From Open Access to Open Libraries: Claims and Visions for Open Academic Libraries

Thomas Krichel
Palmer School
720 Northern Boulevard
Brookville NY11589-1300
krichel@openlib.org
http://openlib.org/home/krichel

Michael E.D. Koenig
Palmer School
720 Northern Boulevard
Brookville NY 11548-1300
michael.koenig@liu.edu

Introduction

This paper looks at the future of digital libraries for academic documents. By academic documents we mean those that writers do not expect to be paid for, that deal with topics that are so specialized that only a small audience is interested in them, and that don't contain advertising. Naïve economic intuition suggests that there should not be much money to be made out of publishing such documents. This intuition is quite wrong, of course. According to the Association of Research Libraries' "Create Change" web site at http://www.createchange.org, the primary publishing industry is worth several billion US dollar annually[1].

Recently, the primary academic publishing industry has become much preoccupied with the idea of open access to its output. Open access becomes possible as the Internet has reduced margin costs of giving out an additional copy to virtually zero. Very simple economic theory suggests that when welfare is optimal, prices are equal to marginal costs. Such a theory thus supports the idea that open access has favorable welfare properties.

Our paper however is not about open access to primary research papers per se. Despite all the hype about open access, we think that the access method is not crucial to the academic digital library. The essence of any digital library is not so much about the access to the documents; rather it is more about how such documents are organized. There has to be some human organization in the digital library. Let us call this idea "claim 0". In other words, this paper is about the secondary data relating to the research documents. Such data has historically been provided by the abstracting and indexing industry. We are interested in an abstracting and indexing equivalent of open access publishing. We examine the idea of the open library. For now, just think about the open library as an open access collection of metadata about documents.

One way to achieve an intermediate step towards an open library is to make all academic papers available through open access on the Web, and have their full text indexed by a Web indexing service such as Google. Let us call this the vacuum cleaner approach. Arms (2000) questions the wisdom of going beyond the vacuum cleaner. We agree that, if the aim is merely to provide access to documents, a simple computer-generated index will suffice for most user needs. The popularity of Google proves this point. However, recall that from claim 0, this will not be a digital library – the structure is missing. The question then remains if claim 0 with its structure is a useful point to make at all, i.e. is it worth aiming for?

To answer that question, we need to step back from the open access / open library setting

[1] If one, as a benchmark, adds up the 1997 net sales for Wolters Kluwer, Reed Elsevier, J. Wiley & Sons, Plenum, and Thomson (Wyly 1998), the total is 17.35 billion dollars per year. Granted, this is an overestimate, since not all the products of these companies represent scholarly publishing, but still, scholarly and professional publishing is what these companies are primarily known for, and to a degree offsetting that overestimation is that there are many other scholarly publishers (numerous but much smaller). Thus one may conclude that several billion US dollar annually is, if anything, likely to be an understatement.

to examine the final purpose of academic publishing be it primary or secondary. We claim that academic publishing is part of scholarly communication. As such, and this is our "claim 1", it serves essentially the interest of authors, rather than the interest of readers. It is crucial to the evaluation of academic activity, and makes or breaks academic careers. See Roosendaal and Geurts (1997) for a more detailed analysis of scholarly communication that corroborates our claim 1. Since anybody can put anything up on the Web, and since computers can not know who has put up what, claim 1 implies that the vacuum cleaner approach does not serve crucial functions of scholarly communication. Thus we need something beyond the vacuum cleaner. Ideally, we would like to achieve an open academic library. This basically is the idea of an open library applied to academic documents.

The Open Academic Library

The open academic library as we conceive it, is a freely available abstracting and indexing dataset. In its final stage, an open academic library contains descriptions of all academic documents, with every document linked to the publication channels (i.e. journal, conference etc) in which it appeared, with all the authors identified rather than simply named, with citation information that identifies each cited paper etc. Note that the access to the documents full-text many be restricted (recall that this paper is not about open access to primary documents). The library is be a dataset, presumably using an XML syntax, available to multiple user services via a range of application layer transports protocols such as the Open Archives Initiative Protocol for Public Metadata Harvesting or HTTP. Publishers contribute to the open library using a protocol that the open academic library has established. Use of the library is free for contributors and users. In this way the open academic library disseminates more effectively than any toll-gated counterpart. We can say that it has good dissemination properties. But dissemination is only one purpose for which an open academic library exists.

The problem with this vision of an open academic library is that it requires constant maintenance and human effort. This is obvious when looking at examples. Only a human can adequately and consistently determine if a document is part of a collection, i.e., it appears in a specific journal or it has been given at a certain conference. Only a human can decide if two collections are the same. This can be an ambiguous decision because of the rearrangement of journals over time, such as through name changes and splits. Only a human can decide if two authors of two papers are the same person. There are many other examples. The problem with constant human intervention is that it is very expensive.

Our main claim comes now. We claim that, despite the expense, it is possible and desirable to build open academic digital libraries. This is our "claim 2", our principal claim. We devote the remainder of this paper to arguments that back up this claim.

We proceed in two parts, by demonstrating two sub-claims. First, we assume that an open academic library exists, and we claim that it can be sustained. The argument in favor of the sustainability of an open academic libraries runs as follow. We already have established the effective dissemination properties of a freely-available database. Therefore we argue that publishers will want to contribute since the library provides advertising for products that they may want to sell. The dissemination effect of the digital library will pay for the efforts that have to be made to contribute to it. Authors will also want to contribute to the library by maintaining the structure that identifies authors and the institutions that authors work for. This structure can be used to prepare quantitative evaluations of an authors' work. This can be done through aggregating usage measures of individual documents for all the documents that the author has written. as well as traditional, i.e. citation based quantitative evaluations that one can construct

for the impact of the authors, and even, through aggregation of authors of an institution, the work and the impact of the institution. We can call this the accounting function of the academic digital library. This in turn implies strong incentives for authors and institutions outside the library to join, since if they do not participate, their scholarly contributions will appear to be non existent. By a similar argument, we can demonstrate that insiders have strong incentives to keep the information about them up-to-date. We recognize that individuals are frequently reluctant to participate in quantitative evaluation procedures, but recognize that within an institution there will be multiple levels of pressure points encouraging participation – team leaders, department heads, and deans will all want their domain to show well; it is not just that the institution wants to shine well on National Research Evaluations or on U.S. News and World Reports rankings. Thus once an open academic library is built to a sufficient scale, once it reaches a certain critical mass, it will sustain itself. Once providers follow some specific procedures for the formatting of the data, the remaining coordination work can be accomplished by a few volunteers. Indeed, we can foresee that there may soon come a time in which working as a volunteer coordinator in an open academic library carries a prestige similar to that of the volunteer who works to edit an academic journal.

Now that we have demonstrated the sustainability of an open digital library, we now turn to the second sub-claim. This deals with the chicken and egg problem that is present when the open digital library is started, i.e. will the open digital library scale to and beyond the point of critical mass? At the beginning, when the scale of the library is modest, the incentives for contributors are weak. If incentives are weak, contributions are likely also to be modest, and with only modest contributions, the library may not reach the scale required for sustainability.

Open academic libraries can nevertheless be built. Note that claim 2 does not claim that a single open academic library will be built, but that there may be many of them. We make no claim as to how many of them. The open academic library is a multi-faceted concept that may be built by assembling various independently constructed facets. To understand this idea of many digital libraries, let us go right back to claim 1. Recall that by virtue of claim 1, scholarly communication is author-driven. But authors don't act in isolation. The very idea of “scholarly communication” implies that there are several parties involved. Each author, however, communicates predominantly only with other authors that are within some similar subject area. Thus, authors are surrounded by a fuzzy set of other authors (or potential authors, such as students) that are potential readers of their output. Thus, dissemination does not need to go very far, it is just sufficient that the library disseminates to the right people. A similar argument can be thought of when looking at the evaluation function of the academic library. Evaluation of academic work is usually not done with respect to all other academic work. Rather evaluation is conducted with respect to academic work that is comparable. A similar fuzzy set of comparable documents exists for each document. We can call the union of some aggregate fuzzy sets of authors and of some fuzzy sets of documents a discipline. Thus it is sufficient to organize an open digital library at the scale of a discipline.

A discipline is a community of authors and their institutions held together by a set of documents. Over many years, some disciplines, such as chemistry and history, have established themselves quite clearly. However it is very difficult to split academia into a set of non-overlapping disciplines. Thus the existence or non-existence of an open academic library will be, in part, due to the arbitrary choices being made at the outset of an open academic library. Thus, to prove that open academic libraries can be built, it is sufficient to find an effective example for one discipline, particularly a digital library that has scaled up almost entirely through volunteer efforts within the community and without large infusions of grant and financial support. The RePEc digital library is precisely such an example.

An example open academic library: RePEc

RePEc is an open academic library for economics. Economics is a discipline with a working paper culture. Working papers are early accounts of recent research results. They are issued by academic economics departments and other institutions that do economics research, such as central banks, for example. In the print days, working papers were circulated by exchange arrangements between issuers. Electronic dissemination started in April 1993, when Thomas Krichel put the first every electronic working paper in economics out on a Gopher server. This was the start of collection of electronic working papers that he kept as a hobby project. For some of them, he had the full-text for, for others he linked to the full text. This small collection, named WoPEc, was complemented by a much larger collection of bibliographic references to working papers provided by Fethy Mili at the University of Montreal. This collection was called BibEc. BibEc and WoPEc continued to be the largest access points for economics working papers in the Internet until 1997. In that year the RePEc project was founded whereby BibEc, WoPEc and a small number of other initiatives created a platform to exchange data. The data was encoded in a purpose-built format called ReDIF, see Krichel (2001), and exchanged using a purpose-built transport protocol called the Guildford protocol, see Krichel (1999). Both are still used in RePEc. At the time of writing, in June 2004, RePEc describes 270,000 items of interest to economists. There are over 175,000 of these available online. There are130,000 working papers, 139,000, journal articles, 1,000 software components and 750 book and chapter listings described. These data are provided by close to 400 RePEc archives that provide ReDIF data laid out according to the Guildford protocol. In addition, RePEc contain 5,000 author contact and publication listings, and 8,000 institutional contact listings. The latter two are the crucial components. The institutional contact listings are comparatively stable, and can therefore be compiled by one person, Christian Zimmermann, an economics professor at the University of Connecticut. The author contact and publication listings come from the RePEc author service. The RePEc author service replaces an earlier service called HoPEc, see Barrueco Cruz et al (2000). Authors provide the service with contact data, and then find the articles that they have written that are included in the database. At the time of writing, over 75,000 items have at least one identified author.

RePEc data is used in many services. We will mention here only the most important ones. EconPapers at http://econpapers.hhs.se and IDEAS at http://ideas.repec.org are web interfaces to the entire RePEc database. “NEP: New Economics Papers” at http://nep.repec.org is a human-edited current awareness service for new working papers in RePEc. LogEc at http://logec.hhs.se shows data about the popularity of items in RePEc. These data have been compiled from all RePEc user services, and access by autonomous agents such as web crawlers has been removed. RePEc archive maintainers receive usage data for the documents in the archive that they maintain. More importantly, RePEc author service registrants receive a monthly email with the usage data for the papers they claim to have written. The flurry of registration updates that follow the sending out of email shows that the authors take active note of the usage data that they are getting.

Given the crucial importance of RePEc author service, it comes as no surprise that there is an active development plan for it, funded by a grant from the open society institute. Plans include allowing authors to submit documents to it, as well as allowing authors to associate with citation data. Autonomous citation data for open access documents in RePEc is provided by the CitEc project at http://netec.ier.hit-u.ac.jp/CitEc. Note however that this grant support has been received only after RePEc has established itself as a viable open academic library, and that the development work will be available for other open academic libraries. That observation leads to

two important corollary points:

- Each facet of the open academic library makes the creation of other facets easier.
- The facets however need to be constructed similarly from open source software.

Conclusions

Let us start the conclusions by examining the consequences of open academic libraries for digital libraries in general. As a new field digital libraries take much of their inspiration form the world of physical libraries. Much attention is spent on the interaction between users and documents and providers of documents. In some way, this is limiting We speculate that in the future, digital libraries will be more like interactive communication tools that allow users to interact with each other through document data. The document will loose its central role, and the distinction between users and providers will become more blurred. This is exactly the case in RePEc through the RePEc author service.

Let us close the conclusions by thanking you, the reader, for having read this far, unless, of course, you have jumped to this conclusion, in which case we urge you to read from the beginning. You have paid for this paper with a very valuable commodity, your attention. As you have a finite lifespan and you can not change the past, the attention you paid to this paper can not be reproduced, it is a sunk cost. The same argument can be applied to abstract and indexing databases. Such databases essentially advertise academic work. We expect that soon there will be new abstracting and indexing databases that will outperform the existing ones in terms of both accuracy and timeliness and yet be freely available. Even controlled-vocabulary indexing could be provided in a decentralized fashion provided the community can agree on a vocabulary to use. Overall, free availability goes hand in hand with accuracy and timeliness, because the free availability of the data ensures good dissemination and the prospects of good dissemination ensure accurate work of the providers. Open academic libraries will revolutionize scholarly communication.

References

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, Jose 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

Krichel, Thomas, (1999) "Guildford protocol", available at ftp://netec.mcc.ac.uk/pub/NetEc/RePEc/all/root/docu/guilp.html

Krichel, Thomas (2001) "ReDIF version 1", available at ftp://netec.mcc.ac.uk/pub/NetEc/RePEc/all/root/docu/redif_1.html

Roosendaal, Hans E. and Peter A.Th.M. Guerts, (1997) "Forces and functions in scientific communication: an analysis of their interplay", available at http://www.physik.uni-Oldenburg's/conferences/crisp97/roosendaal.html

Wyly, Brendan J. (1998) "Competition in scholarly publishing? What Publisher Profits Reveal" ARL Newsletter, # 200, available at http://www.arl.org/newsltr/200/wylytable.html