RePEc, an Open Library for Economics *

24 March 2000

Thomas Krichel
Palmer School of Library and Information Science
Long Island University
720 Northern Boulevard
Brookville, New York 11548–1300
USA
Thomas.Krichel@liu.edu
http://openlib.org/home/krichel

Abstract

After arXiv.org, the RePEc economics library offers the second-largest source of freely downloadable scientific preprints in the world. RePEc has a different business model and a different content coverage than arXiv.org. This paper addresses both differences.

As far as the business model is concerned, RePEc is an instance of a concept that I call the "Open Library". An Open Library is open in two ways. It is open for contribution (third parties can add to it), and it is open for implementation (many user services may be created). Conventional libraries—including most digital libraries—are closed in both directions.

As far as the content coverage is concerned, RePEc seeks to build a *relational* dataset about scholarly resources and other content relating of to these resources. This basically means the identification of all authors, all papers and all institutions that work in economics. Such an ambitious project can only be achieved if the cost to collect data is decentralized and low, and if the benefits to supply data are large. The Open Library provides a framework where these conditions are fulfilled.

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

*The work discussed here has received financial support by the Joint Information Systems Committee oftp://swopec.hhs.se/RePEc/hhs/soft/remi/f the UK Higher Education Funding Councils through its Electronic Library Programme A version of this paper was presented at the PEAK conference at the University of Michigan on 2000–03–24. I am grateful to Ivan V. Kurmanov for comments on that version. In March 2001, I revised and updated the paper following suggestions by Jeffrey K. MacKie-Mason and Emily Walshe. Neither of them bear no responsibility for any remaining errors.

1 Introduction

Digital or digitisable data is supplied by publishers, to be consumed by readers. Reports of research results in research "papers" form the bulk of academic digital or digitisable data, and I will refer to these as documents in what follows.

In this chapter I am not concerned with the demand for documents, nor am I not concerned with the supply of documents. Instead, I focus on the supply of information about documents. For some documents, holding detailed information about the document is as good as holding the document itself. This is typically the case when the document can be accessed on the Internet without any access restriction. Such a document will be called a public access document. Collecting data about documents is therefore particularly relevant for public access documents.

The main idea that is brought forward in this paper is the "Open Library". Basically, an open library is a collaborative framework for the supply and usage of document data. Stated in this way the idea of the open library is quite trivial. To fully appreciate the concept, it is useful to study one open library in more detail. My example is the RePEc dataset about Economics. In Section 2, I introduce RePEc as a document data collection. In Section 3, I push the RePEc idea further. I discuss the extension of RePEc that allows one to describe the discipline, rather than simply the documents that are produced by the members of the discipline. In Section 4, I make an attempt to define the open library more precisely. The example of RePEc demonstrates the relevance of the open library concept. I conclude the paper in Section 5.

The efforts of which RePEc is the result go back to 1992. I deliberately stayed away from a description of the history of the work to concentrate on the current status. Therefore, insufficient attribution is given to the people who have earned historic merits by contributing to the RePEc effort. See Krichel (1997) for an account of the early history of the NetEc projects. These can be regarded as precursors of RePEc.

2 The RePEc document dataset

2.1 Origin and motivation of RePEc

A scholarly communication system brings together producers and consumers of documents. For the majority of the documents, the producers do not receive a monetary reward. Their effort is compensated through a wide circulation of the document and a mark of peer approval for it. Dissemination and peer approval are the key functions of scholarly communication.

Scholarly communication in economics has largely been journal-based. Peer review plays a crucial role. Thorough peer review is expensive in time. According to Trivedi (1993), it is common that a paper takes over three years from submission to publication in an academic journal, not counting rejections. From informal evidence, slowly rising publication delays have stabilized in the past few years as journal editors have fought hard to cut down on what have been perceived to be intolerable delays.

Researchers at the cutting edge cannot rely solely on journals to keep abreast of the frontiers of research. Prepublication through discussion papers or conference proceedings is now commonplace. Access to this informally disseminated research is often limited to a small number of readers. It relies on the good will of active researchers to disseminate their work. Since good will is in short supply, insider circles are common.

This time gap between informal distribution and formal publication can only fundamentally be resolved by reforming the quality control process. The inconvenience resulting from the delay can however be reduced by improving the efficiency of the informal communication system. This is the initial motivation behind the RePEc project. Its traditional emphasis has been on documents that have not gone through peer review channels. Thus RePEc is essentially a scholarly dissemination system on the Internet. It is independent of the quality review process.

2.2 Towards an Internet-based scholarly dissemination system

The Internet is a cost-effective means for scholarly dissemination. Many economics researchers and their institutions have established web sites. However, they are not alone in offering pages on the Web. The Web has grown to an extent that the standard Internet search engines only cover a fraction of the Web, and that fraction is decreasing over time (Lawrence and Giles 1999). Since much of the Economics research uses common terms such as "growth", "investment" or "money", it is likely that a subject search on the entire Web would yield an enormous amount of hits. There would be no practical way to find which pages contain economics research. Due to this low signal to noise ratio, the Web *per se* does not provide an efficient mechanism for scholarly dissemination. An additional classifying scheme is required to segregate references to materials of interest to the economics profession.

The most important type of material relevant to scholarly dissemination are research papers. One way to organize this type of material has been demonstrated by the arXiv.org preprint archive, founded in 1991 by Paul Ginsparg of the Los Alamos National Laboratory with an initial subject area in high energy physics. Authors use that archive to upload papers, which remain stored there. ArXiv.org has now assembled over 150,000 papers, covering a broad subject range of mathematics, physics and computer science, but concentrating on the original subject area. An attempt has been made to emulate the arXiv.org system in economics with the "Economics Working Paper Archive" (EconWPA) based at Washington University in St. Louis. Its success has been limited. There are a number of potential reasons:

- Economists do not issue preprints as individuals; rather, economics departments and research organizations issue working papers.
- Economists use a wider variety of document formatting tools than physicists. This reduces the functionality of online archiving and makes it more difficult to construct a good archive.
- Generally, economists are not known for sophisticated practices in computer literacy and as such, they

are more likely to encounter significant problems with uploading procedures.

- There is considerable confusion as to implications of networked pre-publication on a centralized, high-visibility system for the publication in journals.
- Economics research is not confined to university departments and research institutes. There are a number of government bodies—central banks, statistical institutes, and other—who contribute a significant amount of research in the field. These bodies, by virtue of their size, have more rigid organizational structures. This makes the coordination required for a central research paper dissemination more difficult.

An ideal system should combine the decentralized nature of the Web, the centralized nature of the arXiv.org archive, and a zero price to end users. I discuss these three requirements in turn.

The system must have a decentralized storage of documents. To illustrate, let us consider the alternative scenario. This would be one where all documents within a certain scope, say within a discipline, would be held on one centralized system. Such a system would not be ideal for three reasons. First, those authors who are rejected by that system would have no alternative publication venue. Since economics is a contested discipline, this is not ideal. Second, the storage and description of documents is costly. The centralized system may levy a charge on contributors to cover its cost. However, since it enjoys a monopoly, it is likely to use this position to extract a rent from authors. This would not be ideal.

The centralized nature of the arXiv.org system is the ability to have a one-stop-shop where all the papers relevant to economics are accessible. Again, to show that this a requirement for an ideal dissemination system, imagine a situation where that would not be the case. In that case, the search for documents would be difficult and therefore the distribution of documents not optimal.

To explain why the end-user access to the dissemination system should be free, it is useful to refer to the distinction between trade authors and esoteric authors, as done by Harnad (1995). Authors of academic documents are esoteric authors rather than trade authors. They do not expect payments for the written work; instead, they are chiefly interested in reaching an audience of other esoteric authors and to lesser extent, the public at large. Therefore the authors are interested in wide dissemination. If a tollgate to the dissemination system is set-up, then the system as such falls short of an ideal one.

Having established the three criteria for an ideal system, let me turn to the problem of implementing it. The first and third objectives could be accomplished if departments and research centers allow for public access to their documents on the Internet. But for the second, we need a library to hold an organized catalog. The library would collect what is known as "metadata": data *about* documents that are available using Internet protocols. There is no incentive for any single institution to bear the cost of establishing a comprehensive metadata collection, without external subsidy. However, since every institution will benefit from participation in such an effort, we may solve this incentive problem by creating a virtual collection via a network of linked metadata archives. This network is open in the sense that persons and organizations can join by contributing data about their work. It is also open in the sense that user services can be created from it. This double openness promotes a positive feedback effect. The larger the collection's usage, the more effective it is as a dissemination tool, and thus more authors and their institutions join as participation is open. The larger the collection, the more useful it becomes for researchers. This leads to more usage.

Bringing a system to such a scale is a difficult challenge. Man is an animal of habit. Scholarly communication systems have evolved time. Academic careers are directly dependent on the results of the scholarly communication. Therefore, change in the this area is slow because it involves important aspects of the lives of those who are the potential implementors of the change. A scholarly dissemination system on the Internet is more likely to succeed if it enhances current practice, but it does not 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 organize this process on the Internet.

2.3 The architecture of RePEc

RePEc can be understood as a decentralized academic publishing system for the economics discipline. RePEc allows researchers' departments and research institutes to participate in a decentralized archival scheme which makes information about the documents that they publish accessible via the Internet. Individual researchers may also openly contribute, but they are encouraged to use EconWPA.

Each contributor needs to maintain their own collection of data using a set of standardized templates. Such a collection of templates is called an "archive". An archive operates on an anonymous ftp server or a Web server controlled by the archive provider. Each archive provider has total control over the contents of its archive. There is no need to transmit documents elsewhere. The archive management retains the liberty to post revisions or to withdraw a document.

2.3.1 An example archive

Let us look at an example. The archive of the OECD is at http://www.oecd.org/eco/RePEc/oed/. In that directory we find two files. The first is oedarch.rdf:

```
Template-Type: ReDIF-Archive 1.0
Handle: RePEc:oed
Name: OECD Economics Department
Maintainer-Email: eco.contact@oecd.org
Description: The working papers of the
 Economics Department of the OECD
URL: http://www.oecd.org/eco/RePEc/oed
```

This file gives basic characteristics about the archive. It associates a handle with it, gives an email address for the maintainer, and most importantly, provides the URL where the archive is located. This archive file gives no indication about the contents of the archive. The contents list is in a second file, oedseri.rdf:

```
Template-type: ReDIF-Series 1.0
Name: OECD Economics Department working
  papers
Type: ReDIF-Paper
Provider-Name: OECD Economics Department
Provider-Homepage:
 http://www.oecd.org/eco/eco/
Maintainer-Email: eco.contact@oecd.org
Handle: RePEc:oed:oecdec
```

This file lists the content as a series of papers. It associates some provider and maintainer data with the series, and it associates a handle with the series. The format that both files follow is called ReDIF. It is a purpose-built metadata format. Appendix B discusses technical aspects of the ReDIF metadata format that is used by RePEc. See Krichel (2000) for the complete documentation of ReDIF.

The documents themselves are also described in ReDIF. The location of the paper description is found through appending the handle to the URL of the archive, i.edot at http://www.oecd.org/eco/RePEc/oed/oecdec. This directory contains ReDIF descriptions of documents. It may also contain the full text of documents. It is up to the archive to decide whether to store the full text of documents inside or outside the archive. If the document is available online—inside or outside the archive—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 files, e.g. a dataset or a computer program.

Participation does not imply that the documents are freely available. Thus, a number of journals have also permitted their contents to be listed in RePEc. If the person's institution has made the requisite arrangements with publishers (e.g. JSTOR for back issues of Econometrica or Journal of Applied Econometrics), RePEc will contain links to directly access the documents.

2.3.2 Using the data on archives

One way to make use of the data would be to have a web page that lists all the available archives, and allow users to navigate through the archives on the search for documents that they may be interested in. However, that would be quite a primitive way to access the data. First, the data as shown in the ReDIF form is not itself hyperlinked. Second, there is no search facility, no filtering of contents, etc..

The provision of services that allow for convenient access of users is not a concern for the archives, but for user services. User services render the RePEc data in a form that make it convenient for a user. User services are operated by members of the RePEc community, libraries, research projects etc.. Each service has its own name. There is no "official" RePEc user service. A list of services in at the time of writing may be found in Appendix A.

User services are free to use RePEc data in whatever way they see fit, as long as they observe the copyright statement for RePEc. This statement places some constraints on the usage of RePEc data:

```
You are free to do whatever you want with
this data collected on the archives that
are described here,  provided that you

(a) Don't charge for it or include it in
    a service or product that is not free
    of charge.

(b) When displaying the contents of a
    template (or part of a template) the
    following fields must be shown if
    they are present in the template:
    Title, Author-Name, File-Restriction
     and Copyright (if present).

(c) You must contribute to RePEc by
    maintaining an archive that actively
    contributes material to RePEc.

(d) You do not contravene any copyright
    statement found in any of the
    participating archives.
```

Within the constraints of that copyright statement, user services are free to provide all, or only a subset of, the RePEc data. For example, one service may only show papers that are available electronically, another may restrict the choice to act as a quality filter. In this way services implement constraints on the data, whether they be availability constraints or quality constraints.

The RePEc data may not be sold or incorporated into a product that is sold. Therefore all RePEc services are free. 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.

2.3.3 Building user services

The provision of a user services usually starts with putting frequently updated copies of RePEc archives on a single computer system. This maintenance of a frequently updated copy of archives is called "mirroring". Everything contained in an archive may be mirrored. For example, if a document is in the archive, it may be mirrored. If the archive management does not wish the document to be mirrored, it can store it 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.

An obvious way to organize the mirroring process overall 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. Some may update daily, while some may only update weekly. One disadvantage of this system is that it is not known how long it takes for a new item to be propagated through the system.

2.4 The documents available through RePEc

Over 160 archives in 25 countries currently participate in RePEc, some of them representing several institutions. Over 100 universities contribute their working papers, including U.S. institutions such as Berkeley, Boston College, Brown, Maryland, MIT, Iowa, Iowa State, Ohio State, UCLA, and Virginia. The RePEc collection also contains information on all NBER Working Papers, the CEPR Discussion Papers, the contents of the Fed in Print database of the US Federal Reserve, and complete paper series from the IMF, World Bank and OECD, as well as the contributions of many other research centers worldwide. Last, but not least, RePEc also includes the holdings of EconWPA. In total, at the time of writing in March 2001, over 37,000 items are downloadable.

The bibliographic templates describing each item currently provide for papers, articles, and software components. The article templates are used to fully describe published articles. They are currently in use by the Canadian Journal of Economics, Econometrica, the Federal Reserve Bulletin, and IMF Staff Papers, the Journal of Applied Econometrics, the RAND Journal of Economics. These are only a few of the participating journals. Participation does not imply that the articles are freely available.

The RePEc collection of metadata also contains links to several hundred "software components"—functions, procedures, or code fragments in the Stata, Mathematica, MATLAB, Octave, GAUSS, Ox, and RATS languages, as well as code in FORTRAN, C and Perl. The ability to catalog and describe software components affords users of these languages the ability to search for code applicable to their problem—even if it is written in a different language. Software archives that are restricted to one language, such as those maintained by individual software vendors or volunteers, do not share that breadth. Since many programs in high-level languages may be readily translated from, say, GAUSS to MATLAB, this breadth may be very welcome to the user.

3 The ReDIF metadata

From the material that we have covered in the previous section, we can draw a simple model of RePEc as

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

The term "RePEc" is initially an acronym; it stands for Research Papers in Economics. In fact the term should now to be a literal, because RePEc is about more than the description of resources. It is probably best to say that RePEc is a relational database about economics as a discipline.

One possible interpretation of the term "discipline" is given by Karlsson and Krichel (1999). They have come up with a model of the discipline, as consisting essentially of four elements arranged in a table:

resource	collection
person	institution

A few words may help to understand that table. A "resource" is essentially any output of academic activity: a research document, a dataset, a computer program, or anything else that an academic person would claim authorship for. A "collection" is a logical grouping of resources. For example the act of peer review may be represented by a resource being included in a collection. A "person" is a physical person or a corporate body who acts as a physical person in the context of RePEc.

These data collectively form a relational database that not only describes papers, but also the authors who write them, the institutions where the authors work, and so on. All this data is encoded in the ReDIF metadata format. I illustrate this in Subsection 3.2 and Subsection 3.3 for the institutional and the personal data, respectively.

3.1 A closer look at the contents

To understand the basics of ReDIF it is best to start with an example. Here is a—carefully selected—piece of ReDIF data at 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 resembles a standard bibliographical format, with authors, title etc.. The only thing that appears a bit mysterious here is the "Author-Person" field. This field quotes a handle that is known to RePEc. This handle leads to a record maintained 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
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
```

In this record, we have the handles of documents that the person has written. This record will allow user services to

[1] I suppress the Abstract: field to conserve space.
[2] I leave out a few fields to conserve space.

list the complete papers by a given author. This is obviously useful when we want to find papers that one particular author has written. It is also useful to have a central record of the person's contact details. This eliminates the need to update the relevant data elements on every document record. In fact the record on the paper template may be considered as the historical record that is valid at the time when the paper was written, but the address in the person template is the one that is currently valid.

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:

```
Template-Type: ReDIF-Institution 1.0
Primary-Name: University of Surrey
Primary-Location: Guildford
Secondary-Name: Department of Economics
Secondary-Phone: (01483) 259380
Secondary-Email: economics@surrey.ac.uk
Secondary-Fax: (01483) 259548
Secondary-Postal: Guildford, Surrey GU2 5XH
Secondary-Homepage:
 http://www.econ.surrey.ac.uk/
Handle: RePEc:edi:desuruk
```

It would take us too far here to discuss this record in more detail. It is probably more interesting to know where these records come from.

3.2 Institutional registration

The registration of institutions is accomplished through 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 this writing. All these records are included in RePEc. For all institutions, data on their homepage is available, as well as postal and telephone information. For some, there is even data on their main area of work. Thus it is possible to find a list of institutions where—for example—a lot of work in labor economics in being done. At the moment, EDIRC is mainly linked to the rest of the RePEc data through the HoPEc personal registration service. Other links are possible, but are rarely used.

3.3 Personal registration

HoPEc has a different organization from EDIRC. It is impossible for a single academic to register all persons who are active in economics. One possible approach would be to ask archives to register people who work at their institution. This will make archive maintainers' work more complicated, but the overall maintenance effort will be smaller once all authors are registered. However, authors move between archives, and many have work that appears in different archives. To date, there is no satisfactory way to deal with moving authors. For this reason, the author registration is carried out using a centralized system.

A person who is registered with HoPEc is identified by a string that is usually close to the person's name and by a date that is significant to the registrant. HoPEc suggests the birth date but any other date will do as long as the person can remember it. When registrants works with the service, they first supply some personal information. The data that is requested is mainly the name, the URL of the registrant's homepage, and the email address. Registrants are free to enter data about their academic interests—using the Journal of Economic Literature Classification Scheme—and the EDIRC handle of their primary affiliation.

When the registrant has entered this data, the second step is to create associations between the record of the registrant and the document data that is contained in RePEc. The most common association is the authorship of a paper. However, other associations are possible, for example the editorship of a series. The registration service then looks up the name of the registrant in the RePEc document database. The registrant can then decide which potential associations are relevant. The authentication methods are weak. HoPEc relies on honesty.

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 delivers 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. Last but by no means least, there are all kinds of privacy issues involved in the composition of such a dataset. For example, Sune Karlsson has informed me that setting up a database such as HoPEc would be illegal in Sweden.

To summarize, HoPEc provides information about persons' identity, affiliation and research interests and links these data with resource descriptions in RePEc. This allows to identify persons and update their metadata in a timely and cost efficient way. These data could also fruitfully be employed for other purposes, such as maintaining membership data for scholarly societies or for lists of conference participants. It is hoped that the HoPEc data will be used as a shared pool of common personal data.

4 The open library

This section of the chapter is somewhat more theoretical. It sets out a body of thought that is built on the experience of RePEc. It is an attempt to find a general theory that could apply in a wide set of circumstances in which similar systems are desirable. I call this general concept, "open library". The parallel to the "open source" concept is intentional. It is therefore useful to review the open source concept first.

4.1 The open source concept

There is no official and formal definition what the term "open source" means. On the Open Source Initative at http://www.opensource.org an elegant introduction to the idea is found

> The basic idea behind open source is very simple. When programmers can read, redistribute, and modify the source code for a piece of software, the software evolves. People improve it, people adapt it, people fix bugs. And this can happen at a speed that, if one is used to the slow pace of conventional software development, seems astonishing.
>
> We in the open source community have learned that this rapid evolutionary process produces better software than the traditional closed model, in which only a very few programmers can see the source and everybody else must blindly use an opaque block of bits.

Open source software imposes no restrictions on the distribution the source code of a software. The source code is the code that is required to build a running version of the software. As long as users have no access to source code, they may be able to use a running version of the software, but they can not change the way that the software behaves. The latter involves changing the source code and rebuilding the running version of the software from the source code. Since building the software out of the source code is quite straightforward, software that has freely available source code is essentially free.

4.2 Open Source and open library

The open source movement claims that the building of software in an open, collaborative way—enabled by the sharing of source code—allows to build software better and faster. The open library concept is an attempt to apply the concept of open source to a library setting. We start off with the RePEc experience.

Within the confines of RePEc as a document collection, it is unrealistic to expect a free distribution of document source code. Such source code is, for example, the word processor file of an academic paper. If such source code would be available for others to change, then the ownership of the intellectual property in the document would be dissolved. Since intellectual property over scientific ideas is crucial in the academic reward system, it is unlikely that such source code distribution will take place. Within the confines of RePEc's institutional and personal collection, there is no such source code that could be freely shared.

To apply the open source principle to RePEc we must conceptualize RePEc as a collection of metadata. The term "metadata" literally means data about data. Strictly speaking its use is inappropriate in the context of RePEc because some of the objects of description of RePEc are not data but physical objects. However, I will continue to use the term metadata for the kind of data that is collected by RePEc.

In terms of the language adopted by the open source concept, the metadata record is the "source code". The way the metadata record is rendered in the user interface is the "software" as used by the end user. We can the define the open library as a collection of metadata records that has few special properties.

4.3 The definition of the open library

An open library can be defined as follows. An open library is a collection of metadata records that has the following characteristics

- *Every record is identified by a unique handle.* This requirement distinguishes the library from an archive. It allows for every record to be addressed in an unambiguous way. This is important if links between records are to be established.

- *Records have a homogeneous syntax of field names and field values.* This requirement constrains the open library to appear like a database. If this requirement would not be present, all public access pages on the Web would form an open library. Note that this requirement does not constrain the open library to contain a homogeneous record format.

- *The documentation of the record format is available for online public access.* For example, a collection encoded in MARC format would not qualify as an open library because access to the documentation of MARC is restricted. Without this requirement the cost of acquiring the documentation would be an obstacle to participation.

- *The collection is accessible on a public access computer system.* This is the precondition to allow for the construction of user services. Note that user services may not necessarily be open to public access.

- *The collection is contributable to without monetary cost.* There are of course non-monetary costs to contribute to the open library. however the general principle is that there is no need to pay for either contributing or using the library. The copyright status of data in an open library should be subject to further research.

4.4 The open library and the Open Archive

Stimulated by work of Van de Sompel, Krichel, Nelson, et al. (2000), there have been recent moves towards improving the interoperability of e-print archives such as arXiv.org, NCSTRL, RePEc etc. This work is now called the Open Archive Initative at http://www.OpenArchives.org (OAi), a term coined by Stevan Harnad. The basic business model proposed by the OAI is very close to the RePEc project. In particular, the open archive technical protocols allow for the separation between data provision and data implementation that is a key feature of the open library model, as pioneered by RePEc since 1997. In addition, because of their ability to transport multiple metadata sets, the open archive protocols allow for several open libraries to establish on one physical system.

4.5 The conceptual challenge raised by the open library

The open library as defined in Subsection 4.3 may be a relatively obvious concept. It certainly is not an elaborate intellectual edifice. Nevertheless, the open library idea raises some interesting conceptual challenges.

4.5.1 Supply of information

To me as a newcomer to the Library and Information Studies (LIS) discipline, there appears to be a tradition of emphasizing the behavior of the user who demands information rather than the publisher—I use the word here in its widest sense—who supplies it. I presume this orientation comes from the tradition that almost all bibliographic data were sold by commercial or not-for-profit vendors, just as the documents that they describe. Libraries then see their role as intermediaries between the commercial supply and the general public. In that scenario, libraries take the supply of documents and metadata as given.

The open library proposes to build new supply chains for metadata. If all libraries contribute metadata about objects that are local to them—what that means would have to be defined—then a large open library can be built.

An open library will only be as good as the data that contributors will give to it. It is therefore important that research be conducted on what data contributors are able to contribute; on how to provide documentation that the contributor can understand; and on understanding a contributor's motivation.

4.5.2 Digital updatability

For a long time the library profession has purchased material that is essentially static. It may be subject to physical decay but the material that it contains is immutable. Digital resources have made mass appearance only a few years ago. These resources may be changed at any time. The change from static to dynamic resource is a major challenge for the LIS profession. Naturally the inclination has been to demand that the digital resources be like the non-digital resource in all but their physical medium. The debate on digital preservation is a result of that demand. Thus the dynamic nature of digital metadata has been seen more as a threat rather than as an opportunity. The open library is more concerned with digital updatability than digital preservation.

4.5.3 Metadata quality control

In the case of a decentralized dataset, an important problem is to maintain metadata quality. Some elements of metadata quality cannot be controlled by a computer. For example, each record must a structure of fields and values associated with these fields to be interoperable with other records. In some cases the field value only makes sense if it has a certain syntax. This is the case, for example, with an email address. One way to achieve quality control is through the use of relational metadata. Each record has an identifier. Records can use the identifiers of other records. It is then possible to update elements of the dataset in an independent way. It is also quite trivial to check if the handle referenced in one record corresponds to a valid handle in the dataset. Highly controllable metadata systems are an important research concern that is related to the open library concept.

5 Conclusions

To my knowledge, Richard Stallman was the pioneer of open source software. He founded the GNU project in 1984 to write a free operating system to replace Unix. At the time few people believed that such an operating system would ever come about. The same may hold for my audience today, when I am calling for an open library. But remember that in the late 1990s the Open Source movement has basically realized Stallman's dream.

Building GNU took a long time. But the obstacles facing the open source movement are much more daunting then the obstacles facing the open library movement:

- The structural complexity of the operating system of a modern computer is much higher than the structural complexity of a metadata collection.
- Computer programming is a highly profitable activity for the individual who is capable of doing it; therefore the opportunity cost of participating in what is essentially an unpaid activity is much higher. These costs are much lower for the academic or the academic librarian who would participate in an open library construction.
- There is a network effect that arises when the open library has reached a critical mass. At some stage the cost of providing data is much smaller than the benefit—in terms of more efficient dissemination—of contributing data. When that stage is reached, the open library can grow without external public or private subsidy.

It remains to be seen how much inroad the open library concept will make.

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.

Harnad, Stevan (1995). The Postgutenberg Galaxy: how to get there from here. available at http://www.cogsci.soton.ac.uk/~harnad/THES/thes.html.

Karlsson, Sune and Thomas Krichel (1999). RePEc and S-WoPEc: Internet access to electronic preprints in Economics. presented at the Third ICCC/IFIP Conference on Electronic Publishing in Ronneby, May 10–12 May 1999, available at http://gretel.econ.surrey.ac.uk/papers/lindi.pdf.

Krichel, Thomas (1997). About NetEc, with special reference to WoPEc. *Computers in Higer Education Economics Reviev 11(1)*, 19–24. available at http://netec.mcc.ac.uk/doc/hisn.html.

Krichel, Thomas (2000). ReDIF version 1. available at http://openlib.org/acmes/root/docu/.papers/redif_1.a4.pdf.

Lawrence, Steve and C. Lee Giles (1999). Accessibility of information on the web. *Nature 400(8 July)*, 107–109.

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

Van de Sompel, Herbert, Thomas Krichel, Micheal L. Nelson, et al. (2000). The UPS Prototype project: exploring the obstacles in creating a cross e-print archive end-user service). Old Dominion Computer Science Tech Report, available at http://openlib.org/home/krichel/papers/upsproto.ps.

A The main user services

I list them by order of historical appearance.

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. The mySQL database is also used for the control of the relational components in the RePEc dataset. BibEc and WoPEc are based at Manchester Computing in Japan and the United States.

EDIRC at http://ideas.uqam.ca/EDIRC
provides a Web pages that represent the complete institutional information in RePEc.

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.mcc.ac.uk/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 junior researchers. They work as volunteers. On 14 March 2000, there are 2753 different email addresses that subscribe to at least one list.

Tilburg University working papers & 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 name of the database is "repref". The attribute set is Bib-1, and the record syntax supported are USmarc, SUTRS, GRS-1 (only string tags, tag type 3).

socionet at http://socionet.ru
is a server in Russian. It offers search facilities to Russian users. Its maintainers also provide 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.

HoPEc at http://netec.mcc.ac.uk/HoPEc.html
provides a personal registration service for authors and allows to search for personal data.

The "Tilburg University working papers & 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 Research. All the other user services are operated by junior academics.

B The ReDIF metadata format

The ReDIF metadata format is inspired by Deutsch, Emtage, Koster, and Stumpf (1994) 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: 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 when ReDIF was designed in 1997, XML was not available. We are still convinced that the template syntax is more humanly readable and easier to understand. However the computer can not find which attributes correspond to the same cluster unless some ordering is introduced. Therefore 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. I do not think that this is an important shortcoming. Instead, I 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. This approach is discussed extensively in the main body of the paper.

ReDIF is a metadata format that comes with tools to make it easy to use in a framework where the metadata is harvested. A file that is simply harvested from a computer system could contain any type of digital content. Therefore the harvested data must be parsed by a special software that filters the data. This task is accomplished by the rr.pm module written by Ivan V. Kurmanov. It parses ReDIF data and validates its syntax. For example, any date within ReDIF has to be of the ISO8601 form *yyyy–mm–dd*. A date like “14 Juillet 1789” would not be recognized by the ReDIF reading software and not passed on to application software that a service provider would use.

The *rr.pm* software uses a formal syntax specification redif.spec. This formal specification is itself encoded in a purpose-built format code-named spefor. Therefore, it is possible for ReDIF-using communities to change the syntax restrictions or even design a whole new ReDIF tag vocabulary metadata vocabulary from scratch.