Open Archives and Free Online Scholarship[*]

submitted to JCDL2002

Thomas Krichel
Palmer School
Long Island University
720 Northern Boulevard
Brookville, NY 11548–1300
USA
krichel@openlib.org
http://openlib.org/home/krichel

Simeon M. Warner
arXiv
Cornell University
4130 Upson Hall
Ithaca, NY 14853
USA
simeon@cs.cornell.edu
http://www.cs.cornell.edu/people/simeon/

Abstract

The potential for free access to scholarly documents on the Internet continues to occupy the minds of all actors in scholarly communication. While there is much agreement that free access is desirable, there is little agreement about how this will come about. We have been actively involved in this transition through our work on two major initiatives in this area. These are arXiv, which covers Physics, Mathematics and Computer Science and RePEc, which covers Economics. We discuss the Open Archives Initiative (OAI) and the Academic Metadata Format (AMF). These discussions inform our proposal of a conceptual framework for the transition to free online scholarship. We pay particular attention to the rôle that digital libraries play in the transition process.

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

[*] The authors are grateful for comments on this paper. This paper is dedicated to the memory of Sigrid Röckner.

1 Introduction

Electronic commerce over the Internet is now commonplace. In a 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 a 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, at least as far as the payment for papers in concerned. We assume that authors will still be prepared to grant the right to publish their papers without getting a monetary reward. Their aims will be the wide dissemination of their output and 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 research papers. This argument essentially compares two steady states. One this the current state, in which the scholarly literature is given to publishers who charge access fees in exchange for formatting, peer-review and distribution services. Such access fees are necessary in the "Gutenberg" era where there are positive marginal distribution costs. Authors accept restricted access in exchange for peer-review and distribution services provided. This steady state is being upset by a technological shock in the form of the Internet. The Internet allows for marginal distribution costs —i.e. the costs of distributing one further copy of a paper—that are zero. Furthermore, formatting costs have been reduced by electronic submission and inexpensive hardware means that archival costs are very low. In this "post-Gutenberg" era the dominant cost is peer-review and, even assuming this is maintained, one can imagine other steady states. One is that the publishers will charge authors for the peer-review services that they provide. Authors and readers will benefit from free access to scholarly work and thus the preservation of access tolls is suboptimal. In the post-Gutenberg social optimum, publishers offer free access to scholarly documents, and concentrate on the rôle of providing peer-review intermediation. This argument is well-understood. The key feature of this argument and many other post-Gutenberg models is that there is free access to the scholarly literature. This is the starting point for our work.

In this paper, we examine the transition from the existing steady state to another. We concentrate on how to achieve free online scholarship. It should be clear that by free scholarship, we mean that the results of scholarship are freely accessible. Clearly there will be some remaining costs. If these costs are low, they can be absorbed by other activities. Examples of output that is freely received but is costly to produce abound. Religious services are an example, television broadcasting is another.

We are motivated by two underlying convictions. First, we are convinced that the freeing of the scholarly literature is optimal but not inevitable. It is possible for society to remain stuck in sub-optimal equilibria. The example of the Scholes typewriter—also known as the "qwerty" keyboard—illustrates this. Parks (2000) has put together an analysis of the powerful forces that will keep the toll-gating publishers in place. Second, we are convinced that concentrating on free access to papers is a considerable oversimplification. We believe that the whole of the scholarly communications system will need reform for free scholarship to establish. Concentrating on the aspect of free access to papers without addressing other aspects is too restrictive.

Our convictions are founded in our contributions to the two important discipline-based scholarly communication projects. These are arXiv and RePEc. These projects have very different backgrounds and modes of operation. We review them in section 2 and section 3, respectively. Our convictions are also the basis for our contributions to the Open Archive Initiative, and, more recently, to efforts to create a metadata format for scholarly communication. These are presented in section 4 and section 5, respectively. In a brief section 6, we discuss the current achievements of free online scholarship from a historical perspective. In section 7 we speculate what shape informal free scholarship will take. Section 8 discusses how formal scholarly communication can be freed. section 9 introduces a managerial model for formal free online scholarship. Section 10 discusses the rôle of digital libraries. section 11 concludes.

2 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 the sense that all submissions and the master database are made at one site. In 10 years, arXiv has grown to 33,000 submissions per year and serves an estimated 80,000 users. arXiv is described by Ginsparg (2001) and Warner (2001).

Two important factors have helped the growth of arXiv. 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.

A second factor behind the growth of arXiv is long-term funding. arXiv has been funded by the US National Science Foundation (NSF) and the US Department of Energy. Since arXiv's move to Cornell in summer 2001, it has been funded by the NSF and Cornell University Library. Its 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.

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 conventional 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. In the longer run, there may emerge a scenario where authors and journals rely on arXiv to providing a digital archive. Journals will then simply contain reviews of papers that are deposited at arXiv. Geometry and Topology at http://www.maths.warwick.ac.uk/gt/ is an example of such an "overlay journal". The presence of a central archive and a range of decentralized overlays will realize free access to fully peer-reviewed literature.

3 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 that was founded in 1993. Krichel (1997) has a summary of the early history of this project. 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 are stored on a file system following a simple protocol called the Guildford protocol. Harvesting software is used to collect the data. Shortly after the implementation of the protocol, several user services appeared that were built on the data. At the time of writing, RePEc has over 200 archives that contribute data and metadata, and ten different user services operating in seven countries. There are about 55,000 downloadable paper cataloged in RePEc.

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. EDIRC is mainly linked to the rest of the RePEc data through the HoPEc personal registration service, see Barrueco Cruz, Klink, and Krichel (2000). This service can be used by economists to register themselves as authors of the documents that are contained in RePEc. To date 10% of all papers have at least one of the authors who is 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

Recently efforts have been made to improve the collection of access and download statistics across user services through the LogEc project. It aims to provide academics with direct evidence of how well the system disseminates papers. Work is currently under way to build a citation linking system. This will allow extensions to HoPEc to collect citation evidence as well, enabling registered authors to directly access information on which of their papers have been cited, whom they have be been cited by etc. From experience, we know that authors are very interested in that type of information.

Thus RePEc is probably the most ambitious project in Internet scholarly communication to date. The final aim is that every author, institution and document in economics will be registered in a database. Thus the project will need constant attention and never be finished. The project has to levy volunteer efforts of academics to supply data. The NetEc project received £129,000 funding from the Joint Information Systems Committee (JISC) of the UK Higher Education Funding Councils. RePEc without any external funding. Running such a large-scale operation with volunteer power only is a remarkable technical and organizational achievement.

4 The Open Archives Initiative

In the Summer of 1999, Van de Sompel, Krichel, Nelson, et al. (2000), conducted an experimental study to set up a common search interface to e-print archives, known as the Universal Preprint Service (UPS) prototype. A call for a UPS meeting was issued by Paul Ginsparg, Rick Luce and Herbert Van de Sompel. The motivation was to improve the interoperability of e-print initiatives. At that time, these were, by order of size, arXiv, RePEc, NCSTRL, and CogPrints. In addition, the electronic dissertation network NDLTD, the digital Highwire Press, the Physics reference service SLAC-SPIRES, and a few others were represented at the meeting. An initial proposal was on the table for a "Santa Fe Agreement" that would outline usage and access conditions to metadata of archives. This proposal was rejected as being too legalistic. Instead it was decided that the permission to access should be given indirectly, through the construction of a technical interface that is designed to provide access to metadata. This idea was the basis of the Santa Fe Convention that was drafted after the meeting and is documented in vandesompel00santa

While the OAI started as an eprint interoperability framework, interest from other communities soon appeared, thus its scope broadened considerably. The result was the OAI Protocol for Metadata Harvesting, see http://www.openarchives.org/OAI_protocol/openarchivesprotocol.html. This can be used for the interoperability of any type of digital library and was designed to provide a low-barrier to interoperability. Key features include:

- support for multiple metadata formats,
- requirement for unqualified Dublin Core (DC) meta-

data as means of global interoperability,

- use of Hypertext Transfer Protocol (HTTP) transport, and
- use of Extensible Markup Language (XML) encoding.

There are three reasons why the OAI is important for scholarly communication. Neither of them are technical, all concern the underlying vision. First, the OAI encourages the sharing of data and metadata. Digital libraries are no longer viewed as closed entities. Second, the OAI adopted the business model, pioneered by RePEc, that separates data providers and service providers. Furthermore, it allowed for multiple metadata formats and cleanly separated the metadata from the transport protocol. Third, the OAI marked a change from the vision of centralized, discipline-specific archives to decentralized and perhaps institution-based archiving. Related to that, OAI created the opportunity for the library community to enter as providers of freely available scholarly literature in institution-based digital archives. These archives will contain research results produced in an institution—from all disciplines—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.

5 The Academic Metadata Format

The Academic Metadata Format (AMF) is an outgrowth of the OAI technical meeting in September 2000, when Thomas Krichel and Simeon Warner were asked to draft a successor to the metadata format that was used with the Santa Fe Convention. The initial aim was to have a metadata format that would plug in with the OAI compatible eprints servers. The current specification of AMF is described in Brody, Jiao, Krichel, and Warner (2001).

The general ansatz is that is not possible to construct a simple, and detailed format that can be applied to multiple domains. The Dublin Core Metadata Element Set, (DCMI 1999), sacrifices detail for generality. AMF sacrifices generality to concentrate on the descriptive needs needs of scholarly communication. At the heart of the requirements analysis stands the question of what actually needs to be described. Krichel and Warner (2001).

argue for four classes of entity

1. resources
2. collections of resources
3. people
4. organizations

In this model, resources are either digital objects, are physical objects for which a digital substitute can be found. AMF currently only deals with textual resources, understood as in the DCMI Type Vocabulary (DCMI 2000), but the entity model is more general. However it is not as general as the entity model of the Resource Description Framework (RDF) promoted by the World Wide Web Consortium as set out in Lassila and Swick (1998). There a resource is anything that has identity. Collections of resources take their meaning form the DCMI Type Vocabulary. For example, an academic serial is a group of a resources. People and organizations are not covered by the Dublin Core.

AMF allows for specification of properties for instances of the entity classes. We will call such instances "AMF instances" in the following. AMF also allows for the specification of relations between them. As far as the syntax is concerned, the AMF design is constrained by the OAI's choice of XML responses. Consequently, AMF makes no use of RDF. Instead, AMF borrows from natural language. The XML elements that represent the fundamental entity classes are called "nouns". XML elements that give properties to nouns are called "adjectives". Some of them admit other adjectives as children, but most of them admit no children. Just as in natural language, adjectives are used to qualify a noun. To make a relationship between two nouns, AMF uses "verbs". A verb must have at least one noun as child element. Just as in natural language, verbs are used to relate two nouns.

There are two important innovation in this framework. First, the metadata is not none homogeneous type, but of several types. Instances of each type can be called up at will. A second innovation, is that there is ample opportunity for decentralization of maintenance. Records may be maintained by different persons, in different files. The record about the person may be maintained by the person concerned. AMF can express that some record is authoritative for a certain handle, made by whoever is responsible for it, and by for non-authoritative records, provided by somebody else. AMF provides no means to document these responsibilities, however.

The most important innovation of AMF, however, is not technical. Rather, it is the vision behind the descriptive model. First, AMF is designed to describe the academic world as process that relates resources to non-resources. Thus, it goes beyond resource focused formats such as the Dublin Core or MARC. Second, AMF accepts that people, resources and organizations are best described using different formats. A common framework allows relationships to glue these different entities together.

6 Formal archiving of research papers

Making the world's scholarly literature in a specific discipline freely available over the Internet has been the dream that animated the founders of arXiv and RePEc. They took action in their home disciplines as an intermediate step for the community. Characteristically the disciplines concerned have a pre-publication tradition. We say that there is a pre-publication tradition if researchers have a habit to circulate account of recent research findings in an informal manner. There are two kinds of pre-publication disciplines. In the "preprint" disciplines, it is the tradition for the author of a research report to issue a preprint to colleagues who may be interested in the results. In the "working paper" disciplines, it is the tradition for institutions to issue working papers, sometimes also called discussion papers or tech reports. These are exchanged between institutions. There are broadly four disciplines that have a prepublication tradition. Two of them, economics and computer science, have a working paper tradition. The two others, mathematics and physics have a preprint tradition. The differences in the mode of operation between arXiv and RePEc can—in part—be traced to back to difference in pre-publication traditions. We think that the emergence of centralized or decentralized archive systems depends on the communication pattern prior to ubiquitous access to computer networks. In the working paper disciplines, it seemed natural for departments to continue to issue papers in electronic form. On the contrary, in the preprint disciplines, it made more sense for authors to submit directly to a central system and thus reach a wider audience.

For a long time arXiv had virtually a monopoly position in the free online scholarship world. Its centralized discipline-specific model—where all papers that are relevant to a certain discipline are stored on one server—became the only recognized business model for free online scholarship. Two important points were completely overlooked at that time.

First, the centralization of arXiv was a gradual process. Before 1994, archives for some new subject areas were started at other sites. These sites used the same software as arXiv. 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 of organization and access.

Second, while it is theoretically possible that the arXiv model could be successfully applied to all other disciplines, the historical evidence casts doubts on the practicalities of such plans. 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, since 1996, and the Economics Working Paper Archive at the University of Washington at St. Louis since 1993. Neither has grown beyond 1,500 documents. In addition, arXiv has found it difficult to expand beyond Physics and Mathematics. CoRR, the Computer Science section of arXiv was added in 1998. It has grown only very slowly indicating reluctance within the community.

What will happen in the disciplines that neither have a working paper nor a preprint tradition is not clear. We advance the hypothesis that neither the centralized nor the decentralized discipline-specific systems will find much acceptance. Instead, a cross-disciplinary institutional archiving strategy may be more appropriate. This is the implicit model of the OAI as applied to institutional libraries. One could imagine that libraries might replace publishers altogether. In that scenario, the library of the authors' institution would make the authors' work available, and the library of the readers' institution would ensure that the work could be found and would be accessible. Access would be free. In practice however a system of entirely library-driven scholarly communication system would have many obstacles to overcome.

First, it can be assumed that within academic institutions, no author can be forced to deposit their paper with the library. Such conduct has to be achieved an a voluntary basis. Submitting a paper to any archive consumes some time. The constituent full-text files have to be assembled and a metadata record must be composed. In the pre-publication disciplines, the library can take over an activity that is already done within individual departments. However the transfer of this activity to the library is likely to be resented as a loss of autonomy by the department. It also has the whiff of authoritarian control of output quantity. In addition, many departments have some form of vetting for pre-publications, which complicates submission procedures considerably. Outside the pre-publication disciplines, libraries will have to do a lot of convincing work because there is no tradition of publishing prior to formal channels. In particular, authors may be concerned that their work will be plagiarized, or that they will run into copyright problems if they later want formal publication to achieve peer-recognition. It is therefore doubtful that a set of public servers of academic papers can be build without an incentive device that will make authors collaborate with the archiving policy of the institution.

The creation of incentives for authors is a problem for library-based archiving. We conjecture that it is impossible for libraries to achieve something in this area without some link of the archive to a review of some form. There are two promising areas. First, libraries in institutions where a department offers a free journal can offer to back up the journal. We are aware that this is the case with the "Economics Bulletin", a new free electronic journal that is part of RePEc. Although such an archive would not archive directly institutional material, it will go a long way to help the budding culture of free journals on the Internet. Second, libraries can use alternatives to peer review. We will come back to this point at the end of section 9.

At the conclusion of this section, one proposal emerges. It is not possible to gather online papers in significant numbers in formal archives—such as the ones operated by libraries—unless there is a reform of the whole of the scholarly communication process which gives scholars an incentive to participate. Since much of the reform of scholarly communication depends on papers being available—cf. the overlay journals of arXiv—we have a true catch-22 situation. We must look deeper at the whole of the scholarly communication process to find a way forward.

7 The vacuum cleaner scenario

If formal archiving fails, the existing scholarly publishing infrastructure will survive, but it might be undermined by what we call the vacuum cleaner scenario. In this scenario, there is a free layer of research documents made available on the web by their authors. They may be withdrawn at any time. There is no bibliographic organization of these papers other than that which can be done automatically. Papers can be found through generic Web search engines, or possibly through a specialized engines such as inquirius.

But, 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. While there could be some system of registering and storing copies of these web pages, it seems more likely that there will remain a toll-gated layer of certified, quality-controlled, literature. We assume that these quality-controlled collections of research papers will have access restrictions. Most of them will only be accessible to subscribers. This toll-gated layer will have good bibliographic descriptions that are specific to each vendor. It does not appear likely that there will be a common catalog of these works that will be freely accessible.

This scenario has been defended by Arms (2000). He envisages 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.

It is clear that the vacuum cleaner scenario does not provide for the access function of scholarly communication. It can not register a new claim because there is no deposit of the paper to a place where it can no longer be altered. Long-run archiving of the system seems cumbersome. Since papers are deposited on the web, we need an archive of the complete web to archive all papers. A reliable method to do that and a reliable implementation are not immediately forthcoming. The registration and certification function will have to be performed by traditional publishers or some other organizations of the same nature. Some of these publishers will limit access to published output through toll gates.

Despite the weakness as a full scholarly communication system, the vacuum cleaner scenario provides a powerful disruptive technology—in the sense of Christensen (1997)—to the existing scholarly communication system. The existence of author web pages is a measure of the dissatisfaction authors have with the limited dissemination provided by the existing system. The existing scholarly communication system has no right to survival. It could be that, over time, journals are evicted by "what is new? what is cool?" link lists on the home pages of top academics. Archiving may become part of web archiving. However, it is likely that the need for certified knowledge will keep some publishers in business. It is also likely that at least some of the output of publishers will not be freely available.

8 Free Scholarly Communication

Let us step back for a time and look at the theory of scholarly communication. We use a simplified version of Roosendaal and Geurts (1997) model of scholarly communication. They consider four functions of scholarly communication.

1. *Registration* allows researchers to claim a new scientific finding.

2. *Certification* allows for other researcher to approve the finding as a new scientific finding.

3. *Awareness* allows for the community of interested researches to become aware of the new finding.

4. *Archiving* allows the access to a historic record of the discovery of the new finding in the future.

Traditionally, the former two functions are associated primarily with publishers acting as agents of authors, and the latter two functions with libraries as agents of readers. Another historical fact is the chronology of the functions. Papers were submitted to journals and/or issued as preprints. This was the registration process. When the journal accepted to publish the paper, the acceptance of the paper in journal was an expression of certification. The distribution of the journal issue was the point at which awareness-raising was started. Back issues stored in libraries ensured the archiving.

To set up a free publication system, it is likely that this chronology should be inverted. That is, that first papers should be archived on a server that is independent of the author's control. This act also allows for the registration of the new claim. In addition the fact that appears in an archive will allow immediate circulation. Peer-review can appear separately. This point is made most effectively by Smith (1999) in his "Deconstructed Journal Proposal". He envisages Subject Focal Points (SFP) that review papers. Storage and review of papers are separated. SFP's simply review papers. In terms of AMF, they prepare collection level metadata. To do this, they must have a close collaboration with archives.

Some evidence is available from the prepublication disciplines that such a model can be implemented there. This is best shown through the overlay journals of arXiv. In other disciplines too, there is a growing number of small start-up free journals. Most are doing pioneering work with little more than a web server and the enthusiasm of a small editorial team. However, evolution is very slow, and it is therefore not likely that free journals will drive out toll-gating publishers anytime soon. In the meantime, the vacuum cleaner scenario may undermine completely the whole formal scholarly communication system.

More generally, we can perceive that peer review is partly sense an artifact of the Gutenberg universe, just like access toll gates. If there are positive marginal costs associated with publishing a paper, it must first be evaluated to see if the publication costs are justified. The process of "peer review" is traditionally a process of "pre-review" i.e. before general awareness of the paper through publication. It is not likely that this process will disappear. Even under the vacuum cleaner model, some form will survive, but in a different form, as a process that runs concurrently with the awareness process.

We can add the pre-review process through peer review a process of post-review of the impact of a paper. This impact review evaluates a paper through the impact that it has had within the scholarly community. There are several criteria that can be used to measure impact of a paper

- the number of the paper has been accessed or downloaded;

- the citations that it has received;

- which papers it has received citations from;

- any other distinctions, such as prizes, conference presentations etc

In addition, an impact review for each author can be prepared by aggregating for all the papers that she has written. This is most important. Researchers do not expect payment for research, but they are very keen on the impact that their work is having. Their market value is much dependent on the total impact of all the papers that they have written. Fortunately, impact measures of papers can be better aggregated across the papers of an author than peer review results. It is therefore quite possible to develop author rankings, once all the papers that an author has written are known. From there it is possible to develop institutional rankings, once the institutional affiliation of all authors is known. There can be little doubt that impact review is a powerful driving force for free scholarship. Lawrence (2001) is a pioneering study that proves—through citation analysis—that online work has a higher impact. The development of free online scholarship is likely to go hand in hand with a decline of the importance of peer review and a rise in the importance of impact review.

9 A managerial model

Roosendaal and Geurts (1997) have given us a model of the functions of scholarly communication. In this paper, we identify four tasks that have to be accomplished to provide online scholarly communication.

1. *Deposit*: to place a paper where it is publicly accessible, can no longer be modified by the author, and where it will stay for the foreseeable future.

2. *Describe*: to compose a metadata record about a paper or other AMF instance.

3. *Identify*: to state that a particular item is unique among a collection of descriptions. It implies the deduplication of multiple descriptions.

4. *Relate*: to establish a link between two described and identified elements.

We will discuss each task in turn and see if the output of the task can be made freely available. We assume the use of the AMF descriptive model, however, our arguments generalize naturally to other descriptive models.

A set of archives, as provided in by OAI compatible archives, a set of RePEc-style archives, but in fact any web server, can take over the deposit task. As we have seen, deposit is the basis of free scholarly communication. The papers must be freely downloadable on the Internet, and their locations must be stable. Location is the set of Curls or similar at which all elements of the full text are accessed. Experience with arXiv shows that the cost of the deposit function and is very low (a few dollars per paper).

The describe task is already more abstract. If there is a need for human-generated metadata, then a question of how this metadata is to be generated arises. This is the level where the OAI comes in. The archives will deliver metadata, at least up to unqualified Dublin Core, for the papers that are in the archive. But this is not the only form of description that may be available. The SFPs of the deconstructed journal proposal will make additional descriptions available. Such secondary description may not be available for public access. But for the moment let us assume that it will be, or narrow the focus of our attention to the stock that is freely available. The same author pressure that pushes papers to be available over time will push descriptions to be freely available.

It is then often assumed that a set of interoperable archives that identify and describe papers will complete a scholarly communication system on the Internet. That is true if one limits the information set to papers and their peer review. But in order to implement impact review—which appears as the enabling condition to free online scholarship—two other tasks that need to be accomplished. The first is identify. The task to identify arises for all AMF entity instances that are involved in an impact review. If there are papers by the same author in different paper archives, then the author needs to be identified and the papers that she has written need to be aggregated in the author record. Similarly, to evaluate the work of a department or research unit, it is crucial to gather an institutional record that lists all its researchers.

The tasks of implementing impact review is close to the work of the existing secondary databases such as the Web of Science product purveyed by the Institute for Scientific Information. However, this product is not relational in the sense that journals and authors are identified, and it is not is not free, of course. To purvey such services on a free basis we need a new kind of organization. We will call this an "aggregator" for the moment. All identifiers that are being assigned by one aggregator should go to the some AMF entity instance. But of course, in a world with several aggregators working independently, there would be nothing to prevent several aggregators to convey different identifiers to the same AMF instance. In a technically ideal situation, a monopoly aggregator would register every AMF instance on the planet. There are, however, a number of reasons why a monopoly aggregator is socially difficult to implement

- Large monopoly organizations have a tendency to be inefficient.

- The funding of the aggregator would require large financial resources gathered by an international collaborative effort.

- AMF instances that are not registered would loose a lot of visibility. This creates a lot of problems for contested material.

Therefore it appears that discipline-specific aggregators are a way forward. Disciplines group AMF instances together. It is understood that that disciplines do not partition the set of all possible AMF instances. Instead there are some instances that will belong to several disciplines. They are likely to be picked up by different aggregators. The emergence of aggregators and their scope will depend on the historical accidents, the presence of entrepreneurial individuals in discipline communities and the nature of the discipline. It would go beyond the scope of this paper to discuss this in further detail.

One discipline aggregator is already in operation: the RePEc project. The CiteSeer autonomous citation system could become an aggregator if it strengthened its archival component. Therefore there appears to be evidence that it is possible to operate aggregators who make their data freely available. These aggregators will be a powerful addition to the services offered by free scholarship initiatives.

10 Digital libraries

Computer geeks have managed to make a complete operating system available on the WWW. At the same time, academics have not been able to make a coherent index to their works available, let alone organize the free availability of the full text of these works. Building such a bibliography is, from a technical point of view, a much simpler task than the distributed maintenance of a computer operating system, yet it has not been built.

Free online scholarship should be an ideal showcase for digital libraries. The only large-scale digital library lead by the digital library community to date was the the Network Computer Science Technical Reports Library, NCSTRL, pronounced as ancestral. This was a research project led by Cornell University that had collected about 10000 online documents by the time it ended in 2000. NCSTRL has recently been revived by a collaboration that intends to create a sustainable system using OAI protocols. This is an interesting full-circle since the OAI protocol is strongly influenced by the Dienst protocol used by NCSTRL. Experience with NCSTRL led to Oar's choice of a much simpler protocol and a genuinely decentralized structure.

There is a great future for digital libraries to support the work of aggregators. To get active in this area digital librarians have to think more about helping contributors of data, rather than users. The HoPEc project for the registration of authors is one example, the CiteSeer system is another. This work implies a slight change in the emphasis of digital libraries as tools to organize contents, rather than an tools to make organized contents available to end users.

11 Conclusions

In this paper, we have added some new themes to the debate on free online scholarship. First we introduced the concept of impact review alongside peer review. Then we proposed a managerial model as a way of thinking about the transition to free online scholarship. Finally we introduce the rôle of aggregators.

Our aim has been more normative than positive. We wish to achieve a transition to free scholarship. We seek to find out how this can be done. We have resolved one important problem that bedevils the debate about online scholarship, the question whether archives should have a discipline or an institutional scope. We find that this debate is a red herring. It is probable that both will exist. However, eventually, the master archives are likely to be institutional with aggregators—who technically are nothing but another form of OAI compatible archive—that are discipline based.

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) has developed protocols that improve on this state of affairs by permitting interoperability between archives. AMF is a able to encode academic output as a process, rather than a set of resources. Large collections of AMF data will open the door to applications in the area of impact review of academic work. All this has very exciting potential.

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.

Barrueco Cruz, José Maunel, 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.

Brody, Tim D., Zhuoan Jiao, Thomas Krichel, and Simeon M. Warner (2001). Syntax and Vocabulary of the Academic Metadata Format. available at http://amf.openlib.org/doc/ebisu.html.

Christensen, Clayton M. (1997). *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail*. Harvard Business School Press.

DCMI (1999). Dublin Core Metadata Element Set, Version 1.1: Reference Description. available at http://www.dublincore.org/documents/dces/.

DCMI (2000). DCMI Type Vocabulary. available at http://www.dublincore.org/documents/2000/07/11/dcmi-type-vocabulary/.

Ginsparg, Paul (2001). Creating a global knowledge network. available at http://arXiv.org/blurb/pg01unesco.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.

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 and Simeon Warner (2001). Design of a metadata framework to support scholarly communication. available at http://openlib.org/home/krichel/papers/kanda.html.

Lassila, Ora and Ralph R. Swick (1998). Resource Description Framework (RDF) Model and Syntax Specification. W3C Recommendation, http://www.w3.org/TR/REC-rdf-syntax/.

Lawrence, Steve (2001). Online or Invisible? available at http://www.neci.nec.com/~lawrence/papers/online-nature01/.

Parks, Robert P. (2000). The Faustian Grip: A dismal essay on the Status Quo in Academic Publishing. mimeo.

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

Smith, John T.W. (1999). The Deconstructed Journal—a new model for Academic Publishing. *Learned Publishing 12(2)*. available at http://library.ukc.ac.uk/library/papers/jwts/d-journal.htm.

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.

Warner, Simeon M. (2001). Open Archives Initiative protocol development and implementation at arXiv. arXiv paper cs.DL/0101027, available at http://arXiv.org/pdf/cs.DL/0101027.