# OAI AND AMF FOR ACADEMIC SELF-DOCUMENTATION

Pavel I. Braslavsky
Institute of Engineering Science
Ural Branch, Russian Academy of Sciences
Komsomolskaya 34
620219 Ekaterinburg
Russia
pb@imach.uran.ru

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

## Introduction

The traditional way to communicate academic research results have been academic journals. Academic journals have two types of costs. They cost time and money. They are cost time because publication delays can be very long. It is not uncommon for papers to take many years for publication. The financial costs of journals are well documented through the literature on the "serials crisis". Nowadays this established business model of publication in peer-reviewed journals is under pressure from the authors who can publish their work independently from the peer review process.

The Internet has given new rise to possibilities to publish contents at the marginal distribution costs that are virtually zero. Any organization or individual can become a "publisher" in the sense that they can make documents public. However it does not replace the quality control function of the established publishing outlets. Self-publishing is desirable because it furthers equal access to scientific documents for anyone with an Internet access. In Russia, there are a number of grass-roots initiatives in this area. It is however clear that the mere accumulation of distributed data cannot provide for dynamic scientific communication. There is a requirement to organize the data to provide valuable services related to them.

Currently many documents on the web are indexed by search engines. The ratio between the surface Web (i.e. accessible for search engines) and the deep Web (i.e. invisible for search engines) may well be in decline as search engines have managed to index non-html files such as Postscript and PDF documents. Even if most of scientific data were indexed by search engines, it would not provide for a scholarly communications system. The problem, in a nutshell, is that search engines are pure services. They do not have responsibility for the contents that they provide. Search engines are general-purpose solution, while academic documents require more attention.

Although no standard business model for the open access to scientific documents in digital form has been established yet, independence and decentralization are expected to be its most important features. By decentralization we mean that the provision of contents must be the work of many providers. Under these conditions, the objective becomes not to concentrate and store data in one place but to build services upon distributed contents. Since we desire more than mere indexing, we need quality metadata. Free high-quality academic services can be built upon metadata provided by the individual providers who absorb the cost of data provision.

Three things need to be established before decentralized provision can take place. First of all, there needs to be the will to provide such data. Second, there needs to be agreement on what kind of data will be provided by the individual contributors. Third, there needs to be a way for the data to be "harvested", i.e. collected from the different providers.

## Open Archive Initiative Protocol for Metadata Harvesting

The last problem can be easily solved by using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [6]. The current version 2.0 was published on June 14, 2002. The OAI-PMH provides a technical framework for the harvesting of metadata contents. The main feature and advantage of the protocol is that is relatively (compared to say Z39.50) easy to implement.

There are two classes of participants in the OAI-PMH framework:

- *Data Providers* administer systems that support the OAI-PMH as a means of exposing metadata;
- *Service Providers* use metadata harvested via the OAI-PMH as a basis for building value-added services.

A harvester is a client application operated by a service provider. It issues OAI-PMH requests to repositories maintained by data providers. OAI-PMH requests are expressed as HTTP requests. The OAI-PMH defines the following six requests:

| **Request** | **Expected Response** |
|---|---|
| `Identify` | Information about a repository |
| `ListRecords` | List of metadata records from a repository |
| `GetRecord` | An individual metadata record from a repository |
| `ListIdentifiers` | List of metadata record headers |

| | |
|---|---|
| ListMetadataFormats | List of metadata formats available from a repository |
| ListSets | The set structure of a repository, useful for selective harvesting |

The data returning from the repository must be formatted in XML. The OAI-PMH mandates a version of simple unqualified Dublin Core [3] as a common metadata format. Since all Dublin Core elements are optional, however, this does not require any semantic structure on the records. The OAI-PMH allows any optional metadata formats encoded in XML for extensibility and for community specific enhancements.

By now, about 100 registered data providers and about 11 service providers are listed on the OAI home page. It is to expect that more participants will join the initiative in the near future.

## Academic Metadata Format

The second problem (what to encode) is more challenging. To find a common semantic standard for the description of academic activity is very difficult. Each discipline and organization has specific descriptive needs. Established bibliographic standards have their roots in offline documents accessible through card catalogs. They are not suitable for current technology; They are focused on the description of documents. The later problem is particularly acute. If we want to get self-archiving going, we need to create incentives for academics to advertise themselves through their document. That is we need to have a stab at solving the first problem that is built-in to the solution of the second problem. Thus, we need to focus on the description of authors and their institutions, rather than merely on the documents that they produce.

The Academic Metadata Format (AMF) [1] is a modular metadata model for academic authors, institutions, documents, and collections of documents. It uses standard vocabularies wherever possible and simply builds an XML framework for their usage. AMF can be used to build descriptions of complete academic disciplines that relate authors to their institutions, to the documents that they have written and to the organization of documents into collections.

AMF is encoded in XML. There are four noun elements in the AMF:

| | |
|---|---|
| person | a physical person |
| organization | an entity that has physical persons as its members |
| text | a text resource |
| collection | a collection of resources |

The word “text” is understood here in the sense of the Dublin Core (DC) type vocabulary. It is possible to add further, non-textual resource types to AMF but that is not a priority, as most academic documents fall within the DC text category. Each instance of a noun element in AMF data that is not an empty element is called an AMF record. An AMF record can have child elements, which are optional and repeatable. An AMF record admits two types of child elements. The first type is “adjective” elements. Adjectives give further information about nouns. The second type is “verb” elements. Verbs relate one noun to other nouns. Each verb must have one or more nouns as children. Verbs must not have adjectives as direct children.

Figure represents a sample AMF record (shortened to conserve space). It appears readable without additional comments.

```
<amf>
    <collection id="csfhrd">
    <title>Classification Scheme for Human Rights Documentation</title>
    <homepage>http://www.huridocs.org/clasengl.htm<homepage>
    <haseditor>
        <person>
        <name>Ivana Caccia</name>
        <email>icaccia@web.apc.org</email>
        </person>
    </haseditor>
    <haspart>
        <collection id="csfhrd:GEN II.10"><title>
        natural justice </title></collection>
        <collection id="csfhrd:GEN II.20"><title>
        universality / relativism </title></collection>
        <collection id="csfhrd:GEN II.30"><title>
        philosophy & human rights </title></collection>
        <collection id="csfhrd:GEN II.40"><title>
        political theories & human rights </title>
        <haspart>
            <collection id="csfhrd:GEN I.41"> <title>democracy</title>
            </collection>
            <collection id="csfhrd:GEN I.42"> <title>liberalism</title>
            </collection>
            <collection id="csfhrd:GEN II.45"><title>marxism</title>
            </collection>
        </haspart></collection>
    </haspart></collection>
</amf>
```

**Figure. A sample AMF record**

## AMF and OAI-PMH

OAI-PMH and AMF interoperate on three levels.

First, they can be used to collect bibliographic data from servers to build large collections of bibliographic data. The records can then be identified through removal of duplicate descriptions. These bibliographic collections form useful services by themselves.

At the second stage, the items in the bibliography can be related to personal data. Thus, it is possible to have users registering with the system to provide data about papers that they have written or collections that they are editing.

At the third stage, evaluative data can be gathered from the dataset. These evaluative data concern page views of documents, full-text download information, as well as citations data that can be gathered out of the full text. These evaluative data are crucial to create incentives for authors and institutions to contribute data. If the contribution of data helps the improvement of the authors' ranks in some evaluative system – however silly that system may be – we can be confident that they have incentives to contribute.

At every level an OAI-PHM compliant archive can be used to collect and distribute data, and at every level AMF can be used to support the composition and retrieval of contents.

## Social and Economic Issues

Upon discussing technical framework of self-documentation process, we come to the initial issue. How should the self-documentation process in academic domain be promoted? How can incentives for authors and institutions to contribute data be created? The problem seems to be rather social and economic than technical.

A widespread business model for digital libraries—inherited from the print world—is a centralized collection of data. It is not likely that any one institution on its own could support the cost of providing for the centralized collection. To fund such an operation, a subsidy must be levied from the user or contributor communities. In the absence of a political decision to levy such a subsidy, a centralized collection does not work. Therefore, we need to find a number of regular contributors to the system and we need find ways in which the work of different contributors can be put together. The latter problem is mainly technical, but the former problem is mainly of an economic nature. Too many digital libraries focus on means to achieve results for the user. They neglect the contributor. The study of contributor motivation is crucial for the development of any digital library. It is a neglected area of digital library research. The history of digital libraries is littered with examples of test bed collections that were funded with research funding. These collections did not establish a sustainable contribution mechanism. They were closed when the research funding ran out. The NSF/DARPA funded NCSTRL project provides an illuminating example.

Our approach is a digital library that would be maintained by providers and users themselves, using free software. In that case, the cost of collecting the digital library is distributed. All contributors can internalize the cost of their own contributions. Just like the World Wide Web, the collection would need no subsidy. To get this to work, we need to create a community of contributors and users. Both users and contributors must have incentives to use the collection. In particular, contributors whose work we want to make freely available must have good incentives to contribute to the collection.

Free provision of documents is already one of the key features of academic life, in the area of research papers. Authors are not paid for the research papers that they write, often, they themselves—or their institutions—have to pay for publication process. The key to that free provision is an incentive mechanism within the academic world that rewards dissemination of work. Thus, within the world of research, authors make the research reports freely available. Of course, the documents are then appropriated by publishers who generally impose access restrictions.

Working on good user interfaces has been central to the work of digital libraries. Now, we must pay more attention to contributors' needs and intentions. All our contributors are academics. Academic writers are both highly individual and highly social. They are individual in the sense that their reputation as an individual determines most of their professional value. They are highly social in the sense that their position is only observable through acts of their discipline peers. Therefore, to impact on academics, services must be that exploit the urge to define the position of an individual academic within a competitive environment of other academics.

The conceptualization of such services could be called "aggregative evaluation". To impact on academic learning and research cultures, discipline-specific data aggregation have to be built. Institution-wide approaches are not sufficient, because contributor perception is more related to their status within the discipline rather than within the institution. The presence of aggregative data is necessary for the construction of evaluative data. But the opposite is true too. Evaluative data provides crucial incentives for academics to supply labor to the aggregation process. Authors will have good incentive to maintain an organized collection of their documents as long as the collection is publicly seen as an official evaluative record of their activities.

## RePEc and Socionet

The RePEc [8] and Socionet [9] projects can serve as examples of the described approach.

The RePEc project is a pioneering effort into building a decentralized academic publishing system for the economics discipline. After arXiv.org, the RePEc economics library is the second-largest library of freely downloadable scientific papers in the world. Since 1997, RePEc is based on the collaboration of the archives that provide simple "attribute: value" data templates in static files. The files can then be harvested from http and ftp servers where they are stored. A central collection is limited to a list of all available archives. At the time of writing there are over 250 such archives. They provide about 200,000 records in the domain of academic economics.

The Socionet project is based on full RePEc database and works on extending RePEc to the social sciences in Russia. By now, the Socionet database includes materials from the six social science disciplines: economics, sociology, political sciences, demography, law, and psychology.

The interesting feature of the Socionet is its user services. For example, a user can specify his/her personal information robot (i-robot) for automated controlling contents of input data flows. A customized i-robot excludes not relevant archives and series and filters the input flow of new additions according to the user's profile of interests. The i-robot creates reports with specified regularity and delivers it by e-mail and/or as static web pages. [4]

The Socionet services are built as implementation of ideas to develop the general RePEc concept from the initial state as “global electronic catalog” to an integrated digital information environment for community of social scientists. Its construction, the Internet technologies that it uses and its user services allow easy and flexible modifications of the database structure in a decentralized manner according to the natural development of community needs.

Socionet services help users not only to access the publications they need for their research; they also assist in putting back their research results into the same common information environment. This approach is implemented by Socionet in the Open Online Archives, which allow publishing research materials.

Until now RePEc and Socionet have used purpose-built metadata format called ReDIF for the description of documents, authors, institutions, and collections (see [7] for the complete documentation of ReDIF). The RePEc project converts all of its own and Socionet’s holdings to AMF and provides an OAI compliant archive for the collection as a whole. A gateway is available at http://oai.repec.openlib.org

## Conclusion notes

The OAI-PMH–AMF bond can provide an easy and effective solution for scholarly data circulation and academic communication. The pair solves the problems of metadata harvesting and academic process description, as well promotes the self-documentation efforts.

We should not forget that there are social and economic issues lying beyond and affecting the technical solutions.

However, some important unsolved issues lie beyond the scope of this paper. Among them, we could mention the problems of metadata reliability, identity control, record duplication, keyword sets and classification schemata standardization.

1. Academic Metadata Format, http://amf.openlib.org/doc/ebisu.html
2. Braslavsky P. How should we arrange for academic information retrieval on the Net? (in Russian) Proc. Memorial Lyapunov Conf., Novosibirsk, 2001, http://www.ict.nsc.ru/ws/Lyap2001/2320/

3. Dublin Core, http://www.dublincore.org/documents/dces/
4. Krichel T., Parinov S.I. The RePEc Database and its Russian Partner Socionet. Russian Digital Libraries Journal, 2002, Vol. 5, No. 2, http://www.elbib.ru/journal/2002/200202/KP/KP.en.htm,
Russian version: http://www.elbib.ru/journal/2002/200202/KP/KP.ru.html
5. Krichel T., Warner S.M. A Metadata Framework to Support Scholarly Communication. Proc. Intl. Conf. on Dublin Core and Metadata Applications, Tokyo, 2001. P. 131-137.
6. Open Archive Initiative, http://www.openarchives.org
7. ReDIF ver.1, http://openlib.org/acmes/root/docu/redif_1.html
8. RePEc, http://www.repec.org
9. Socionet, http://www.socionet.ru