Personal data in a large digital library[*]

Accepted at ECDL2000

1 July 2000

José Manuel Barrueco Cruz
Biblioteca de Ciències Socials
Universitat de València
46071 València
Spain
jose.barrueco@uv.es
http://www.uv.es/~barrueco

Markus J.R. Klink
Equant Application Services
Weyside Park
Godalming GU7 1XE
United Kingdom
markus.klink@gmx.net

Thomas Krichel
Department of Economics
University of Surrey
Guildford GU2 7XH
United Kingdom
T.Krichel@surrey.ac.uk
http://openlib.org/home/krichel
RePEc:per:1965-06-05:THOMAS_KRICHEL

Abstract

The RePEc Economics library offers the largest distributed source of freely downloadable scientific research reports in the world. RePEc also contains details about Economics institutions, publication outlets and people working in the field. All this data forms a large relational dataset.

In this paper we describe HoPEc, a system that allows to implement access control records for personal data within RePEc. The bulk of these data describe the authors of documents. These data are maintained by the authors themselves. We discuss the technical and social aspects of this system.

This paper is available online at http://openlib.org/home/krichel/phoenix.html.

[*] The work discussed here has received financial support by the Joint Information Systems Committee of the UK Higher Education Funding Councils through its Electronic Library Programme. Markus J.R. Klink is especially grateful for the support and helpful comments of his supervisor Stefan Kuhlins of Mannheim University. Thomas Krichel thanks the Digital World Research Centre at the University of Surrey for sponsoring his attendance at ECDL2000. All authors are grateful to an anonymous referee who has made excellent suggestions that helped to improve this paper.

1 Introduction

The identification of authors is a problem that has a long and distinguished history in Library and Information Science. There are two basic problems with using author names to identify authors. The first is that the same physical person may be referred to through different names. The same person's name may have different varieties, for example

- Clinton, William Jefferson Blythe III
- Clinton, William J.
- Bill Clinton

Some people may even change their name during their lifetime. For non-European names the translation into a suitable character set may add a further layer of difficulty. Thus even if the name is the same, it may appear that it has been transcribed to the Latin alphabet in different ways.

A second problem is that two physically different people may have the same name, or at least, that some spelling varieties of the name are the same. In the Economics research area, there are two Michael Devereux. One is Michael B. Devereux, the other one is Michael P. Devereux. Any occurrence of the term "Michael Devereux" or "Devereux, M." would be ambiguous between the two persons. This is not a difficult problem if we are aware that there are these homonyms in the library. However, it is very difficult to establish if all occurrences of a fixed name string point to the same physical person, i.e. to show the absence of a homonym. Even if there is a small number of person names in the dataset, an appearance of a name string does not identify a person.

To summarize, there are two problems to achieve a one-to-one mapping between name and physical person. This paper is about addressing both problems simultaneously within the context of a large digital library of academic documents. Before we discuss our work, it is useful to have a brief look at traditional approaches to the problem and set out our approach in Section 2. The collection that we are using is the largest distributed source of freely downloadable research papers in the world, the RePEc project. "RePEc" originally stood for "Research Papers in Economics". However the term should be understood as a literal, because the mission of RePEc goes well beyond the description of documents. RePEc builds a public-access documentation of Economics research. We describe this project in Section 3. General questions for the management of personal data are addressed in Section 4. In Section 5, we describe the HoPEc project at http://netec.mcc.ac.uk/HoPEc/ that implements the registration and search service. Our experience with running this service is the subject of Section 6. Section 7 concludes this paper.

2 Access control in a digital library

In conventional catalogues, attention is paid to harmonising the name for all spelling varieties who—to the best knowledge of the cataloger—are the same person. This process is one of the prime functions of authority control.

In the library tradition the authority control for author names is the process of choosing between the different variants of an author name that exist or can be used, a single one that will be used by the library. An elaborate set of rules ensures that the official name is the same for any library. Links from the unofficial variants to the official variant will be created.

This tradition comes from running a card catalogue. Here users looking for "Devereux, Michael" would be advised to consult the entries for "Devereux, Michael B." and "Devereux, Michael P.". Hoffman and Hatch (2000) is a good source of recent contributions to traditional authority control.

With the advent of electronic catalogues, the approach to author identification has shifted away from authority control towards access control. With access control, there is no official version of the author name that all holdings in the collection would use. Instead, all variants of a person's name are linked to one author record. In that author record all relevant data of the author is collected. A recent implementation of the access control is Snyman and Jansen van Rensburg (1999). They propose an "International Standard Author Number" (ISAN) that would identify each author. The ISAN would be awarded and maintained by national bibliographic authorities. Each national authority would share the data with the others within an international cooperative framework.

In computer science terms, the access control model implements a relation between documents and their authors. From a technical point of view, this is a rather trivial innovation over the authority control model. The problem with access control is the organisational structure that supports the creation and maintenance of the author database. It appears difficult to imagine that this problem could be completely solved without an international database of all living people. There are political, legal, and economic obstacles that imply that such a database is not forthcoming in the foreseeable future. Even a collaborative international database about important authors would require resources that are beyond of what national bibliographic authorities—who are considerably stretched by the appearance of digital resources—could afford.

The technological change from physical libraries to digital libraries is the main force driving the change from authority control to access control. In a digital library, the implementation of access control demands a similar attention than the authority control in the library that consists of physical objects. If the digital library is large, then it is very costly to implement access control since this process can not be fully automated. We are not aware of a digital library that has complete access control for its author data. Thus the digital library revolution does not help to solve the author identification problem. If anything, the availability of resources on the Internet makes the author identification problem more pressing. It is part of the mismatch between the richness of resources on the Internet and the paucity of metadata about these resources. This mismatch is a problem that has received a lot of attention.

While a global solution of author identification is not forthcoming, there is nothing to prevent a group of people from implementing a local author identification within a geographical or subject domain. The smaller the local author base, the least likely is the appearance of the name-identity problem. Thus resolving the problem is a less pressing issue in a local setting. However the collected personal information may be used to other ends. For example the access control records may contain contact information like email addresses, homepage URLs and telephone numbers. These additional data are typically relevant when authors and readers form a community where the same people have both author and reader rôles and where an exchange between both groups is an ongoing concern. In such a setting it is useful to involve non-librarians in the access control process. This papers describes an attempt—the first to our knowledge—to involve the authors directly in the creation of access control data. We take a set of digital library data, and we ask authors to tell us which papers they have written.

Naturally, this strategy to get the authors involved in producing their own access control data will only work if the authors have incentives to supply such data. Since academic authors are interested in the visibility of their work, they will have good incentives to supply data to a database that is frequently consulted by potential readers. For any database to achieve such a status, it must be relatively large and available at low cost. The RePEc dataset of Economics research is a good candidate.

3 The RePEc system

3.1 A scholarly dissemination framework

RePEc is a system to improve scholarly communication in Economics using the Internet. Scholarly communication has two functions. It disseminates scholarly output and it adds a quality certificate through peer review. In Economics, peer-review is particularly severe. As a result, the delay of formal publication—the attribution of the quality certificate—is very long. According to Trivedi (1993), it is common that a paper takes over three years from submission to publication in the same academic journal, not counting rejections. Thus researchers can not rely on the formally approved work alone because this material is out of date. As a consequence there is a flourishing culture of informal publication. Clearly the exchange of such publications may take advantage from the Internet. This is the initial motivation of the RePEc project. It is a digital library that disseminates Economics research.

A scholarly dissemination system on the Internet should start by enhancing the pre-Internet practice rather than attempting to replace it. The distribution of informal research papers in the past has been based on institutions issuing working papers. These are circulated through exchange arrangements. RePEc is a way to organise this process on the Internet. Its business model can be summarized as follows

$$\text{Many archives} \Longrightarrow \text{One dataset} \Longrightarrow \text{Many services}$$

RePEc allows Economics departments and research institutes to participate in a decentralized archival scheme which makes information about the documents that they publish accessible via the Internet. A contributor places metadata about its documents on a public access computer system. This is usually an anonymous ftp server or a web server. Each participating institution has total control over the contents of its archive. The archive management retains the liberty to post revisions or to withdraw a document. There is no need to transmit documents elsewhere. Participation does not imply that the documents are freely available. Thus commercial publishers can contribute. If a document is available online, a link may be provided to the place where the paper may be downloaded. Note that the document may not only be the full text of an academic paper, but it may also be an ancillary file, e.g. a dataset or a computer programmes.

In April 2000, about 130 archives in 21 countries participate in RePEc, some of them representing several institutions. Over 80 universities contribute their working papers. Some important non-academic institutions like the US Federal Reserve, the IMF, the World Bank and the OECD are also present. A number of scholarly journals also contribute data to RePEc. There are over 70,000 resources described. About 20,000 are accessible on the Internet without access restrictions.

Users access RePEc data through user services. Appendix A of Krichel (2000) lists a range of user services. Note that the RePEc data may not be sold or incorporated into a product that is sold. User services compete through quality rather than price. All RePEc archives benefit from simultaneous inclusion in all services. This leads to an efficient dissemination that a proprietary system can not afford.

3.2 A relational metadata set

The contributors of bibliographical data supply it in a special format called ReDIF. ReDIF is a template format inspired by Deutsch, Emtage, Koster, and Stumpf (1994) also known as the IAFA template. To understand the basics of ReDIF it is best to start with an example. Here is a—carefully selected—piece of ReDIF from ftp://www.econ.surrey.ac.uk/pub/RePEc/sur/surrec/surrec9601.rdf.[1]

```
Template-Type: ReDIF-Paper 1.0
Title: Dynamic Aspect of Growth and Fiscal
 Policy
Author-Name: Thomas Krichel
Author-Person:
 RePEc:per:1965-06-05:thomas_krichel
Author-Email: T.Krichel@surrey.ac.uk
Author-Name: Paul Levine
Author-Email: P.Levine@surrey.ac.uk
Author-WorkPlace-Name: University of Surrey
Classification-JEL: C61; E21; E23; E62; O41
File-URL: ftp://www.econ.surrey.ac.uk/pub/
 RePEc/sur/surrec/surrec9601.pdf
File-Format: application/pdf
Creation-Date: 199603
Revision-Date: 199711
Handle: RePEc:sur:surrec:9601
```

When we look at this record, the ReDIF data appears like a standard bibliographical format, with authors, title etc. The only thing that appears a bit mysterious here is the Author-Person field. This is a legal field but it is as yet very sparingly used. The field quotes a handle that is known to RePEc. This handle leads to a record at ftp://netec.mcc.ac.uk/pub/RePEc/per/pers/RePEc_per_1965-06-05_THOMAS_KRICHEL.rdf [2]

```
Template-Type: ReDIF-Person 1.0
```

[1] We suppress the abstract to conserve space.

[2] We leave out a few fields to conserve space.

```
Name-Full: KRICHEL, THOMAS
Name-First: THOMAS
Name-Last: KRICHEL
Postal: 1 Martyr Court
 10 Martyr Road
 Guildford GU1 4LF
 England
Email: t.krichel@surrey.ac.uk
Homepage: http://openlib.org/home/krichel
Workplace-Institution: RePEc:edi:desuruk
Author-Paper: RePEc:sur:surrec:9801
Author-Paper: RePEc:sur:surrec:9702
Author-Paper: RePEc:sur:surrec:9601
Author-Paper: RePEc:rpc:rdfdoc:concepts
Author-Paper: RePEc:rpc:rdfdoc:ReDIF
Handle: RePEc:per:1965-06-05:THOMAS_KRICHEL
```

This record is the access control record for the author "Thomas Krichel". We will discuss this record in detail in the next section. For now, note that in the person template, we find another RePEc identifier in the "Workplace-Institution" field. This points to another record at ftp://crefe.dse.uqam.ca/pub/RePEc/edi/inst/desuruk.rdf that describes the institution. The maintenance of these records is the work of the EDIRC project. The acronym stands for "Economics Departments, Institutions and Research Centers". This dataset has been compiled by Christian Zimmermann, an Associate Professor of Economics at Unversité du Québec à Montréal on his own account, as a public service to the Economics profession. The initial intention was to compile a directory with all Economics departments that have a web presence. Since there are many departments that have a web presence now, a large number are now registered, about 5,000 of them at the time of writing. All these records are included in RePEc.

It is clear that when it comes to the collection of personal data we can not follow the approach of EDIRC, i.e. a single person collecting data "off the web".

4 Organising personal data

4.1 The relational information structure

The RePEc model implies that the flow of information always travels bottom up. The contributing archive is the authoritative source of information. User services have read-only access to the data as provided by the archive. Within that framework, one possible approach would be to ask archives to register people who work at their institution. This will make archive maintainers' work more onerous initially, but the overall maintenance effort will be smaller once all authors are registered. However, authors move between archives. Many have work that appears in different archives. To date there is no satisfactory way to deal with moving authors. Therefore the author registration is carried out using a centralized system. The first step to provide personal data is to open a RePEc archive that houses such data.

Before the creation of that central archive, all personal information within RePEc has been composed into the paper templates. This compositional model implies that the author data is part of the resource data. It does not form an independent entity. If the paper disappears, the author data disappears as well. If the author writes two papers, the author data is keyed in twice and the data attached to the author may be not be the same for both papers. We refer to this data as the composed personal data. This compositional logic comes from the traditional library catalogue model. What we wish to achieve is a movement from a compositional representation of personal data to a relational—as opposed to a compositional—representation. To create a fully relational model we need to collect some basic personal data (name, email address etc.) in a personal record and then we refer in the document templates to that personal data. This is done in the record for the paper RePEc:sur:surrec:9601, but to date this is one of very few records that contain such an entry. Organisational features are to blame for the lack of usage of person identifiers in document records. Recall that we have rejected the model where archive maintainers maintain personal data for "local" authors. Having made that choice, we rely on archive maintainers to convince local authors to register with the centralized personal registration service. It is only when this registration is achieved that the archive maintainer may quote the resulting handle in the paper templates. This change of working practice will take time to implement. At the time of writing, we have not even suggested to archive maintainer to implement this change.

The transition from the compositional to the relational model does not rely on the "Author-Person" tag in the document template. Instead, we use on "Author-Paper" field as demonstrated in the RePEc:per:1965-06-05:THOMAS_KRICHEL record. An important advantage of this approach is that the registration service can ensure that are only valid document handles are used. The disadvantage is that the relational structure is not as complete as one would like it to be. The "Author-Paper" entry tells us that the registered person is an author of the paper. But in the case where the paper has multiple authors we do not know which of the co-authors of the paper is described in the person template. This is an important conceptual limitation of the proposed model. However it has only limited practical implications since a comparison of name strings should allow to find which of the co-authors is the one the personal record refers to. Heaven forbid the two Michael Devereux co-authoring a paper.

4.2 Handle structure

The registration process associates a unique handle to each registered person. Conformity with RePEc tradition and its template analysis software requires that the first two components—delimited by colon—should have constant length. These first two components are respectively "RePEc" and "per". Thus "per" is the code for the RePEc archive that was opened to store the personal data collected by the personal data archive. The personal data is available through the conventional file structure set out in the "Guildford protocol". This is the convention on how RePEc archives store files.

The handles of the personal data records that are created start with the archive handle where the record is maintained. There are many options for the contents of the remainder of the identifier. In general, to build a unique identifier for persons is a problem that has never been completely solved. In our case, we feel that a pragmatic solution is needed that is not too mnemonic and not too cryptic. The combination of the name and date of birth—already widely used in the library world—appears to be a good starting point. However, it should be noted that some registrants may not wish their birth date to be known publicly. Thus requiring the date of birth would have reduced the acceptance of the service. What we require instead is a date in lifetime of the registrant that the registrant would be able to remember. We will refer to that date as the "significant date" of the registrant.

4.3 Dealing with compositional personal data

The registration service also attaches an internal handle to all available personal data that is composed into the existing resource metadata as collected by RePEc. These handles can not have the same structure as the handles for registered persons, because the significant date is not known, and neither is the real identity of the person. The internal handle has three requirements. It should be unique, it should be possible to build it only from the resource metadata, and finally it should be stable. These requirements are satisfied as long as the internal handle combines data from both the handle of the resource and from the names of the authors. For the example of the template RePEc:sur:surrec:9601 the internal author handle RePEc: sur:surrec:9601:Thomas_Krichel is derived. This method will fail to produce unique handles only if there is a document authored by two or more persons with the same name, which is most likely the result of an data input error.

A further concern with the composed personal information is that author names are not normalised. To search for authors is more efficient if names are split into first names and last names. An important task for the code that implements HoPEc, is to try to normalise person names that are found in composite author name data. Any name field may—in contravention to the principles of ReDIF—contain several author names, say "Markus Klink and Thomas Krichel". First, potential titles like "Prof.", "PhD" etc. are eliminated. The resulting string is recursively decomposed until there are no further name separators. These separators are colons, semicolons, repetition of blanks and words that indicate a separation like "and". Each of these components is then examined to find out where the first name and last name are located. First, all first names are normalized (e.g. Klink M $\mapsto$ Klink M.), then elements that are in brackets are removed (e.g. José Manuel Barrueco Cruz (editor) $\mapsto$ José Manuel Barrueco Cruz). If there is a comma, it is assumed that the construct is *lastname, firstname*. Otherwise a structure *firstname lastname* is assumed. However, if the last name is much shorter than the first name, a structure *lastname initial* is assumed, and first and last name are assumed to be in the opposite order. The complete algorithm was developed under pragmatic assumptions to deal with the realities found in the dataset.

5 The HoPEc service

HoPEc started its life as a collection of links to economists' homepages that contained downloadable research papers, maintained by Barry Schachter. This collection started in 1995 as a spare time project, when Barry was a Financial Economist at the US Comptroller of the Currency. In 1997, the collection was integrated under the name HoPEc into the NetEc consortium of projects. At that time the name "HoPEc" stood for "Home Page Papers in Economics". Barry moved on to become Vice President of Chase Manhattan Bank, and HoPEc moved on to become the "Home of RePEc Person Registration". The HoPEc service allows registrants to maintain their personal RePEc data. It is more fully described in Klink (1999). It was opened in September 1999. It has two basic functions. The "search" function allows to search for personal data in the RePEc dataset. The "registration" function allows for persons to register.

5.1 The search function

The search function provides convenient access to personal information when only the name of the person is known. As a result of a search all the data relating to what the system presumes is the same person appears at the same place. This is very easy when a registered person is found. At the moment however, the search function mainly deals with the composed personal data. For these composed data, the task of the search is to find a take a set of name strings, and group those strings together that it believes are the same. Different aggregations of name strings will lead to different numbers of presumed persons.

	number
total number of persons	142,848
registered	544
with personal data	7,416
with workplace data	14,567
numbers of aggregations with common	
lastname	22,608
first & last name	44,292
first & last name & archive	54,092

Within the composed personal data, there are very few person that have any additional data either about themselves or the institution that they work for. By "additional" we mean any non-blank data field that is not the name, for example the email address or homepage URL. It is therefore clear that these additional data can not be used to identify people within the dataset, because of the limited amount of data available. In the best possible case, where all the persons with additional data would be identical, this would reduce the total number of persons only by about 15,000. Therefore we need to analyse the identity of persons with the same name structure. The following strategies can be envisaged

- When we look for a person's name, we are only concerned with the person's last name. This leads to a great aggregation of data.

- If differentiate further by using the first name—considering that an initial is different from the full name—a great aggregation of names is still possible.

- If the provision of the data through archives is taken into account, the number of aggregates does not increase by much. Thus we can conclude that there are only few persons who have publications in different archives.

Since we would like to use all information that we have available, we decide to use the third method. That implies that we consider that a person is identical to another if (s)he has the some name (first and last name) and if (s)he publishes in the same archive. This assumption would lead to an error if there are two persons with the same name who publish material in the same archive and who are physically different. In this very unlikely case, the maintainer of the archive, who has supplied RePEc with the data, could ensure to distinguish the two persons through attributing different spelling forms of the name. It should be noted that we overcome this problem through our author registration service.

The following table displays all the responses to an approximate query for the last name "Devereux" and first name "M".

last name	first name	archive
DABROWSKI	M.	fth
DABROWSKI	MAREK	wop
DE BROECK	MARK	imf
DEVEREAUX	MICHAEL	boc
DEVEREUX	M.	fth
DEVEREUX	M.	wop
DEVEREUX	M. B.	fth
DEVEREUX	M. P.	fth
DEVEREUX	MICHAEL	ifs
DEVEREUX	MICHAEL	nbr
DEVEREUX	MICHAEL	wuk
DEVEREUX	MICHAEL B.	nbr
DEVEREUX	MICHAEL P.	ifs
DEVEREUX	MICHAEL P.	wuk
DEVEREUX	MP.	wop
DVORAK	MICHAEL W.	fip

Here the search finds some names that are similar to "Devereux". For a subject specialist—who knows both authors and the areas they usually work in—an inspection of the record for Devereaux reveals that this paper is in fact by Michael P. Devereux. Note that he is also the "Michael Devereux" who publishes in the "wuk" and "ifs" archives, but not the one in the "nbr" archive. For a person from outside the discipline it would be difficult to aggregate these authors correctly. Therefore person registration provides a significant enhancement of the RePEc digital library.

5.2 The registration function

The registration process consists of two stages. In the first stage, the registrant is asked to supply personal information. In the second stage the registrant may create associations between the personal record and the resource records in the RePEc database. A person may only create associations to resources that are currently described in the RePEc dataset. It is not possible to create associations with potential future resources.

The current implementation of HoPEc considers associations with resources that appear in four different templates types. These are "ReDIF-paper", "ReDIF-article", "ReDIF-software" and "ReDIF-series". Persons can have different associations with these templates. Persons can be authors of papers, articles or software, and they can be the editors of a series. In the current model, the relationship between resource type and association type is injective. When the resource type is known, the association type can be deducted, but the opposite does not hold.

A person is identified by the combination of a name and a significant date. The email address of the registrant must also be known. These three elements are the minimum data elements that are required for registration. If two persons that are physically different were to enter the same data for all three fields—name, date and email—the system would consider that they are the same.

To make the association with the resources easier, the system will suggest a number of resources for the registrant to associate with. These are the documents that a person has associated with before—if any—as well as any other documents with the same or a similar last name and the same first name or corresponding initial. The selection of associations uses checkboxes. It is also possible to enter the handle of resources to create an association with in case that this association is not proposed.

Any change of the data has to be confirmed. The registrant receives a four-digit confirmation number by email. To complete the registration process, the registrant must confirm the registration using the confirmation number. This avoids the administration of passwords. We do not use passwords to avoid the administrative burden that users who have forgotten their password place on the maintenance of the system. If a person has forgotten a password, she might wish to register anew. In that case we would have two records for the same person. This is a situation that we wanted to avoid. We think that the combination of names, significant date and email address offers sufficient security.

5.3 Illustration

The functionality of HoPEc as described in Subsections 5.1 and 5.2 is illustrated in Figure 1. It illustrates that readers and authors of documents access the system to search or register. Searching is conducted across all archives. Search results from the document archives result in aggregated data according to our aggregation strategy. Search results from the personal data archive are especially highlighted to indicate that HoPEc found a registered author. Only registered authors—and are "members" of the person archive—are able to provide information about the working papers they wrote. This is illustrated by the connecting lines.

5.4 Metadata

In order to be able to share personal data between the different services which make use of the RePEc dataset, the personal data must be stored in the form of ReDIF templates. The format of these templates has already been discussed above. Additionally the registration service generates data output in two different formats. These are a native XML format and an embedded RDF (see Lassila and Swick (1999)) format, which is currently embedded as metadata in the generated HTML files.

The RDF itself does not define a metadata standard to describes resources. It aims are rather to allow to use a multitude of standards simultaneously within a uniform structure. We employ a subset of the metadata tags suggested by the Dublin Core Metadata Initiative (1999). Where necessary we add our own metadata tags. The abbreviated example below describes the resource at the URL http://netec.mcc.ac.uk/WoPEc/data/Papers/sursurrec9601.html.

```
<rdf:Description
about="http://netec.mcc.ac.uk/WoPEc/data/
 Papers/sursurrec9601.html" bagID="bag_0">
<DC:Title="Dynamic Aspect of Growth and
 Fiscal Policy">
<DC:Creator> <rdf:Bag
 rdf:_0="LEVINE, PAUL"
 rdf:_1 resource="./gemini.cgi?submit=id&
 HANDLE=RePEc:per:1965-06-05:THOMAS_KRICHEL"/>
</DC:Creator></rdf:Description>
<rdf:Description aboutEach="#bag_0"
 HOPEC:attributedto="Hopec Person
 Registration Project">
<rdf:type resource="http://www.w3.org/1999/
 02/22-rdf-syntax-ns#Statement" />
```

At the moment only title and author information about the paper are given. The paper has one title and two authors. Author information is given in the BAG construct. One author is unregistered (the literal "Levine, Paul") and one author is the registered author Thomas Krichel. The latter is not represented by a string literal, but by his own resource. The description of the bag serves an interesting purpose. We refine the statement of the authorship of this paper by giving responsibility information about the first statement (the author bag). In words: The HoPEc Registration Project states that Paul Levine and Thomas Krichel are the authors of the above mentioned resource. This process is known as reification.

6 Preliminary results on the usage of HoPEc

The HoPEc service opened for public registration in October 1999. Therefore at the time of writing, we can look over a seven-month period of work for the service. Clearly this service is only valuable if it is maintained indefinitely. Our main concern here is therefore to look at issues that affect the sustainability of the service.

There are several significant problems that a service like HoPEc faces. First since there is no historical precedent for such a service, it is not easy to communicate the raison d'être of the service to a potential registrant. Some people think that they need to register in order to use RePEc services. While this generates valuable information about who is interested in using RePEc services—or more precisely who is too dumb to grasp that these services do not require registration—it clutters the database with records of limited usefulness. Here is a rather striking example of a record that has been removed

```
Template-Type: ReDIF-Person 1.0
Name-Full: MARLEY, BOB
Name-First: BOB
Name-Last: MARLEY
Email: johniblaze@excite.com
Workplace-Name: mcdonalds
Workplace-Postal:  101 webb blvd
  new bern nc 28532
Workplace-Email: hotboy@hotmail.com
Workplace-Fax: 3344447554
Handle: RePEc:per:1980-10-16:BOB_MARLEY
```

We have taken steps to warn registrants that if they do not belong to our target group, registration is a waste of their time. In particular, we tell them that registration does not give them any priviledged access to RePEc services.

Having taken these steps, the problem of spurious registration does not appear to be important. From manual inspection, we find that in about one in eight of all registered persons, there is no evidence that the registered person belongs to the academic Economics research community. We think that probably less than ten percent of the records will turn out to be spurious. In the future we will look at records that have not been updated for three years, and remove them if they do not contain any links to documents. The person concerned is free to register again.

In Figure 2, we list the registrants by year of the significant date. It appears that there are some registrants who choose the date of registration as the significant date. While we have not prohibited this, there is a concern that some registrants may forget their significant date. For basic security reasons, the significant date is not directly visible on the search pages. If registrants wish to update their records and do not know the significant date, they will take either one of two actions.

- They will mail the HoPEc helpline to ask what the date is.
- They will register again and create a duplicate record.

Both of these will create some manual maintenance work. Therefore we are pleased to see that the large majority of registrants have chosen a date that looks to us like it is their birth date. If it is, then the median RePEc registrant is a junior researcher in her thirties. This confirms data that we have gathered informally from other sources.

From our experience the name/date registration system works well. However it should be noted that the problem of double registration is not completely solved with the registration procedure, and some manual effort will have to be done to control for double registration. We plan to add a warning screen when a user registers with the same name as a user who is already registered.

7 Conclusions

The Internet empowers those people who have access to it to create information architectures that are completely new. HoPEc presents such a radical innovation. Our demonstrated success in running HoPEc is ground for optimism that we will be able to build a sophisticated relational academic documentation that will be open for public access on the Internet.

To date, we have about 1,800 resources that have at least one registered author. RePEc user services use these data to group papers with the same author together. Other RePEc-based information services will greatly benefit from reliable author identification data. For example, a citation analysis service is planned that will gather citation data for registered authors.

References

Deutsch, Peter, Alan Emtage, Martijn Koster, and Markus Stumpf (1994). Publishing Information on the Internet with Anonymous FTP. Internet draft, expired March 1, 1995.

Dublin Core (1999). *Metadata Element Set Version 1.1*. available at http://purl.org/dc/documents/rec-dces-19990702.htm.

Hoffman, Shannon L. and Deborah Hatch (2000). Web World of Authority Control. available at http://www.lib.byu.edu/dept/catalog/authority/.

Klink, Markus J.R. (1999). Digitale Bibliotheken: Überblick über Metadatenstandards und praktische Anwendung im Bereich der Personenidentifikation. available at http://openlib.org/home/krichel/gemini.pdf.

Krichel, Thomas (2000). Working towards an Open Library for Economics: The RePEc project. presented at the "PEAK 2000 Conference: The Economics and Use of Digital Library Collections", available at http://openlib.org/home/krichel/papers/myers.html.

Lassila, Ora and Ralph R. Swick (Eds.) (1999). *Resource Description Framework (RDF) Model and Syntax Specification*. World Wide Web Consortium. available at http://www.w3.org/TR/REC-rdf-syntax/.

Snyman, M.M.M. and M. Jansen van Rensburg (1999). Reengineering name authority control. *Electronic Library 17(5)*, 307–311.

Trivedi, Pravin K. (1993). An analysis of publication delays in Econometrics. *Journal of Applied Econometrics 8(2)*, 93–100.

Figure 1: Illustration of HoPEc system

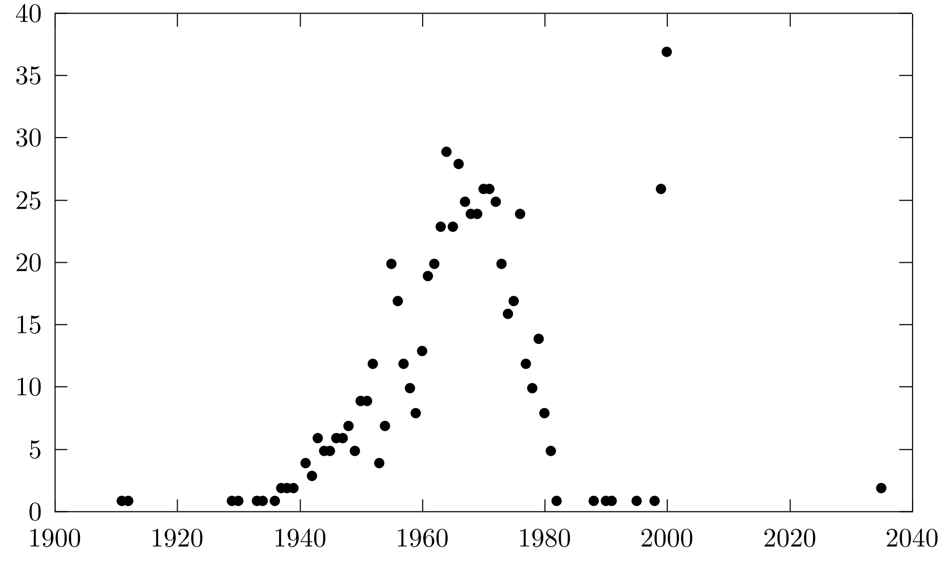

Figure 2: 670 indentifed by significant date used at registration