University of Connecticut

Department of Economics Working Paper Series

The Economics of Open Bibliographic Data Provision

Thomas Krichel
Palmer School, Long Island University

Christian Zimmermann
University of Connecticut

Abstract

In this paper, we discuss the provision of bibliographic data as an extension of the open source concept. Our particular concern is the sustainability of such endeavors. We describe the RePEc (Research Papers in Economics) project, probably the largest "open source" bibliographic database. It demonstrates that open-source bibliographic data collection is sustainable.

1 Introduction

The Open Source model is most commonly linked to the provision of free software by a network of volunteers. There, the source code is provided so that anyone can join in the development effort. This paper shows that this model can also be applied to open academic libraries. In such libraries, academic document archives provide information freely about their contents and allow anyone to use the collected data. We define academic documents as texts that authors do not expect to be paid royalties for, that are targeted towards a very specialized audience and that do not contain advertising. Elementary economics would tell that in an environment where nobody is to gain financially from publishing such documents, there would be no way for them to be disseminated. This intuition is wrong, of course. The commercial academic publishing industry is estimated to be worth seven billion dollars annually[1].

The advent of the Internet has pushed marginal costs of online publishing near to zero. The web, in particular, is currently dramatically changing the publishing industry. Some academics become vocal over the publishers' insistence on a pricing model that appears to mimic the old print world. While obviously there is concern—on both sides of the debate—on how things will evolve, we want to focus here on another side of the academic publishing question: how to make the information about publications in their various formats available and, especially, how to organize this information.

Organizing bibliographic information is a monumental task, as any librarian will testify. There is considerable human effort necessary to enter information, categorize it and verify it. But once this information has been compiled, it can be replicated at very little cost, just like software. The question thus becomes who would be willing to venture into cataloging information. New entrants into this market will be deterred by large fixed costs and low marginal costs. One avenue is to follow Google's lead, i.e. crawl the web for all sorts of information and try to organize it in some automated way. This is useful if the goal is just to provide access to texts, see Arms (2000). But doing this in a more organized and especially monitored way is both more useful and more costly. Another avenue is the open source model. Over the following pages, we show this is possible. We use an example in Economics, the RePEc (Research Papers in Economics) project.

---

[1] As documented by the Association of Research Libraries' Create Change web site at http://www.createchange.org/.

2   Making Research Output Available

There are mainly three ways through which research output is made available to the community: scholarly journals, conferences proceedings, and preprints. As a specialist researcher, one can usually find one or several journals that are relevant to one's topic. In addition, one does look at the most famous general journals. Those outlet as often priced in a discriminatory way. For example, specialized journals in medicine are much more expensive than those in history.

Conference proceedings are usually published in a much more erratic ways. There are few regular outlets for them. Therefore, a specialist needs to be constantly on the lookout. This applies also to the opportunity to contributing to those proceedings. One needs to be aware that conferences are held. Some are of small scale and limited to a more or less closed network of cognoscenti.

Pre-prints are the jungle of academic publishing. But publication delays through the journal system has made pre-prints an essential tool of research diffusion in some fields. This has serious shortcomings. Without an organized diffusion process, researchers are on their own both for getting their results known and for learning about others' results. Again, informal networks of insiders make it possible for a select few to be ahead of the others. Some research centers or departments are better organized, with mailing lists or subscriptions to list their pre-prints. But this remains a costly system with high noise to output ratio that certainly does not provide an equitable and open access to recent research.

In the face of these various research outlets, how is the bibliographic metadata about these publications brought to the public? Currently, the main purveyors of such data are abstracting and indexing (A&I) databases, like PUBMED, ABI/Inform or PsycINFO. Of course, gathering the metadata and organizing it is expensive[2]. They act essentially as brokers of information. The majority of A&I database providers finance their operation through the sale of the access to their database. In addition to these datasets, there are also the portals of commercial publishers, like Ingenta, ScienceDirect or SpringerLink. The latter usually limit themselves to their own research output. These conventional commercial publishers fund their operations through selling subscription packages.

The advent of the Internet has not changed the scenario much. While it makes it is easier to make research results available, they are drowned in a lot of irrelevant noise. Google may have an impressive database and a very efficient matching algorithm, it cannot sort the wheat from the chaff like a researcher would like. Google Scholar is an improvement, but still has serious deficiencies, in particular

---

[2]Arms (1999) estimates the cost of creating and disseminating a single bibliographic record at $50.

with respect to the organized open academic libraries we outline below. See Jacsó (2004) for an illuminating discussion.

The troubling aspect of this is that researchers not only have to pay to access their research, which already subject to debate and mobilizations, they also have to pay to find it. This does not necessarily need to be so. Open libraries fashioned like open source co-operations can provide both publications and the metadata about them. In this paper, we concentrate on the "metadata about" aspect.

3   The Incentive Problem: How to Make it Happen

Clearly, an open academic library faces a particular hurdle: how to provide for free what A&I databases provide for a fee. Our point here is that many people are ready to provide the necessary resources for free. One just need a good way harness this willingness.

Those who are willing to volunteer for open academic libraries are those that have the most to gain from them: the publishers and the researchers. They want their work to be read and cited. They want to be recognized in the profession. And academics want to satisfy the service component of their job by doing something useful rather than sitting on committees. This willingness can be harnessed with a good system where researchers (or their department, or their publishers) input the bibliographic metadata themselves, where they maintain their own author profile, and where they check that automatically generated citation data is accurate. Once the open library is recognized as a good place to find papers, people will want to have their work listed, publishers do no want to be left out, and authors will push publishers to participate. Every participant has a strong incentive to keep all information up-to-date, as the contents of the open library provide to the community a window of the research activity, in particular once it is used for rankings or promotion. This is a textbook example of network economies.

One may have two objections to such a proposal. The first is that such a system can be rife with abuse, especially once its citation records become used for tenure or promotion purposes. Our experience shows that there is surprisingly little of it. It is easily monitored. The community watches quite carefully over abuse.

The second objection is that research documentation cannot be that highly organized. With increased organization, the cost to build the system to a critical mass is prohibitively high. We can not reject that objection generally. But we describe a working and successful example of such a system in place that documents that the has overcome the critical mass problem despite a high degree of internal organization.

4 An Example of an Open Academic Library: RePEc

Economics is a discipline that suffers from extremely long publishing delays. This comes from a combination of high rejection rates in journals, a refereeing process that often takes over a years for a single journal, frequent repetitive rounds of revisions that are resubmitted to referees, and yearlong waits from acceptance to printing. This process is supposed to guarantee the quality of peer-refereed journals. Its main consequence is that journals essentially act like historical records. The published research is typically several years old.

As a reaction to this adverse development, pre-prints have become an essential tool for the dissemination of leading edge research. As mentioned above, there are problems. The main concern is that pre-prints lead to the formation of insider networks. These informal networks lead to a concentration of top notch research in a few universities.

With the advent of the Internet, and in particular the web, this was allowed to change. With researchers placing their pre-prints on their personal or their departments' web pages, anyone can now access the latest research. Academics who want to be widely read place more material online. But this alone does not provide equitable access to research. One still needs to find the relevant home pages, and the bias is still in favor of the top universities. Indeed, this is where you would start searching, and the chances of research from lesser institutions to be sought is still minimal.Hodgson and Rothman (1999) and Kocher and Sutter (2001) have shown that citations continue to be dominated by top departments.

Clearly, what is needed is a way to concentrate the information about all these pre-prints. In Economics, this started in April 1993 through the collaboration of Féthy Mili, then a librarian at the Université de Montréal, and Thomas Krichel, then lecturer at the University of Surrey. Krichel put in place a gopher server with two collections: WoPEc contained electronic texts, or links to them, and BibEc contained the information about hard copies that Mili collected. Managed as pet projects, WoPEc and BibEc moved to the web, gradually grew and became popular sites for economists searching for the latest developments in their field.

For the four years that followed, Krichel tried to persuade other to join the effort. Some joined early on. For example, the US Federal Reserve Banks gave him a copy of their "Fed in Print" database as early as 1995. The breakthrough came in 1997. DEGREE, the national project to collect working papers in the Netherlands agreed to give their data to BibEc and WoPEc as well as to using the data that these projects already had. A national initiative for Scandinavia, the S-WoPEc project also joined the effort. Thus a system to exchange data was required. This was the birth of the RePEc project.

Over the years, RePEc has grown to impressive dimensions. At the time of this

writing, over 430 archives in 40 countries are participating, from major commercial publishers to economics departments in teaching colleges. A total of 145,000 working papers and 155,000 articles are described. 1,500 software components as well as 750 books and book chapters are thrown in for good measure. About 2/3 of the material is available online through links both to pay sites and free sites. Indeed, commercial publishers have became interested in participating in RePEc given the exposure it provides. Interestingly, some publishers that normally only provide bibliographic metadata to subscribers for a fee make it available for free through RePEc.

Each participating repository provides metadata about its contents in a purpose-built format called ReDIF (Krichel 2001). This metadata is exchanged through the so-called Guildford protocol (Krichel 1999) that organizes where the metadata is to be found and how it is structured. Thus, the participating archives deposits on a web server or an anonymous ftp server the metadata of all the publications they would like to be listed with RePEc. They can modify the data as as when they see fit. They cannot blame someone else if data is not current.

For example, the current paper would appear in the following form on the the web site or anonymous server of the publisher:

```
Template-type:  ReDIF-Paper 1.0
Author-Name:  Thomas Krichel
Author-Email:  krichel@openlib.org
Author-Homepage:  http://openlib.org/home/krichel/
Author-Workplace-Name:  Palmer School, Long Island
 University
Author-Workplace-Homepage:  http://palmer.cwpost.liu.edu/
Author-Name:  Christian Zimmermann
Author-Email:  christian.zimmermann@uconn.edu
Author-Homepage:  http://ideas.repec.org/zimm/
Author-Workplace-Name:  Department of Economics, University
  of Connecticut
Author-Workplace-Homepage:  http://www.econ.uconn.edu/
Title:  The Economics of Open Bibliographic Data Provision
Abstract:  In this paper, we discuss the provision of
 bibliographic data as an extension of the open source
 concept.  Our particular concern is the sustainability of
 such projects.  We describe the RePEc (Research Papers in
 Economics) project, probably the largest "open source"
 bibliographic database, and show that open source
 bibliographic data collection is sustainable.
Classification-JEL: L39
```

Similar templates describe the pre-print series or the journals, and RePEc services mirror these files onto their own server and then display the collected information to the public.

It should be noted that RePEc is a way to organize the collection of the metadata. To make it available to the "common user", RePEc encourages others to use it. First there are a group of services that act as portals for the RePEc data. Several such services try to woo users by displaying the RePEc data on the Web. The result is a healthy competition. The most notable user services are EconPapers at http://econpapers.repec.org/, IDEAS and http://ideas.repec.org/, Inomics at http://inomics.com/ and Socionet at http://socionet.ru/.

A more small-scale, but highly innovative RePEc user service is "NEP: New Economics Papers". Here recent additions to the working paper paper stock of RePEc are circulated to a group of volunteer editors. These editors filter the new additions data into subject specific reports. Each editor typically looks after one single subject. Currently, there are about 60 reports covering most areas of economics. The report issues are circulated by email. Data about past inclusions in report issues are fed back to the portal services. This can further help in categorizing the bibliographic metadata. The NEP service provides a further illustration how the work of volunteers can be used to further enhance the RePEc dataset. See Barrueco Cruz, Krichel and Trinidad (2003) for an illuminating discussion of NEP.

As powerful and popular as all these end-user services are, they are not the most interesting aspect of RePEc. This spot should be taken by some additional intermediary services.

The first is CitEc at http://citec.repec.org/. It provides for autonomous citation analysis for many papers listed in RePEc. Essentially, papers that are available online are converted to text, references are then extracted and matched with listed

7

works. At this point, about 50,000 works could be analyzed, citing 60,000 items. As services use this data to provide links to references and citations. Some publishers now even provide ready-made bibliographies along with their metadata to increase their visibility.

The second intermediate service is the logging service LogEc at http://logec.repec.org/. This service is based on web traffic information from the portal services and NEP. LogEc consolidates and aggregates these data. This undertaking is not as simple as it appears, as any traffic from robots, even undeclared ones, as well as repeat access from identical IP addresses need to be vetted out. The data are circulated by email to archive maintainers. They give crucial feedback to those maintainers on how well RePEc is disseminating their work.

A third add-on is the RePEc Author Service at http://authors.repec.org/. Authors register and provide their contact details and affiliation(s). Then they claim their works from the RePEc metadata. Thus, they create a profile that links to their works and vice-versa. This makes it them possible to provide each author individually with traffic and citation statistics. Currently, over 6,000 authors are registered, claiming about a third of the listed papers and articles.

The RePEc Author Service, CitEc and LogEc give authors a precise idea of how much their papers are used. The data can be used to construct rankings of economists and departments. Thus authors have high incentives to keep their profiles current. They also push their publishers to participate. There publishers can be their local Economics department, a research center or even a commercial publisher. New journals are now typically first indexed in RePEc, as it is easy and free to do.

None of this has been done with any significant budget. RePEc benefits from sponsored hardware in various locations but has no paid staff. The only exception is a programmer working on the RePEc Author Service under a grant from the Open Society Institute. The software suite to manage this service is released for use by others. All services described above are managed by economists and librarians throughout the world, always as a side job, never paid except for very few teaching releases. All software used is open source as well: linux, perl, mySQL, among others.

But the data input, i.e. the collection of bibliographic data about over 300,000 items of research (and growing by several thousands a month) is done by a large network of local volunteers, graduate students, faculty, secretaries, or IT professionals, who just follow a simple framework to organize their bibliographic data. Their individual cost in doing so is small, but once they realize their benefits in the circulation of their works, they are ready to do so.

RePEc has become very popular in the Economics community for several reasons: the first is that it is freely available to anyone. Thus academics in developing

countries whose library cannot afford a commercial indexing service just need an Internet connection to replicate a fee-based service. Others in richer areas just like the convenience. Second, RePEc does not just index printed research but has a very substantial collection of pre-prints. Third, RePEc can be much more timely than another other service: in most cases, a text is integrated in RePEc within a day of being added to an archive. The decentralized nature of RePEc makes this possible.

5   Can this Model be Applied Elsewhere?

In principle, an open library of academic documents could sustain itself in any discipline. The imperatives of publish or persish play out in similar ways across discipline. While not every discipline has a preprint culture—in fact Economics, Mathematics, Physics, Computer Science and Physics are the only ones with a preprint culture of note—all have some form of informal communication. Conferences play a key role in many disciplines. Funding application documents can also be of interest because they forecast future research activity. Some disciplines have built shared input facilities, such as gene databanks for example, that could act as a starting facility for shared research output.

What one needs to start with is a single dedicated volunteer per discipline. That volunteer needs to be a discipline insider who has

- an excellent grasp of information technology, including XML and scripting languages;

- an ability to grasp the incentive structures and publication habits in the discipline;

- a vision of what the library will be;

- an unerring determination to build it;

- and the ability to communicate all the above to others in order to recruit volunteers for the tasks ahead.

This is a tall order, indeed. It is not likely that it will come along in many areas in the near future. But the need for academic evaluation will not go away. There will be no automated technology that can do it. No automated technology will be able to distinguish two authors, or two institutions.

One important reason for research evaluation is research policy. National funding for education and research has to be based on an implicit or explicit research policy. Some national evaluation schemes, such as the UK Research Assessment

Exercise, put the onus on universities to gather data. This data is limited in scope and in scale. If the model could be reproduced in other countries and coordinated on a world-wide scale remains to be seen. Thus the role of governments as an external force to create open academic digital libraries remains limited.

Some people in the library community feel that it should be a driving force in scholarly communication change. Unfortunately, we are not aware of any library based project that has a great deal of impact on academic scholarly publishing. While there are some fine library-based databases, such as PubMed and ERIC around, they are centrally funded and stay at the library side of the business. That is, they are essentially catalogs of resources. If there is just one special aspect of the RePEc project, it is its insistence on "non-resources", i.e. the authors and the institutions. If we want to get open digital academic libraries to go in other fields, we cannot do it by limiting ourselves to resources. It is only when authors and institutions are documented that they really start to make serious efforts in participating.

One thing where the RePEc has influenced the wider library community is through the Open Archives Initiative's Protocol for Public Metadata Harvesting (OAI-PMH). Work on the OAI-PMH had its origin in an effort coordinated by Herbert Van de Sompel, Thomas Krichel and Michael L. Nelson to build a cross archive eprints service. Basically, the OAI-PMH is a more general version of the Guildford protocol for a situation where XML Schema is used to describe metadata formats. This XML technology was not available at the time of Guildford protocol. A technical protocol has significance beyond its technical remit. In the case of OAI-PMH, it can be hoped that the fact there is a standard technology to harvest metadata will encourage more people to share metadata. Such shared metadata can be an important building block of collections such as RePEc.

In the meantime, a new project is under way to build a collection similar to RePEc for computing and library and information science. The title the project is written as "rclis" which is pronounced as "reckless". It's web page is at http://rclis.org. Computing is a "neighboring" discipline to economics, in the sense that, it too, has a preprint culture. But rather than being strengthened by the Internet, in the sense that departments will keep organized collections on the web, computer science has been weakened by the advent of the web. Departmental collections have lost importance. Researchers have made papers available through web sites. An automated system, CiteSeer, at http://citeseer.ist.psu.edu/ has been set up that automatically find papers on the Internet through the Internet. While this centralized systems is an impressive achievement, it does not provide for author registration or institutional evaluation. In addition, computer science has a large bibliographic database called DBLP, see http://dblp.uni-tier.de. But this is a strict bibliography. It does not contain abstracts and classification but contains accurate bibliographic

data that allows to identify documents. These data can then be used to look for the documents on the web, thus building a virtual archive for these papers. This virtual archive, again can be used to build a web service. Once/if the web service becomes popular, it can be used to perform evaluation of the usage of papers. The crowning finish will be an author service, using the same software that powers the RePEc author service.

6 Conclusions

We looked at open academic libraries, i.e. libraries that provide freely information about academic works. We showed that it is possible to create such libraries by following a model similar to that of open software where scattered volunteers help on a common project. We argued that there are strong incentives for the participant to make sure the library is accurate, especially the part that they are in charge of. While the benefit of open software is more diffuse, i.e. a contributor's reputation will not depend on how his scripting contribution will improve the software overall as it is difficult to identify everyone's contribution, in the case of open libraries the incentives are much clearer once some critical mass has been reached. Indeed, authors, the institutions they are affiliated with and the outlets where they publish all strive to have accurate and up-to-date information about their publications in the open library.

7 References

Arms, William Y. (1999). Digital Libraries, MIT Press, available at http://www.cs.cornell.edu/wya/DigLib/MS1999/index.html

Arms, William Y. (2000). Automated Digital Libraries: How Effectively Can Computers Be Used for the Skilled Tasks of Professional Librarianship?, D-lib magazine, available at http://www.dlib.org/dlib/july00/arms/07arms.html

Barrueco Cruz, José Manuel and Krichel, Thomas and Trinidad, Jeremiah C. (2003). Organizing current awareness in a large digital library. In Proceedings Conference on Users in the Electronic Information Environments, September 8–9, 2003 Espoo, Finland., pp. 1–19, Espoo, Finland, available at http://eprints.rclis.org/archive/00000368/

Barrueco Cruz, José Manuel, Markus Klink and Thomas Krichel (2000) Personal data in a large digital library, presented at ECDL2000, available at http://openlib.org/home/krichel/papers/phoenix.html

Hodgson, Geoffrey M., and Harry Rothman (1999). The Editors and Authors of Economics Journals: A Case of Institutional Oligopoly?, Economic Journal,

Royal Economic Society, vol. 109(127), pages F165–86, February.

Jacsó, Péter (2004). Péter's Digital Reference Shelf–December, Google Scholar Beta, December 2004, available at http://www.gale.com/servlet/HTMLFileServlet ?imprint=9999&region=7&fileName=/reference/archive/200412/googlescholar.htm l

Kocher, Martin G., and Matthias Sutter (2001). The Institutional Concentration of Authors in Top Journals of Economics during the Last Two Decades, Economic Journal, Royal Economic Society, vol. 111(127), pages F405-21, June.

Krichel, Thomas (1999). Guildford protocol, available at http://ideas.repec.or g/p/rpc/rdfdoc/guildp.html

Krichel, Thomas (2001). ReDIF version 1, available at http://ideas.repec.org/p /rpc/rdfdoc/redif.html