OAI and AMF for academic self-documentation*

submitted to RCDL2002

2002–05–01

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

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

1 Abstract

The traditional way to communicate academic research results have been academic journals. Academic journals have two types of costs. They cost time and money. It is not clear which type of costs is more expensive. The financial costs of journals are well-documented through the literature on the "serials crisis". Nowadays this established business model is under pressure from authors who can publish their work independently from the peer review process.

The Internet has given new rise to possibilities to publish contents at 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 channels. Self-publishing is desirable because it furthers equal access to scientific documents for anyone with Internet access. In Russia, there are a number of grass-roots initiatives in this area.

Currently most documents on the web are indexed by search engines only. This is a low-cost, general purpose solution. Search engines are pure services, they do not have responsibility for the contents that they provide. Academic documents require more careful attention.

Although no standard business model for the open access to scientific documents in digital form has been established yet, independance 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. Free high-quality academic services can be built upon good metadata provided by the individual providers who absorb the cost of data provision.

Three things need the 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

*We are grateful for comments.

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.

The last problem is the easiest one to solve. The Open Archives Harvesting Protocol, launched by the Open Archives Initiative (OAI), is a technical framework for the harvesting of metadata contents. The current version 2 has been published on June 1, 2002. It is quite easy to implement for providers. It can be used transport any kind of data or metadata as long as they are formatted in XML. It mandates a version of simple unqualified Dublin Core as a common metadata format. Since all Dublin Core elements are optional, however, this does not require any semantic structure on the records.

To solve the second problem is more difficult. 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 root in offline documents accessible through card catalogs; they are not suitable for current technology; and they are focussed on the description of documents. The later problem is particularly accute. If we want to get self-archiving going, we need to create incentives for academics to advertize themselves through their document. Thus we need to focus on the description of authors and their instititutions, rather than on the documents that they produce.

The Academic Metadata Format (AMF) 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 acedemic disciplines that relate authors to their institutions, to the documents that they have written and to the organization of documents into collections.

OAI 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 a 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 instititutions to contribute data. If the contrbiution 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 compliant archives can be used to collect and distribute data, and at every level AMF can be used to support the retrieval of contents.

As an illustration, we will discuss the RePEc and Socionet projects. The RePEc project is a pioneering effort into providing such a collection. 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 230 such archives. They provide about 200,000 records in the domain of academic economics. The Socionet project works on extending RePEc to the social sciences in Russia. The RePEc project converts all of its holdings to AMF and provides an OAI compliant archive for the collection as a whole.

2 Note

The latest version of this document is available at http://openlib.org/home/krichel/papers/yoyogi.html.