Electronic Publishing '01 – 2001 in the Digital Publishing Odyssey
A. Hübler et al. (Eds.)
IOS Press, 2001

Disintermediation of Academic Publishing through the Internet: An Intermediate Report from the Front Line

Thomas Krichel,
Palmer School of Library and Information Science
Long Island University, 720 Northern Boulevard, Brookville
New York 11548-1300, USA
krichel@openlib.org

Simeon Warner,
MS B285 Los Alamos National Laboratory
Los Alamos, New Mexico, 87545 USA
simeon@lanl.gov

Abstract. There has been a lot of discussion about the potential for free access to scholarly documents on the Internet. At the turn of the century, there a major initiatives. These are arXiv, which covers Physics, Mathematics and Computer Science and RePEc, which covers Economics. These initiatives work in very different ways. This paper is the fruit of collaboration between authors working for both initiatives. It therefore reflects the perspective of people working to achieve change, rather than an academic perspective of pure observation.
We first introduce both arXiv and RePEc, and then consider future scenarios for disintermediated academic publishing. We then discusses the issue of quality control from an e-print archive point of view. Finally, we review recent efforts to improve the interoperability of e-print archives through the Open Archive Initative (OAI). In particular, we draw on the workshop on OAI and peer review held at CERN in March 2001 to illustrate the level of interest in the OAI protocol as a way to improve scholarly communication on the Internet.
This paper is available online at http://openlib.org/home/krichel/sants.html

0. Introduction

The exchange of contracts over the Internet is now commonplace. In the majority of cases, delivery of the merchandise occurs off-line. However, purely informational commodities--such as statistical data or pornographic pictures--can both be contracted upon and delivered over the Internet. That affords the holders of such commodities the opportunity to directly contract with customers in the way that was not possible off-line. The new medium thus provides an opportunity for disintermediation.

In the academic world, the debate about the possible extent of disintermediation has concentrated on the rôle of academic publishing. A large part of academic writing attracts no payment from publication. For the sake of simplicity, this paper deals exclusively with scholarly works for which the author receives no payment. These will be referred to as "research papers" or "papers" for short. It is further assumed that the advent of the Internet will not change the reward structure in the academic world. We assume that authors will still want to write without payment, with the exclusive aim of achieving peer recognition.

It has been proposed, most vociferously by Harnad (1995) and in many papers since, that the Internet will lead to free access to academic papers. In 1995 two large-scale initiatives

were already well under way to work on realizing this vision. These were the arXiv archive for Physics, and the RePEc dataset for Economics. This paper has been written by people working within these initiatives. It is a contribution to the debate about "freeing the literature". The volume of papers on this topic is already quite large. What justifies a new paper on the topic?

The process by which the academic literature is to be freed is most often referred to as author-self archiving. This is somewhat of a misnomer, but we will stick with the term for a moment. This paper has been written by practitioners of author self-archiving. Thus it offers practical insight from the process that are often not available to the outside observer. More generally, we have also been involved in recent efforts to coordinate author self-archiving initiatives in the Open Archives Initative. We will therefore take a look beyond the day-to-day running of our work and propose some steps that would bring the process forward. Thus we bring two innovative perspectives to the literature, a practical perspective and a normative one.

The remainder of the paper is organized as follows. In Section 2 we introduce the arXiv archive, and in Section 3 we discuss the RePEc dataset. In Section 4 we suggest scenarios for future Internet-based publication. In sections Section 5 and Section 6 we discuss steps forward and setbacks for the author self-archiving movement. In Section 7 we discuss the question of quality control. In Section 8 we present recent work of the OAI. In Section 9 we offer a few conclusions.

1. The arXiv archive

The arXiv e-print archive is the largest and best-known archive of author self-archived scholarly literature. It is discipline-based and centralized, in that all submissions and the master database are at one site. Here we briefly review the history of arXiv and then discuss its place in the scholarly publishing environment.

The arXiv e-print archive, formerly known as xxx, was started in August 1991 by Paul Ginsparg. The first subject area was theoretical high-energy physics, 'hep-th' and it was initially little more than an email reflector for paper exchange. In 1992 an ftp interface was added and the archive expanded to other areas of high energy physics. Development since then has seen steady expansion of the subject areas to cover all of physics, mathematics, computer science and non-linear systems; and the addition of new interfaces and facilities as new technologies have become available. Significant developments have included: web interface (December 1993); automatic PostScript generation from TeX source (June 1995); PDF generation (April 1996); and web upload (June 1996). Recently, arXiv has been at the forefront of the Open Archives Initiative (OAI).

arXiv serves an estimated 70,000 users in over 100 countries. There were about 13,000,000 downloads of papers in 2000. There were over 30,000 submissions in 2000, and the number of new submissions is growing approximately linearly, with about 3,500 additional new submissions each year. The growth in submissions has had little impact on the administration of the arXiv although there is an increasing reliance on moderators to assure appropriateness. More that 98% of the submissions are entirely automated: 68% of them via the web, 27% via email and 5% via ftp. It is interesting to note that if just 15 minutes were required for each submission then a full-time staff of 7 people would be required to deal with new submissions. Instead, arXiv uses less than one full-time equivalent to deal with day-to-day operations.

The Los Alamos site is funded by the US National Science Foundation and the US Department of Energy. The 15 mirror sites around the world are funded independently, the

cost is just a few thousand dollars for a new machine every few years and a small amount of system administration effort.

The high-energy physics community uses the TeX text formatting system almost exclusively, and this has been very convenient for arXiv. arXiv does not accept preprocessed TeX submissions, authors must submit the source. This allows generation of various types of output including DVI, PostScript in several flavors, and PDF. Expansion into other areas of physics means that there are now an increasing number of non-TeX submissions and this trend is sure to continue. Unfortunately, many common word-processing packages produce very inefficient and sometimes low-quality output unless used expertly. Experience shows that PostScript or PDF submissions require greater screening effort than TeX submissions. This is an example of how the physics and mathematics communities differ from other communities in a way that has favored author self-archiving.

Before 1994, archives for some new subject areas were started at other sites using the same software. In November 1994 the data from the remote sites were moved to the central site, and the remote sites became mirrors. The reason for this reorganization was the need for stability. While a distributed model is appealing, sites relying on a single person, who's main occupation is not the archive, are not stable on a time-scale of years. A user perception of stability is important for the success of an archive. arXiv is once again investigating a more distributed model but any new system will be robust to the disappearance of any site.

arXiv has not been an academic exercise, it was started from inside the community it aimed to serve. At all stages of expansion to cover a wider subject area, arXiv has been guided and promoted by members of the new fields. Nowadays, some commercial publishers tacitly acknowledge the legitimacy of arXiv by accepting submissions where the author simply quotes an arXiv identifier. Policies vary on whether the publisher permits the author to update the arXiv version to reflect changes made during the refereeing process. However, authors often ignore any prohibitions.

arXiv is by far the largest author self-archiving project. Some proponents of change in scholarly publishing have heralded arXiv as the single mode forward for the change of scholarly communication. While it is possible that the arXiv model could be successfully applied to all other disciplines, there are good reasons to question this. We will highlight two issues.

First, there have been attempts to emulate the success of arXiv by building discipline-based archives for other disciplines. Two working examples are CogPrints at the University of Southampton and the Economics Working Paper Archive. Both have been operating for more than four years, yet neither has grown beyond 1,500 documents. Why have they not been more successful? This leads us to question the applicability of the centralized scheme to other disciplines. It should be noted that both initiatives have been run by members of the community they serve. Therefore they can not be dismissed as failing because they are foreign to the discipline and not sensitive to its particular needs.

Second, even though arXiv has been successful in attracting submissions in several areas of physics and mathematics, an attempt to expand to computer science has had limited success. A high-level committee on the future of scholarly communication in computing entered an agreement with arXiv to run a Computer Science Research Repository (CoRR). This has, to date, received rather little attention from computer scientists in spite of various publicity efforts within the community. Are we to conclude that computer scientists care little about formal on-line dissemination of research papers? Are they simply happy to leave their papers on their homepages?

If current trends continue, arXiv will provide and increasingly complete free layer of primary research literature for the mathematics and physics disciplines. This leads us to

ask whether it is likely that, over time, small-scale archiving, be it in departmental or homepage archives, will decline. There is certainly a trend for departmental archives in mathematics to migrate to arXiv. However, the Open Archives Initiative (OAI) has the potential to change the tradeoffs in the choice between centralized and decentralized archives. If cross-archive discovery and awareness tools exist for decentralized archives then institutional and department archives may become as effective as centralized archives such as arXiv.

2. The RePEc database

RePEc is much less known than arXiv and it is also less well understood. There are two reasons for that. First, it is limited to the economics discipline. Second its business model is more abstract.

Historically, RePEc grew out of the NetEc project. It was started by Thomas Krichel in February 1993. In April 1993 he opened a tiny collection of a electronic papers on an ftp server with a gopher interface operating at Manchester Computing. However, it was never the business plan of NetEc to become an electronic working paper archive for the discipline. Instead the idea was to collect data about printed and electronic papers as published by various sites so that they could be searched together on a single site. It was a library of bibliographic data rather than a archive. The project developed mirror sites in Japan (July 1996) and in the US (January 1997). In 1997, the RePEc dataset was created by the NetEc project, and two other projects that were active in the area, DEGREE and S-WoPEc. These projects agreed to exchange metadata in a common, purpose-built format called ReDIF. This metadata can be harvested following a simple harvesting protocol called the Guildford protocol. Shortly after the implementation of the protocol, several user services appeared that were build on the data. Thus the RePEc project is a true forerunner of the Open Archives Initiative. In 1998, data about Economics institutions was In 2000, a personal registration service was opened. It allows persons to register associations between them and the document and institution data in the database. At the time of writing, it has over 160 different archives that contribute metadata, and eight different user services. There are about 40,000 downloadable paper cataloged in RePEc. The NetEc project received £129,000 funding from the Joint Information Systems Committee JISC, it now runs without any external funding. Running such a large-scale operation with volunteer power only is a remarkable technical and organization achievement.

RePEc is not only a database of papers in economics, but it also contains data about economics institutions and academic economists.

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.

At the moment, EDIRC is mainly linked to the rest of the RePEc data through the HoPEc personal registration service. This service can be used by economists to register themselves as authors of the documents that are contained in RePEc. To date 6% of all papers have an at least one of the authors as a registered person. The HoPEc registrations will in the future

be used for building a collection of papers held in the homepages of these authors. Already now, the collection is used to link from the papers of authors to their homepage and for the provision of current contact information. Clearly, current contact information may be quite different from the contact information in the bibliographic data.

Thus RePEc appears as a very ambitious project. In its end stage, every author, institution and document in economics will be registered in the database. What are the chances that this project be completed? Two remarks are on order here.

First, while RePEc have rejected centralization of papers as a means forward for the collection of grey literature, but it happens that RePEc has a commercial competitor in the "Economic Research Network" operated by "Social Science Electronic Publishing". Much of the offerings on their site are free. There is a substantial overlap with the contents of RePEc. At the time of writing, they have reached about 12,000 electronic papers. Thus they are still much smaller, but they seem to be expanding at a faster rate. There is two reasons for the success of the company. First, they have extensive statistical reports on download of papers, something that is next to impossible to collect in the context of a multi-service operation.

Second, the company has started cooperative agreements with departments who wish to sell the working papers that they produce. If this trend expands we will see under the Internet access to things that used to be free in the paper days. Second, while are some commercial publishers that already collaborate with RePEc, it is not easy to forecast that there will be wide-spread collaboration. There are two reasons for that. First in economics, no publisher holds a large amount of material. Many publishers produce economics journals, but for each publisher it is a small amount of their business. Therefore the cost to working with RePEc-specific bibliographic system is relatively high as compared to the returns. Therefore the contents of RePEc is still dominated by working papers.

3. Future scenarios of publication

In the introduction, we pointed out that the term "author self-archiving" is somewhat of a misnomer. In the arXiv scenario it is arXiv that do the archiving. In the RePEc scenario the archiving is performed by the intermediate providers like departments and research centers. To be quite precise, "author self-archiving" would be more appropriately applied to authors depositing their papers in homepages. However, this type of activity can hardly be classified as archiving because the concept of an archive suggests a long-run storage facility which is rarely the case with personal homepages.

Thus we would like to introduce the concept of academic self-publishing as a replacement for author self-archiving. By academic self-publishing we mean a trend for the academic sector to take over the publication of its own output. It should be noted that we understand "publishing" here in the wide sense of the word as "making public", rather than in narrow sense of "making public through an intermediate agent to whom copyright is being transfered". Both RePEc and arXiv fit this description.

Academic self-publishing of research material is currently being experimented with on a per-discipline wide basis. A useful discussion of same initiatives is Kling and McKim (2000). We agree with the basic message of their paper. There will be no unique method by which networked scholarly communication will be running across disciplines. Discipline-specific differences are likely to persist. For each discipline, it is most likely that evolution depends on

- the established communication patterns prior to the proliferation of electronic networks

- the presence or absence of entrepreneurial pioneers to stimulate change
- the political and financial environment of the discipline
- in that order. The persistence of discipline-specific features in academic self-publishing is the first assumption for our discussion here.

A second assumption is that authors will continue to write research papers. These are discreet accounts of individual research findings written by one author or a small group of co-authors. Thus we reject the idea that in the future, work of individual authors becomes dissolved in a global scientific hypertext net. It this case it would be difficult to trace the responsibility for the research results. Academic reward structures depend on precise knowledge of the authorship of research papers.

A third assumption is that the Internet will not go away. It will always be available as a medium for the distribution research papers. Therefore we assume that a free layer of research papers will always exist, even if sparsely populated.

A fourth starting assumption is that in all disciplines some form of toll-gated access to research papers will survive. This assumption can be justified by the idea that there is little chance that all the teaching material will be made freely available. In all disciplines authors of teaching material have always received a monetary reward. Since the border between research and teaching documents is fuzzy, it appears unlikely to us that there will ever be a free access to all research documents.

Having made these starting assumptions, we have two points still to determine

- the extent of coverage of the free layer
- the size of any quality gap between toll-gated and non-tollgate papers.

Here we propose three terminal scenarios. Individual disciplines may adopt one scenario or develop a mixture between scenarios.

In scenario one, there is a free layer of research documents deposited on the web by their authors. They may be withdrawn at any time. There is no bibliographic organization of these papers other than ones organized by machine. In addition, since these papers are in places where they can be modified by authors, it does not appear to be possible to base a certification system on these papers.

Papers in the free layer can be found through general Web search engines, or possibly through a specialized engine like inquirius. In addition, there will be a toll-gated layer of quality controlled, final publication. It will have a good bibliographic description but there will be no common catalog in the public domain. Thus there will be quality controlled libraries that be behind toll-gates.

Although this scenario may appear unlikely, it has been defended by Arms (2000). He imagines the co-existence of an expensive layer of a research library that is powered by humans, with the extensive quality control of the data, and a free layer that is essentially computer generated. Author pressure, he speculates, will make a lot of research papers openly available. But the bibliographic layer, since this is costly to produce, is not likely to be free. Some elements of the construction of the free interface can not be fully automated. This for example concerns the removal of duplicates, dealing with moved, changed or merged collections, maintaining a correct set of author data etc.

In scenario two there will be the free layer of research papers and there will be a free bibliographic system too. The layer of toll-gated and free papers will be decentralized. The bibliographic system will comprise both free and toll-gated papers and indications on which servers they live. The bibliographic layer may be centrally or decentrally organized. In a centralized system all providers of papers will submit to a central index of all papers.

It is an open question how such a server may be organized and funded. In a decentralized bibliographic system, all sites that provide papers will have a metadata layer that can be harvested through a machine interface. In this scenario the problem of funding a central agency does arise only to a small extend. The only central agency that we think will be needed is one that registers participants in the catalog exchange scheme. These participant may either be providers of catalog data or providers of services using the catalog data.

In scenario three there will be a large central site of papers, which comprises the vast majority of papers in a discipline. The quality control of the papers will be done in overlay services. These overlay service may contain links to papers on the central site that have been found fit for inclusion in the overlay service. It may also contain reviews of the free material. We would expect that most overlay services are free but that is not crucial to the scenario.

To compare the three scenarios, let us stress that we imply that there is an implicit quality ladder. Scenario 1 is the least controlled, scenarios 2 and 3 improvement upon this because they allow for better information retrieval and better facilities for quantitative evaluation of the research process. Scenario 3 improves over scenario 2 because it makes the long-run preservation of the archived data much more secure. A centralized system would be likely to find it much easier to implement format conversion than a decentralized system.

4. Steps forward

Having established terminal scenarios we will discuss steps that may lead to them. If nothing is done, by default, we will have scenario 1. The academic sector may find ways to implement scenarios 2 and 3, but it is not clear today what should be done. In this section, we make some suggestions.

First, we believe that change to the scholarly communication will be discipline community driven. Anyone proposing a new service should make sure that they receive the maximum backing of the community of the discipline that they are supposed to serve. The successful services have managed to do that. These statements may say the obvious, but there are examples where such elementary advice has not been followed. The most prominent is Varmus (1999), the initial call for the creation of PubMed Central. This call was clearly a simple translation of the principle of arXiv to the biomedical field. Little thought was given to adapting the arXiv business model to the needs of the biomedical community. Much of the negative reaction to the proposal would have never occurred if its discipline-specific implications would have been thought out before.

Second, proper attention has to be played on the motivation of contributors. Too many schemes have been proposed where a lot of effort has been spent on developing a user interface, but little effort is spent on population the services. This holds in particular for projects that are funded by the library community. Libraries are accustomed to address end-users, and therefore do not pay much attention to the question of contributors. There has been general agreement that public access to scientific documents does wonders to the exposure that these documents receive. But this additional exposure has to be demonstrated to contributors. If it can not be demonstrated, with figures at hand that organized contents is better than unorganized contents, then it is difficult to see much contents organization will taking place.

The first step forward must be the collection of contents. Any collection of contents that is organized with some metadata, as long as it is large enough, will stimulate interest. The problem with most of the existing attempts is that the contents is simply too thin. To stimulate the provision of contents, the collection must be conceived as an advertisement

of the work of its contributors. To do that, the contributing person or institution must be at the center of the collection effort. Again, we need a break with the tradition of libraries. In a library setting, the work is at the center of the descriptive effort. In an academic self-publishing setting, the contributor, rather than the works created by the contributor must be the center of the collection effort.

For most disciplines, it should be possible to gather free research papers that are available on the Internet in a gateway catalog. All disciplines have some form of informal communication channel, and many of these papers in these channels can be made freely available if somebody is willing to "put them up". If a sufficient number of papers in a small discipline can be made available, this can really change the way the sub-discipline works. Such small-scale efforts, opened up to the wider academic word through on OAI compliant archives would be an important step forward. Unfortunately, the current climate in academic institutions is such that, while writing papers is valued, collecting and organizing them is often not valued so such work is not incentive compatible.

5. Steps backward

There are a number of failed collections whose ruin can provide valuable insights into why collection efforts fail. An interesting example is the NCSTRL collection of Computer Science Technical Reports. On that site, we read

> NOTICE: This site is being maintained on a legacy basis and support for it is minimal. As a result, some of the functionality is limited. We will be posting a message in the near future about the future status of this site. Posted 2000-03-05.

This message that was supposed to be forthcoming is not found on the site. Despite watching relevant mailing lists closely, the authors have not found that announcement. It seems that NCSTRL, if not terminated officially, is in serious crisis.

The failure of NCSTRL is worrying for two reasons. First it is a failure of work for the Computer Science area. Many digital library efforts find it very hard to access appropriate computing expertise at times when such expertise is in high demand by other sectors that have more financial resources. But such expertise should have been plentiful at Computer Science departments. Second, the failure comes after years of development and on a collection of considerable scale. NCSTRL started as early as 1993 and has about 10,000 papers available in electronic format. After arXiv and RePEc, this is the third-largest collection of freely downloadable documents.

We suspect that the failure is not of a technical but managerial nature. As outsiders, we can only make some speculations about what the problems are.

One possible explanation in the double nature of NCSTRL. It was conceived as a digital document collection and a digital library test bed. It those circumstances, there is constant experimentation. NCSTRL worked on an elaborate protocol called Dienst to run the collection. Dienst was a language with many instructions call verbs. To allow several versions of Dienst to run on servers, each verb in Dienst had a version number, and more recent versions of the Dienst software could still execute the old verbs. Nevertheless in order to install more recent versions of Dienst software, archives had to do extra work. NCSTRL could not simply be a service without research components without compromising the funding position of the collection. ArXiv has continued funding as a service, RePEc has no funding and is vulnerable because of that.

Another possible reason for failure may have been the NCSTL decentralized architecture. Many participating sites were staffed by graduate students. This implies discontinuities in archive management. This is also a problem for RePEc but the technical architecture of RePEc is much simpler, and therefore the problems are much smaller.

To sum up this discussion, it is crucial that the business plan for any collection effort is well thought out. Everything else depends on it. The problem with author self-publishing in academic circles is that many academic institutions do not have a good record of enterprise, in particular in cases where the activity is unlikely to generate much revenue.

6. The question of quality control

Defenders of the status quo in publishing defend the system because of the quality classification of standard papers that it affords. A free publication system must consider the feasibility of filtering in a world where anybody can publish anything.

To tackle the problem of quality control, our experience from managing self-publication systems suggest a simple model of paper quality. We defined just two levels of quality: standard and inadequate.

Inadequate papers are so poor that almost everybody with a Master's level degree in the discipline, who would read these documents would become aware that they simply do not contain scientific material as it is recognized within the discipline.

Standard papers are the remaining papers, the majority of which will be unremarkable. A few will be read by a sizable group, even fewer will be cited in the future, and still fewer will be considered seminal works.

We claim that it is trivial for a member of the appropriate disciple to separate inadequate papers from standard papers. We have a moderation process instead of the conventional refereeing process. On the other hand, distinguishing between different shades of quality in standard papers is very difficult and costly in terms of time and effort. In the next Section, we will show how the OAI protocols can be used to implement a technical infrastructure that would be useful for this process. In the remainder of this section, we will look at how the filtering of inadequate submissions can be automated.

The object of automated filtering is to reduce the cost of the moderation process which still maintaining sufficient quality control. One approach is to allow submission based on institutional affiliation; our experience shows that a paper from Cornell University is likely to meet the requirements of standard quality. In this approach might also assume some institutional responsibility for submissions. However, in the construction of such systems it is important to cater for cases where unaffiliated authors wish to submit work that meets the requirements of standard quality. It would be a retrograde step for the author self-publication movement to be more prejudiced than the existing system.

Another way to function is to construct registries of personal and registration information for academics who are allowed to publish on open e-print archives. Depending on the discipline these registries might be administered by scholarly societies, academic institutions or the archives themselves, with policies reflecting the views of the community. A system which avoided duplicate registrations would be labor saving for both maintainers or the registries and authors who would have to keep their information up-to-date. A shared registry would have several other benefits. It could contain contact data shared by several organizations in the academic sector, such as scholarly societies. At the moment each scholarly society maintains its own registry of members, with a shared set of personal data, a society would simply maintain the handles of it members in the shared personal dataset.

There could be a mix of these approaches where personal registration might be based partly on the affiliation of a person with a recognized institution.

7. The Open Archive Initiative

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. This work is now called the Open Archive Initative (OAI). When the OAI started, there was little interoperability between e-print initiatives and cross-archive resource discovery was the focus of the initial meeting in Santa Fe. The scope of the initiative has expanded considerably since then.

The basic business model proposed by the OAI was inspired by the RePEc project. In particular, the Open Archives Metadata Harvesting (OAMH) protocol separates data provision and service provision as pioneered by RePEc since 1997. The OAMH protocol is a simplification of the subset Dienst protocol experimented with after the Santa Fe meeting. It was designed to provide a very a low-barrier to interoperability. Key features include: support for multiple metadata formats, requirement for Dublin Core (DC) metadata as means of global interoperability, use of Hypertext Transfer Protocol (HTTP) transport, and use of Extensible Markup Language (XML) encoding. There are four means to select between records

- Datestamp of a record
- Identifier of a record
- Metadata format(s) available for a record
- Sets, which are groups of records

There are three commands to support other protocol requests

- *Identity* gives basic information about an archive
- *ListMetadataFormats* shows the metadata formats that are supported
- *ListSets* shows the sets

There are the three other commands to access the records

- *ListRecords* returns records
- *ListIdentifiers* returns identifiers only
- *GetRecord* returns a single record

This protocol has attracted a lot of interest from digital library communities. It is hoped that it will provide a framework for the solution of many interoperability issues in digital libraries.

We conducted a survey of registered OAI repositories on 8 March 2001 to see how many data providers were operational less than 2 months after the OAMH protocol was announced. In the list below we show the name of the archive and the number of record identifiers returned from a *ListRecords* request.

name	*number*
arXiv	155522
OCLC Theses & Dissertations Repository	102762
NACA	6352

M.I.T. Theses	5196
Oxford Text Archive	1290
Perseus Digital Library	1030
CogPrints	1028
NSDL at Cornell	870
PhysNet	472
Humboldt University of Berlin	464
Resource Discovery Network	388
University of Tennessee Libraries	201
Linguistic Data Consortium	216
European Language Resources Association	183
A Celebration of Women Writers	142
The Natural Language Software Registry	78
California Digital Library	3

The OAI sponsored a workshop on OAI and peer review in Europe between March 22 and 24 of 2001. In was organized by the Access Division of LIBER, the Ligue des Bibliothèques Européenne de recherche. One working group, led by Herbert Van de Sompel, discussed ways in which the OAI framework could be used in the certification , by peer review or some other method, of academic works. We outline some of the ideas from this discussion in the remainder of this section. We hope that it will illustrate the potential of the OAMH protocols for more than just the resource discovery aspect of scholarly communication.

Let us first consider what certification means in the current environment of traditional paper journals. The type of certification is usually peer-review and the indication of this is publication in a particular journal. The journal reference is in essence a certification metadata for the work even if it exists somewhere other than in the journal, for example the author's homepage or in some other archive. We will not here consider the issue of whether this other copy is true to the certified version. The journal publisher provides, in addition to other things, a certification service.

The simplest scenario to consider is the certification of a paper that is first placed on an e-print archive. The certification process could be initiated by pull from the certification service or push from the archive, but is more likely to occur by direct author action (such as an e-mail). If the paper exists on an e-print archive then the certification service can retrieve it, and also harvest a metadata record using the OAMH protocol. After or during the certification process the certification service could expose certification metadata using the OAMH protocol. This could be harvested by the original archive and associated with the e-print provided it contained the identifier of the e-print. The American Physical Society (APS) already accepts submissions from arXiv (by quoting the e-print identifier). There was agreement between the APS and arXiv to test this scheme.

Given that some OAI data providers might then have records for both certified and non-certified papers, how could one harvest information about only certified records within OAI? Technically, there are several ways that this could be accomplished. One way is through multiple metadata sets, perhaps one for discovery metadata, one for the certification, etc. Using OAMH protocol, it is then possible to harvest metadata only for the certified records by requesting the certification metadata set (objects that do not have that metadata will not be returned). Another approach would be for the data provider to expose certified and non-certified sets which could then be harvested separately.

The group also discussed the contents of the certification record. It was thought that any certification metadata scheme (i.e. data `about the certification') should be applicable. across different communities. At the very minimum it must contain a link to the certified

document (or a metadata record for it), information about the certification service (e.g. journal name), and a timestamp for the certification. It might also include a timestamp at which the original harvest too place, and an indication of the current status for an open certification process. There might also be an indication of the certification type. A controlled vocabulary would be desirable for this.

From the above discussion, it is clear that the OAMH protocol could be a very useful tool for the transformation of scholarly communication through the Internet. Using the OAI, we can accomplish new tasks in a collaborative way that are not possible in the established setup of essentially closed initiatives.

8. Conclusions

While there is more and more freely accessible academic content on the Internet, the organization of that content is much less useful than the organization of content in formal archives and libraries. The Open Archives Initiative (OAI) is well suited to develop protocols that improve on this state of affairs by permitting interoperability between archives.

It has been suggested that the OAI framework be used to support institution-based digital archives. These archives will contain research results produced in an institution and archived in the library. The ARNO project (Academic Research in the Netherlands Online) is a small-scale, but pioneering effort to do just this. It remains to be seen how successful it will be. For libraries to take on the rôle to archive, rather than to make material available that is produced by intermediaries outside academia, is a big step. It implies a change in their perceived rôle. It remains to be seen if libraries will take up that challenge.

It may be that the establishment of institution-based archives is a better and faster way to free scholarly literature than waiting for each discipline to embrace author self-publishing in its own proprietary way. However, this is by no means clear and the same OAI framework that supports interoperability between institution-based archives also supports interoperability between discipline-based archives. We expect that, whatever the final outcome, there will be increasing numbers of both institution- and discipline-based archives in the near term. We think that the degree of interoperability between these archives will strongly influence their success.

9. References

Arms, William Y. (2000). Automated Digital Libraries How Effectively Can Computers Be Used for the Skilled Tasks of Professional Librarianship? D-lib Magazine 6. available at http://www.dlib.org/dlib/july00/arms/07arms.html.

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

Kling, Rob and Geoffrey McKim (2000). Not Just a Matter of Time: Field Differences and the Shaping of Electronic Media in Scholarly Communcication. Journal of the American Society for Information Science 51(14), 1306-1320.

Van de Sompel, Herbert, Thomas Krichel, Micheal L. Nel-son, et al. (2000). The UPS Prototype project: explor-ing 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/upsproto.ps.

Varmus, Harold (1999). E-BIOMED: A Proposal for Electronic Publications in the Biomedical Sciences. available at http://www.nih.gov/about/director/ebiomed/ebi.htm.