Cataloging Economics preprints: an introduction to the RePEc project*

José Manuel Barrueco Cruz and Thomas Krichel

December 11, 2003

Jos+latexcommand(\'{e}) Manuel Barrueco Cruz
Biblioteca de Ci+latexcommand(\‘{e})ncies Socials dq(Gregori Maians)
Universitat de Val+latexcommand(\‘{e})ncia
Campus dels Tarongers s/n
46071 València
Spain
jose.barrueco@uv.es
http://www.uv.es/~barrueco

Thomas Krichel
Department of
University of
Stag Hill
Guildford GU2 5XH
United Kingdom
T.Krichel@surr
http://openlib.org/hor
RePEc:per:1965–06–

Abstract

Cataloging scientific papers creates a new educational resource. Collecting that data is a costly process to achieve and manage. In particular the level of granularity that is required is finer than say for a collection of web sites. One possible approach towards cataloging these resources is to get a community of providers involved in cataloging the materials that they provide. This paper introduces RePEc of http://netec.wust.edu/RePEc, as an example for such an approach. RePEc is mainly a catalog of research papers in Economics. It is based on set of over 80 archives which all work independently but yet are interoperable. They together provide data about almost 60,000 preprints and over 10,000 published articles.

José Manuel Barrueco Cruz is a librarian at the Universitat de València. Thomas Krichel is a lecturer in Economics at the University of Surrey. Both welcome comments on this paper, write to wopec@netec.mcc.ac.uk.
This paper is available online at http://openlib.org/home/krichel/papers/shankari.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. We are grateful to Christopher F. Baum, Robert P. Parks, Thorsten Wichmann and Christian Zimmermann for comments on the questionnaires. William L. Goffe and Christian Zimmermann made many helpful suggestions. The topic of Subsection 5.2 was suggested by Jane Greenberg. Sophie C. Rigny kindly pointed out many stylistic and grammatical errors in an earlier version.

1 Introduction

Some scientific disciplines have a preprint tradition. Essentially these are Mathematics, Physics, Computer Science and Economics. Preprints are not issued in the same ways across those disciplines. In Mathematics and Physics preprints are essentially issued by individual academics. In Computer Science and Economics, it is more the department that distributes the preprints.

In this paper we deal with Economics preprints, usually called working papers. Economics is the dismal science. Its bad reputation is founded on two conceptions. The first is that economists never agree on anything. Winston Churchill claimed "If you put two economists in a room, you get two opinions, unless one of them is Lord Keynes, in which case you get three opinions". And on the other side of the pond, President Truman sought to hire a one-armed economist because he would no be able to say "on the other hand". The other conception is that Economics is very theoretical to the point of being totally useless. A popular tale is that of the two economists who sit down to play chess. They study the board for 24 hours and eventually declare a stale-mate.

Fortunately both of these conceptions do not fully apply to all sections of Economics. There is a large mainstream literature that is based on a common set of principles. It is true that this literature is heavily mathematical but that that does not follow that it is completely useless. There are counterexamples. For example the calculation of option values is important for anybody how is dealing with financial options. Trade in such options has only taken off since a pricing formula has been found. Another example are studies relating to competition. These are used by government organizations who work on regulating industries and on anti-trust measures.

Economics research documents are therefore useful to a wide variety of people, not only to students. In the past years more and more Economics departments and research institutions have made their working papers available on the Internet. However in that form the papers can only be found by specialists who know who has been working in a certain area, where that researcher is based and whether there are any papers of that researcher available on the web pages. This is the kind of knowledge that is circulated at scientific conference—usually on the back of business cards and an napkins—and therefore this data is not available to the people outside the research community. The normal mortals will only be able to benefit if a catalog of these papers is available.

In this paper, we describe attempts to build a catalog of online and offline working papers in Economics called RePEc. In Section 2 we introduce the concepts behind it. The RePEc is spread over many archives and these are described in Section 3. Section 4 describes the contents of the dataset. In Section 4 we review the RePEc dataset. Section 4 we consider user interfaces to RePEc. Section 6 concludes.

2 RePEc

The Electronic dissemination of Economics working papers can be traced back to the start of the Working Papers in Economics (`WoPEc`) project in April 1993. By May 1999 this single archive has grown into an interconnected network of over 80 archives holding over 14,000 downloadable working papers and over 50,000 descriptions of offline papers from close to 1,000 series. The network of archives is called `RePEc`. This term is was initially conceived to stand for "Research Papers in Economics". Nowadays it is best understood as a literal, because the objectives of RePEc go way beyond a database of scientific papers.

RePEc data is freely available, in the sense that the provider pays for the provision of the data, not the user. In order to make such a system viable without public subsidy, the cost of providing the data must be spread among many agents[1]. This requirement has been a feature of RePEc right from the start of the collection in May 1997. Each participating provider sets up an archive on a http or ftp server. The archive supports the storage of structural data about objects relevant to Economics, and possibly the storage of some of the objects themselves. All objects in RePEc are uniquely identified following by handles.

RePEc data can be accessed through a plethora of user services. Some are heavily used, for example the "`IDEAS`" user service had one million hits in just over 2 moths in 1999. The main interest of this paper is to examine the collection aspect of the data. The idea that a coherent literature catalog can be put together by a large group of people who are physically dispersed and have very little personal communication without the need of extensive training nor intensive coordination remains to be demonstrated. At the time of writing this paper RePEc is two years old. We feel that this is a good time to review the operations of RePEc and the data that it has collected. Clearly the RePEc data is in a constant state of flux. To keep matters simple

[1] understood here and in the rest of the paper as a person or institution

we took a dump of the data on 1 May 1999. In this paper we are only referring to the state of the data on that date.

There are some aspects of RePEc that this paper does not discuss. We eschew any mentioning of the data on software, books, etc to concentrate on the collection of traditional academic papers be they preprints or published articles. This data forms the bulk of the present collection. We also leave out the personal and institutional data which are is not included in the papers and article templates. We aim to use such data to build a fully relational database system that describes Economics as a discipline. We will report on such efforts in future papers.

The nature of RePEc is not precisely defined. Most people think about it as a collection of archives and services that provide data about Economics. More precisely, RePEc is most commonly understood as referring to three things. First it is a collection of archives that provide data about Economics. Second it is the data that is found on these archives. Third, it is often also understood to represent the set of agents who build archives and channel the data from the archives to the users. In that latter sense RePEc has no formal management structure.

RePEc has two aims. The "cataloging aim" is to provide a complete description of the Economics discipline that is available on the Internet. The "publishing aim" is to provide *free* access to Economics resources on the Internet.

The basic principle of RePEc can be summarized as follows

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

Basic RePEc concepts are: archive, site and service.

- An "archive" is a space on a public access computer system which makes data available. It is a place where original data enters the system. The is no need to run any software other than an ftp or http daemon that makes the files in the archive available upon request. Each archive is identified by a three-letter code. Some elementary metadata about the archive like its name, its url and some basic contents information are polled by a special central archive with the handle RePEc:all, where "RePEc" is the naming authority and "all" is the archive code.

- A "site" is a collection of archives on the same computer system. It usually consists of a local archive augmented by frequently updated ("mirrored") copies of remote archives.

- A "service" is a rendering of RePEc data in a form that is available to the end user.

All archives hold papers and metadata about papers, as well as software that is useful to maintain archives. Everything contained in an archive may be mirrored. For example, if the full text of a paper is in the archive, it may be mirrored. If the archive does not wish the full text to be mirrored, it can store the papers outside the archive. The advantage of this "remote storage" is that the archive maintainer will get a complete set of access logs to the file. The disadvantage is that every request for the file will have to be served from the local archive rather than from the RePEc site that the user is accessing. Of course an archive may also contain data about documents that are exclusively available in print.

There is no need for every site to mirror the complete contents of every archive in the system. To conserve disk space and bandwidth some sites only mirror bibliographic information rather than the documents that an archive may contain. Others mirror all the files of an archive. Others may mirror only parts of a few archives. The software that is used to mirror the archive is provided at RePEc:all. It first mirrors the central archive. This software then reads a configuration file and then writes batch calls to the popular "`mirror`" program for ftp and the "`w3mir`" script for http archives.

An obvious way to organize the mirroring process would be to mirror the data of all archives to a central location. This central location would in turn be mirrored to the other RePEc sites. The founders of RePEc did not adopt that solution, because it would be quite vulnerable to mistakes at the central site. Instead each site installs the mirroring software and mirrors "on its own", so to speak. Not all of them adopt the same frequency of updating. Many update every night, but a minority only updates every week. It is therefore not known how long it takes for a new item to be propagated through the system.

Each service has its own name. A service that is based on mirrored scripts may run on many locations. Within reason, all services are free to use any part of the RePEc data as they see fit. For example a service may only show papers that are available electronically, others may restrict the choice further to act as quality filters. In this way services implement constraints on the data, whether they be availability constraints or

quality constraints. The user service infrastructure is quite well developed, we list the most important ones in Section 5. This distribution via the several user services is undisputedly successful feature of RePEc. It is therefore not given further attention here.

3 The structure of an archive

RePEc stands on two pillars. First, an *attribute*:*value* template metadata format called ReDIF. This acronym stands for *Re*search *D*ocumentation *I*nformation *F*ormat but it is best understood as a literal. ReDIF defines a number of templates. Each templates describes an object in RePEc. It has a set of allowable fields, mandatory, and some repeatable. The second pillar is the Guildford protocol. It fixes rules how to store ReDIF in an archive. It basically indicates which files may contain which templates. It is possible to deploy ReDIF without using the Guildford protocol. But in the following we will ignore this conceptual distinction, because it is easiest to understand the structure and contents of an archive through an example. This is done in Subsection 3.1. Therefore we will list files in the way required by the protocol as well as the contents of the file that is in fact written in ReDIF. This is done in Subsection 3.1. We return to technical aspects of ReDIF in Subsection 3.2.

3.1 The Guildford Protocol

RePEc identifies each archive by a simple identifier or handle. Here we look at the archive RePEc:sur which lives at ftp://www.econ.surrey.ac.uk/pub/RePEc/sur. On the root directory of the archive, there are two mandatory files. The file *surarch.rdf* contains a single ReDIF archive template.

```
Template-type: ReDIF-Archive 1.0
Name: University of Surrey Economics Department
Maintainer-Email: T.Krichel@surrey.ac.uk
Description: This archive provides research papers from the
 Department of Economics of the University of Surrey,
 in the U.K.
URL: ftp://www.econ.surrey.ac.uk/pub/RePEc/sur
Homepage: http://www.econ.surrey.ac.uk
Handle: RePEc:sur
```

In this file we find basic information about the archive. The other mandatory file is *surseri.rdf*. It must contain one or more series templates.

```
Template-Type: ReDIF-Series 1.0
Name:  Surrey Economics Online Papers
Publisher-Name: University of Surrey, Department of Economics
Publisher-Homepage: http://www.econ.surrey.ac.uk
Maintainer-Name: Thomas Krichel
Maintainer-Email: T.Krichel@surrey.ac.uk
Handle: RePEc:sur:surrec
```

These two files are the only mandatory files in the Guildford protocol. If these are the only files present in the archive then all the archive is doing is to reserve the archive and the series codes. All documents have to be in a series. The papers for the series RePEc:sur:surrec are confined to a directory called *surrec*. It may contain files of any type. Any file ending in ".rdf" is considered to contain ReDIF templates. Let us consider one of them, *surrec*/*surrec9601.rdf*[2]

```
Template-Type: ReDIF-Paper 1.0
Title: Dynamic Aspect of Growth and Fiscal Policy
Author-Name: 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
```

[2] We suppress the Abstract: field to conserve space.

```
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
```

Note that we have two authors here. The "Author-WorkPlace-Name" attribute only applies to the second author. We will come discuss this point now.

3.2 The ReDIF metadata

The ReDIF metadata is mainly an extension of the **?**) , commonly known as the IAFA templates. In particular it borrows the idea of clusters from the draft

> There are certain classes of data elements, such as contact information, which occur every time an individual, group or organization needs to be described. Such data as names, telephone numbers, postal and email addresses etc. fall into this category. To avoid repeating these common elements explicitly in every template below, we define "clusters" which can then be referred to in a shorthand manner in the actual template definitions.

ReDIF takes a slightly different approach to clusters. A cluster is a group of fields that jointly describe a repeatable attribute of the resource. This is best understood by an example. A paper may have several authors. For each author we may have several fields that we are interested in, the name, email address, homepage etc. If we have several authors then we have several such groups of attributes. In addition each author may be affiliated with several institutions. Here each institution may be described by several attributes for its name, homepage etc. Thus a nested data structure is required. It is evident that this requirement is best served in a syntax that explicitly allows for it such as XML. However in 1997—when ReDIF was designed—XML was not available. We are still convinced that the template syntax is more human readable and easier understood. However the computer can not find which attributes correspond to the same cluster unless some ordering is introduced. We proceed as follows. For each group of arguments that make up a cluster we specify one attribute as the "key" attribute. Whenever the key attribute appears a new cluster is supposed to begin. For example if the cluster describes a person then the name is the key. If an "author-email" appears without an "author-name" preceding it the parsing software aborts the processing of the template .

Note that the designation of key attributes is not a feature of ReDIF. It is a feature of the template syntax of ReDIF. It is only the syntax that makes nesting more involved. We do not think that this is an important shortcoming. In fact we believe that the nested structure involving the persons and organizations should not be included in the document templates. What should be done instead is to separate the personal information out of the document templates into separate person templates

```
Template-Type: ReDIF-Person 1.0
Name: Thomas Krichel
Email: T.Krichel@surrey.ac.uk
Author-Paper: RePEc:sur:surrec:9404
Author-Paper: RePEc:sur:surrec:9601
Homepage: http://openlib.org/home/krichel
Handle: RePEc:per:1965-06-05:thomas_krichel
```

We can then replace the author information for the first author in the paper template for RePEc:sur: surrec:9601 by

```
Author-Name: Thomas Krichel
Author-Person: RePEc:per:1965-06-05:thomas_krichel
```

The benefits of such a relational structure are clear. There is a much reduced load on administration of the system. When one element of author data—e.g. her phone number—changes, this change has to be registered at only one point in the system. A pervasive use of these relational features will allow the

	ReDIF-paper		ReDIF-article	
field	*all*	*max*	*all*	*max*
template-type	58254	1	10112	1
handle	58251	2	10110	1
title	58235	2	10110	1
author-name	98321	14	13855	6
creation-date	52730	1	8819	1
revision-date	536	8		
publication-date			510	1
abstract	22984	3	1896	1
classification-jel	20194	2	436	1
keywords	39219	3	9084	1
keywords-attent	457	1		
publication-status	6227	3	1568	1
note	9011	1	1479	2
series	4124	2		
number	16021	2	1501	1
price	4175	3		
file-url	17259	22	1853	2
order-url	2417	1		
contact-email	1141	1		
availability	7169	2		
length	33342	12		
pages			7920	1
month			489	1
issue			8705	1
volume			1293	1
year			1293	1
journal			488	1
paper-handle			19	1

Table 1: The data in article and paper templates

resolution of current author information through the current person template of the author. The user of a RePEc service would therefore find the author of the paper even though the contact information on the paper's title page may no longer be current. We leave the implementation of such systems for future work.

4 The total dataset

4.1 Aggregate Contents

In Table 1, we examine the document data in RePEc. For each field we give the total occurrences of the field in the "all" column and the maximum of occurrences that the field has within a single template in the "max" column. The document data appear in the ReDIF-paper and the ReDIF-article templates. There are two characteristics that potentially set articles apart from papers. First the paper can be understood as a preprint. From that point of view the article is a paper that has gone through some sort of peer review. In that case the distinction between paper and article has to do with the contents only. Secondly the distinction between paper and articles could be through their physical manifestation. From that point of view the article would be a document that is bound with others in a journal issue and it would therefore carry page numbers, issue numbers etc. This is the official criterion according to the ReDIF documentation. But it is not neat since the pagination may become redundant if the journal becomes electronic. In the following we will use the term "document" when we wish to refer to papers and articles simultaneously.

Total numbers for documents are given by the "template-type" and "handle" fields. Since each template should have exactly one type and exactly one handle the tiny difference between the two numbers is made up of mistakes in the dataset. The title field is also required. It is encouraging to see that most documents

have a creation date attached to them, because as the dataset grows it will become increasingly important to distinguish between recent and dated documents; only the former are likely to be of much interest. By contrast "revision-date" information is rare. Articles may also have a "publication-date". The difference of this field with the "creation-date" field is not clear. We consider this to be a design error in the template structure.

Let us consider the elements that refine the contents description. We encourage contributors to provide abstracts. The presence of abstracts for about one in three papers is very positive. The abstract field can be repeated. This is desirable when there are abstracts in different languages. A large number of the papers have a `Journal of Economic Literature` (JEL) classification code attached to them. However almost all papers in the offline papers only archive RePEc:fth have the codes and that explains a very large proportion of the classified material. Note that this data has been compiled by a librarian. For the electronic papers there are only two in five papers that have a classification field. We agree that this is a serious limitation to the quality of the data. It would have been possible to require a classification number for each paper right from the start. This would have hampered the collection effort. In particular it would have made it impossible for the WoPEc team to "snarf" bibliographic data from sites where this JEL data was not available. There is also some concern among economists that their areas of work do not match with these codes. The use of more complete and sophisticated classification schemes would not be possible. The main argument against requiring JEL classification codes was, however, that there is considerable opposition against the scheme in the heterodox Econonomics community. They feel that the JEL classification scheme reflects the view of the orthodoxy. Requiring JEL classification codes would have meant excluding these contributors. Then and now only a tiny part of the collection could be grouped as heterodox. However our aim is that RePEc be a broad church. This was the decisive argument against requiring the use of JEL codes.

There is a large number of templates that have keywords. About 50% of these templates come from RePEc:fip where each paper has a keyword. ReDIF allows for both free and qualified controlled vocabulary. This facility is used by for the internal keyword scheme of the `Attent: Research Memoranda` database. They are only used by the RePEc:dgr archive.

The "publication-status" field can be used to indicate where the paper has been submitted to and where the paper has been formally published. This field appears in the data from large research bodies that have been issuing a series of papers for many years and that have data about the formal publication of the paper. The fields "series" and "number" are somewhat redundant since this information should also be available from the handle. The "price" field normally refers to the delivery of a printed copy. The mode of delivery is often just expressed in the "price" field. The "file-url" field refers to the "full text" locus of a part of the full text. Usually it is the complete full text.

The document may have several components in addition to the full text. These can be listed as several "file" clusters. Each may carry an uncontrolled field about its function within the paper. For example the author may wish to supply a computer program that was used to produce the paper. In that case a whole series of files may be made available. However that is not the way the option of having many files is actually exercised. Most of the time it is used to include elements like graphics or tables that the author did not manage to include into the main document file.

The "order-url" field is used to point to an intermediate page that sits between our description and the files of the document. In that case we are not aware if the resource does actually exist online. "order-url" may be used in conjunction with the "file-url" attribute. Note that there is no "order-email" field in the document templates. Such a field figures in the series template, because the ordering of a paper should be the same for all papers in the series. The "contact-email" may otherwise be used to contact the somebody who has any connection with the paper. This field is only used by the contributors to the RePEc:wpa archive. The "availability" is used most of the time to signal that the paper is no longer in print.

Finally a "length" attribute can be used to indicate how many pages the reader has to go through to read the paper. This field is present in all templates provided by RePEc:fth and it seems to appear in a surprisingly large number of other templates.

Articles have a number of specific attributes that are listed at the bottom of the table. Strictly speaking these are not descriptive elements of the articles themselves, they rather relate to the position the article has within the journal. Finally the "paper-handle" allows to point from the preprint version to the article template.

file			person			organization		
name	*all*	*max*	*name*	*all*	*max*	*name*	*all*	*max*
url	19112	1	name	112176	1	name	8598	1
format	19024	1	postal	8	1	postal	2118	2
size	2630	1	homepage	1557	2	homepage	596	2
function	1661	1	email	3166	2	email	1451	3
restriction	2548	1	phone	282	1	phone	164	1
			fax	259	1	fax	197	1
			workplace-name	8598	4			

Table 2: The data in clusters

4.2 The clustered data

The data available in Table 1 is not the complete set of information available in the dataset. It only lists the individual attributes and the key attributes of clusters in the paper and article templates. In Table 2 we have the data that is contained in the clusters in this subset of the RePEc data. This data is therefore consistent with the data in Table 1.

There are three types of clusters, "file", "organization" and "person". The numbers that are present suggest that there are significant possibilities for a relational structure in the dataset between persons and their organizations. An interesting consideration in the person cluster is the high number of workplace templates. Providers of the data seem to attribute more importance to the workplace of a person rather than to her strictly personal data, e.g. her homepage. The only explanation that we can offer here is that most likely the data is provided by an agent of the workplace. The low number of homepages is an indicator which also suggests that in most cases the provider is not the author herself. Note also that the workplace information—when it is present—is much more complete than the corresponding data for the individuals.

5 User services

There would be little point in collecting all that data if there were no users to use them. Note that there is no official user service for RePEc. The implicit ability and explicit intention to allow for many user services at one time is a key features of RePEc. This provides an important selling point once a potential provider understands that submitting data to RePEc means submitting the data to all the user services at once. Here we list the most important user services in Subsection 5.1, before we critically discuss them in Subsection 5.2.

5.1 The main user services

By order of historical appearance, they are

BibEc at http://netec.mcc.ac.uk/BibEc.html &

WoPEc at http://netec.mcc.ac.uk/WoPEc.html

provide static html pages for all working papers that are only available in print (BibEc) and all papers that are available electronically (WoPEc). Both datasets use the same search engines. There are three search engines, a full text WAIS engine, a fielded search engine based on the mySQL relational database and a ROADS fielded search engine. Note that the mySQL database is also used for the control of the relational components in the RePEc dataset. BibEc and WoPEc are mirrored in the United States and Japan as part of the NetEc project.

IDEAS at http://ideas.uqam.ca

provides an Excite index of static html pages that represent all Paper, Article and Software templates. This is by far the most popular RePEc user interface.

NEP: New Economics Papers at http://netec.wustl.edu/NEP

is set of reports on new additions of papers to RePEc. Each report is edited by subject specialists who receive information on all new additions and then filter out the papers that are relevant to the subject of the report. These subject specialists are PhD students and young researchers. They work as volunteers. On 27 June 1999 there were 1766 different email addresses that subscribed to at least one list.

Tilburg University Working papers and research memoranda at http://www.kub.nl/~dbi/demomate/repref.htm
This site also operates a Z39.50 server for all downloadable papers in RePEc is available at dbiref.kub.nl:9997. The database name is "repref". The attribute set is Bib-1, and the record syntax supported are USmarc, SUTRS, GRS-1 (only string tags, tag type 3).

RuPEc at http://www.ieie.nsc.ru/RuPEc
is a server in Russian. It does not only provide search facilities for Russian users but also archival facilities for Russian contributors.

INOMICS at http://www.inomics.com/query/search
not only provides an index of RePEc data but also allows simultaneous searches in indexes of other web pages related to Economics.

The "Tilburg University Working papers and research memoranda" service is operated by a library-based group that has received funding from the European Union. INOMICS is operated by the Economics consultancy `Berlecon`. All the other user services are operated by junior academics.

5.2 The usage of user services

Thomas Krichel founded both the WoPEc user service in 1993 and NEP in 1998. José Manuel Barrueco has been the intensively involved in WoPEc user education. Our experience suggests that the average users from developed countries are at the postgraduate and doctoral level. There are many users in developing countries. In these countries the user community includes more senior levels, i.e. more junior academics and professional researchers rather than students. For them the RePEc user services are one of the very few means to get hold of research papers. We think that this is the most rewarding aspect of our work. The free provision of RePEc helps to reduce the gap between the informationally rich and the informationally poor.

The use of RePEc services among senior academics in the developed countries seems to be low. Is this because these people are too much set in their ways to use these modern facilities? We do not think so. Some people think that the low usage by tenured academics We believe that the current user services do not meet the information needs of these people. Academics do not need large-scale information services that they can search. The larger the scale the more likely they are to find information they did not seek and the less likely they are to find information that they want. Since they are working within a very narrow field and only have little time to read a small amount of literature small-scale information services are more tailored to their needs. In addition the contents of the service should be highly selective. Among the current user services that are built on the RePEc data, NEP comes closest to such services. Our anecdotal evidence suggests that this is the service that has the largest proportion of tenured academics.

RePEc as such can not provide small-scale user services. It can only provide the basis for such user services to exist. We are aware of two approaches to build such services. Section 4 of **?**) describes design features for a current awareness portal system where each researcher could register the subject and type of records that she is interested in. The portal would then be able to inform the researcher about new resources in her field. A second approach is outlined is Section 6 of **?**) . Here the idea is to build peer review web ("SurWeb") services. These are supposed to extend NEP to full peer review. It is too early to speculate if such a system can be put into place.

6 Conclusions

The free provision of educational material can be implemented through a central institution. Such an institution needs to be subsidized by central funds. The alternative is to provide the resources by a large number of agents. Then the cost of providing access can be absorbed within each institution. In that case the question of a comprehensive catalog arises. Such a catalog is needed to provide access to the collection in a unified way.

In this paper we have dealt with the provision of a key resource i.e. academic papers. We have presented a collection of metadata that is provided by decentralized archives. We have found that it is possible to build such a collection to a reasonable degree of accuracy if some archives where mistakes occur are aided by others. There needs to be a small group of people who actively support the collection. However this support can be given in decentralized fashion without the need for much coordination between supporters.

The academic library community in the United Kingdom as a whole has made a important contribution to RePEc by donating funds to the work of the WoPEc project. This has allowed the WoPEc project to

collect metadata about papers that are published by institutions that are not yet contributing to RePEc. This was a vital aspect of WoPEc project. The data collected by WoPEc constituted 90% of the RePEc data when RePEc was founded. However nowadays that proportion is falling. The funding for WoPEc has run out but the WoPEc web site continues to expand because of the contributions by made by RePEc archives. The software is maintained by volunteers.

Librarians should carefully consider the vision of the project. This is a kind of academic self-organization where academics publish and catalog their own work. RePEc benefits from network externalities. The more academics join the more those who have not joined will feel pressure to join. If the data is freely available than authors can communicate with their peers without the need of intermediaries. The providers of intermediation services have every reason to be worried. They include publishers *and* librarians. If librarians do not play a more active part by supporting developments like RePEc there will be no more rôle for them in the future. Write to `RePEc@netec.mcc.ac.uk`.

References