Developing a National Metadata Profile for Institutional Repositories

ARL/SPARC Workshop, Institutional Repositories: The Next Stage
November 2004

Mark Jordan, Simon Fraser University (mjordan@sfu.ca)
Kathleen Shearer, Canadian Association of Research Libraries (mkshearer@videotron.ca)

In the fall of 2002, the Canadian Association of Research Libraries (www.carl-abrc.ca/projects/ir) began a project to implement institutional repositories (IRs) at a number of research libraries in Canada. An important part of this project is the pan-Canadian harvester, which was implemented in order to aggregate and search the collections of all of the participating IRs (http://carl-abrc-oai.lib.sfu.ca/). Although the project is a CARL initiative, it was decided early on that the scope of the harvester would be pan-Canadian rather than be limited to the CARL membership (which represents 31 research libraries in Canada). The Canadian harvester has been a valuable way of gathering information about a number of aspects of the IRs in Canada. In essence, it has become a sort of virtual 'laboratory' for the project. This brief paper will introduce three areas in which we are using the harvester to gather information about Canadian IRs in order to monitor and improve the services being providing by these repositories.

Growth of Institutional Repositories in Canada

The harvester is currently harvesting from eight OAI-compliant institutional repositories in Canada[1] and provides daily statistics about the number of records in each of the harvested repositories. This allows us to monitor the growth of the individual collections in participating repositories. The harvester also provides an indication of the aggregate growth of the collections, as well as the growth in the number of repositories in Canada. At this early stage in the project, only very broad inferences can be made using these statistics. But, it is hoped that in the future these statistics can be extrapolated to provide some insight into the critical success factors and barriers for IRs.

Searching Behaviour

The harvester also provides us with information about user behaviour. Again, because of the low number of records in the repositories, only basic inferences can be made with the information being gathered at this time. One of the statistics that was found to be significant was that 36% of all the searches of the harvested records returned no results. There were two major reasons for this. In some cases, there were no records in the database reflecting the content of the search. This could be expected, given the fact that there are still relatively few records in the participating IRs. In other cases, however, there were relevant records in database, but the expected elements were not present in the metadata. This indicated that there was a need to examine the metadata being assigned in the repositories.

Metadata Quality

The harvester is also being used to monitor the quality of the metadata being assigned at the IRs. An analysis of metadata records conducted in the summer/fall of 2004 found that the metadata

[1] *The scope of the repositories being harvested has been defined by three basic parameters: (1) the material is scholarly in nature; (2) the repositories are multidisciplinary; and, (3) the repositories contain the work that was created at the institution. Defining the scope for the harvested repositories has been, and will continue to be a challenge in this project, as the collection policies evolve.

being harvested by the CARL harvester is both inconsistent and incomplete. It was predicted that we would find that self-archiving will result in an irregular use of Dublin Core metadata by depositors. This was indeed the case with the Canadian repositories. However, we also found in our analysis that there were some significant systematic omissions of metadata elements at a number of the repositories (as outlined in the table below).

Element	No. of Providers that do not include this Element
Title	0
Creator	3
Subject	1
Description	0
Publisher	1
Contributor	2
Date	0

Element	No. of Providers that do not include this Element
Type	0
Format	1
Identifier	0
Source	4
Language	1
Relation	5
Coverage	7
Rights	5

The table shows the number of repositories (out of a total of eight being harvested) that did not include a particular metadata element in their records. For instance, three of the repositories did not include the 'Creator' element in their metadata. This was because of a limitation in the earlier versions of DSpace and been corrected in the DSpace version 1.2. However, the analysis also identified several other important metadata fields that were consistently omitted by the participating repositories, including subject, publisher, format, language and rights elements.

We also found that there was a high level of inconsistency with the harvested metadata. Of particular concern were large inconsistencies in assignment of the 'Date' element. While few of the dates were actually invalid, there was a huge variety in the way the date was recorded. In some cases, only the year was indicated; in others, the year and the month; and in others, the year, the month and the day. However, most worrying of all, was that the date indicated in the 'Date' field often represented the date that the item was deposited, rather than the date of 'publication'. Other inconsistencies were also found in the 'Type' and 'Description' fields. For example, when describing a journal article, we found a number of descriptions such as, article, journal (on-line/unpaginated), journal (paginated), learned journal article, scientific journal article (on-line or printed), and preprint.

Developing a National Metadata Profile

This analysis reflects the experiences expressed by others: While the OAI Protocol is robust; it is not in itself a magic bullet and cannot compensate for poor quality metadata (Cole, et.al. 2003). Metadata incompleteness and inconsistency are presenting a significant challenge to the effectiveness of harvesting and ultimately the searching of harvested records. To address these problems, members of the CARL Institutional Repository Project are seeking to better harmonizing the metadata of their IRs through the development of a National Metadata Profile. The Profile will be based on the needs of the Canadian community (ie. reflect the linguistic duality of Canada), be voluntary, and incorporate existing practices and standards as much as possible. In the next several months, a working group will be formed to begin to develop a profile that will contain: (1) required elements, (2) recommended elements, and (3) application guidelines (to document how specific elements are used).

Cole, Timothy, et.al. *Implementation of a Scholarly Information Portal Using the Open Archives Initiative Protocol for Metadata Harvesting.* Final Report to the Andrew W. Mellon Foundation. July 25, 2003 <http://oai.grainger.uiuc.edu/FinalReport/Mellon_FinalReport.doc>