>levan r r</acis:variation>
<acis:variation>r r levan</acis:variation>
<acis:variation>ralph levan</acis:variation>
<acis:variation>levan ralph</acis:variation>
<acis:variation>r levan</acis:variation>
<acis:variation>levan r</acis:variation></acis:names>
<isauthorof>
<text ref="info:lib/crossref:10.1300/J111v34n03_09">
<title>Searching Digital Libraries</title>
<hasauthor>
<person><name>Ralph Levan</name>
```

```
</person>
</hasauthor>
<displaypage>http://www.informaworld.com/ ... </displaypage>
</text>
</isauthorof>

   ...

<acis:hasnoconnectionto>
<text ref="info:lib/pubmed:11475515">
<title>Declining Medi-Cal coverage leads to increasing uninsured rate
  among California's Asian Americans and Pacific Islanders.</title>
<hasauthor><person>
<name>R. Levan</name>
</person>
</hasauthor><hasauthor>
<person>
<name>M. Kagawa-Singer</name>
</person>
</hasauthor>
<hasauthor>
<person><name>R. Wyn</name>
</person>
</hasauthor>
<displaypage>http://www.ncbi.nlm.nih.gov/pubmed/11475515</displaypage>
</text>
</acis:hasnoconnectionto>
</person>
</amf>
```

Given the problems in assembling a comprehensive dataset across disciplines, and the problems for authors with common name variations to manage large portfolios of rejected documents, another approach may be preferable. In that approach, authors would claim in local systems in which they have published documents with. One example of this would be the existing RePEc Author Service. The advantage is that authors have to wade through fewer rejections. The disadvantage is that cross-disciplinary researchers cannot build an accurate portfolio of their works.

The entire AuthorClaim system has been constructed in such a fashion that it is quite useless in isolation. To really populate the system one needs bibliographic services that use the AuthorClaim data. When RePEc introduced the RePEc Author Service, this was not an issue. Bibliographic services existed and immediately included the data resulting from this new member of the RePEc services family. That means that the bibliographic services immediately picked up author profiles.

In particular, RePEc's evaluation services limit the evaluation of every person to the documents claimed by a person. Within the interdisciplinary set-

ting that AuthorClaim operates, evaluative data are rarely built. So far, each provider has only few documents claimed by AuthorClaim registrants. In such a situation, bibliographic providers have little incentives to use AuthorClaim data. And when the AuthorClaim data are not widely used authors do not have incentives to register and the number of registered authors remains low.

Of course the use of AuthorClaim profiles in bibliographic databases would benefit database providers. Simple usage scenarios are not difficult to implement. Some can run in an automated fashion without need for continued manual maintenance. One is a simple list of publications by an individual listed the service, with links to bibliographic items described remotely. Such data are already available in AuthorClaim profiles. It would be particularly welcome for institutional repositories to provide a comprehensive listing of an author's achievements, including papers not in the local repository but in others. The dearth of inbound links is a reason why papers in repositories don't enjoy good visibility.

5 Competing products

There are non-free author identification products that compete with the author identification service of RePEc and 3lib, the RePEc Author Service and AuthorClaim.

The first interdisciplinary author registration service was researcherID. This is an effort by the Institute for Scientific Information (now part of Thomson Reuters) to compile author identification for the citations datasets that they have been compiling under the trade names "Science Citation Index" etc., now known as the Web of Science. This is a closed system as it is not available for further dissemination. Authors are free to register, though, and they receive individually some information about their citation performance.

Later, for reasons we are not aware of, Thomson Reuters became involved the ORCID initiative that started in 2009. ORCID stands for the Open Researcher and Contributor ID. The name is misleading in the sense that the dataset will not be open. It will essentially be accessible through paid membership. While ORCID is committed to make a CC0 licensed dump of user contributed data available once per year, we do not think this should be considered to be an open system. The web site says[5] "ORCID aims to solve the author/contributor name ambiguity problem in scholarly communications by creating a central registry of unique identifiers for individual researchers and an open and transparent linking mechanism between ORCID and other current author ID schemes."

This statement was established only after many months of debate on what author identification really means. It is certainly an issue that is wider than author claiming. However author claiming is part of the model. However, ORCID thinks that self-claiming alone would not get the organization off the ground with a sufficiently large base of users. Therefore they also allow member

[5] http://about.orcid.org/, accessed 29 May 2012.

organizations to submit data on behalf of people affiliated with the institution. However the registration only becomes official if claimed by an author.

ORCID's database is the CrossRef DOI dataset. This dataset is not open. ORCID's data itself is not open. It is essentially only available to ORCID members that have to pay a $5000 membership fee. The current set of members is dominated by publishers. This coincides with the membership of CrossRef. In a way the ORCID initiative seems to further cement the oligopoly of publishers over scholarly communcation.

SSRN (http://www.ssrn.com/) is an important online publisher that spans most of social sciences and some humanities. While it has author profiles, they are not based on registrations. Many authors have several profiles, especially if they have had several affiliations during their career. It is thus not particularly useful in this respect. The metadata from SSRN is not public.

ResearchGate (http://www.researchgate.org/) and Academia.edu are two sites that can also potentially provide author identification services. Their focus, however, is more on social networks with functionalities like tracking of authors, matching with scientists with similar interests, and other notification services. In both cases, the metadata remains private as well.

Finally, and probably most importantly, Google has introduced a citation claiming process in Google Scholar. It is an author claiming system in anything but name. The benefit of the Google Scholar is that it will show registrants what papers have cited theirs and calculate metrics such as the popular H-index. The problem with the system is the weak bibliographic control. The system is a search engine; it has no concept of manually maintained records. This means that individual papers are difficult to track. For example, a summary page listing papers in a journal can be considered a paper in itself. It then receives a lot of citations. This pushes up the H-index in a rather meaningless way.[6] Still, authors will be happy. But again, the data is not available publicly.

6 Conclusion

We have shown that an author identification service can be provided at remarkably low cost. As the example of RePEc shows, such a service can be successful if incentives are aligned so that both authors and publishers want to participate. This happens if there is sufficient critical mass and something useful is done with the collected data.

Providing such a service on a larger scale is feasible, with a few modifications to prevent authors from being overwhelmed by the quantity of works matching their names across scientific fields. The difficulty is rather to find appropriate material to feed into the project. There are vast areas of academic output that are not systematically indexed. It is disappointed to note that among the many calls for open access to academic papers, the issue of secondary data about papers is overlooked.

[6] For example, Google Scholar considered until 2013 RePEc and ArXiv to be among the top publications in terms of H-index, while they are clearly not publications.

7 References

José Manuel Barrueco Cruz and Thomas Krichel, 2000. "Cataloging Economics preprints: an introduction to the RePEc project," *Journal of Internet Cataloging*, vol. 3(3), pages 227–241.

José Manuel Barrueco Cruz and Thomas Krichel, 2005. "Building an Autonomous Citation Index for Grey Literature: RePEc, The Economics Working Papers Case," *The Grey Journal: An International Journal on Grey Literature*, vol. 1(2), pages 91–97.

Chih-Chung Chang and Chih-Jen Lin, 2011 "LIBSVM : a library for support vector machines." *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pages 27:1–27:27.

Heting Chu and Thomas Krichel, 2003. "NEP Current Awareness Service of the RePEc Digital Library," *Digital Libraries Magazine*, vol. 9(12).

Tom Coup, 2003. "Revealed Performances: Worldwide Rankings of Economists and Economics Departments, 1990-2000," *Journal of the European Economic Association*, vol. 1(6), pages 1309–1345.

Thomas Krichel, 1997a. "Guildford Protocol," available at http://openlib.org/acmes/root/docu/guilp.html

Thomas Krichel, 1997b. "ReDIF Version 1," available at http://openlib.org/acmes/root/docu/redif_1.html

Thomas Krichel, 1997c, "About NetEc, with special reference to WoPEc", *CHEER*, vol. 11(1), pages 19–24.

Thomas Krichel and Christian Zimmermann, 2009. "The Economics of Open Bibliographic Data Provision," *Economic Analysis and Policy*, vol. 39(1), pages 143–152.

Christian Zimmermann, 1997. "Step-by-step instructions for the creation of a RePEc archive," available at http://ideas.repec.org/stepbystep.html