SIGIR Forum, Fall 2003 (Vol 37 #2)

Report on the Metadata Harvesting Workshop at JCDL 2003

Simeon Warner[†] and Michael Nelson[‡]

[†] Cornell Information Science, 301 College Ave, Ithaca, NY, USA (simeon@cs.cornell.edu)
[‡] Old Dominion University, Norfolk, VA, USA (mln@cs.odu.edu)

1 Introduction

The "OAI Metadata Harvesting Workshop" was held on Saturday 31 May as part of JCDL 2003. There were 11 participants including OAI service provider implementers, data provider implementers and researchers, from both the US and Europe. Most participants made short presentations to highlight interesting topics or issues, and time was allocated for discussion following each presentation. This report extends the brief summary earlier presented in [20]. Further details, including slides from the presentations, are available from the workshop web site [19].

The participants were: Donatella Castelli (castelli@iei.pi.cnr.it), Naomi Dushay (naomi@cs.cornell.edu), Ed Fox (fox@vt.edu), Tom Habing (thabing@uiuc.edu), Kat Hagedorn (khage@umich.edu), Terry Harrison (tharriso@cs.odu.edu), Xiaoming Liu (liu_x@lanl.gov), Michael Nelson (mln@ils.unc.edu), Heinrich Stamerjohanns (stamer@uni-oldenburg.de), Jewel Ward (jewelw@lanl.gov), and Simeon Warner (simeon@cs.cornell.edu).

2 Topics

The sections that follow are divided according to the topics presented by each participant. While we give the names of the participant introducing each topic, no attempt has been made to attribute comments and ideas in the resulting discussions except where they relate to particular experiences.

2.1 Deduppping

Tom Habing introduced several issues related to deduppping: 1) the use of information in provenance containers in re-exporting data (including indication of whether data is re-exported verbatim or modified); 2) the recursive structure of harvesting and the possibility of infinite recursion (cycles); and 3) problems dealing with multiple versions of a given record or resource. It was agreed that harvesters and aggregators must be careful not to harvest their own material. One way to do this is to check information in provenance containers.

The deduppping problem is not new to OAI and there are many different opinions about what constitutes a duplicate. Discussion touched on a number of different approaches including calculation of similarity metrics, hashing, vector space models, string distance models, and the use of explicit provenance information.

The notion of set-level provenance was mentioned. The current provenance container does not support this but obviously many of the same concepts apply: there is a baseURL for the originating site, there is a last harvest date, and one could use the attribute altered={true|false} to indicate

if any records have been altered. It was also suggested that the existing friends container can be abused to indicate sites that an aggregator harvests from, though this use requires out-of-protocol understanding. It would also be possible to use an about container to indicate this information. A service showing a harvesting map of which aggregators harvest from where would be useful, perhaps with output in some machine readable form.

2.2 Deleted records

Naomi Dushay talked about the need for aggregators and service to use both 'deleted' status and to track records and items that simply 'go away'. To be really sure one must reharvest to find deleted records (which can be done with ListItems requests, saving bandwidth over ListRecords requests). Aggregators can add value by tracking deleted records when then harvest from data-providers that don't store or reveal that information. They could even re-expose a listing of known deleted records as an additional service.

2.3 Poor quality harvested metadata

Naomi Dushay described a 'knowledge gap' between OAI concepts and the technical details of the OAI-PMH. Areas often not understood include protocol use, XML and metadata. Documentation along the lines of the "... for dummies" series was suggested. Several participants described training efforts and all agreed that a key problem is the misconception that "OAI equals Simple Dublin Core (DC)". Constructive suggestions included adding more introductory material, updating the bibliography, adding a list of products that support OAI (including commercial ones), and adding links to metadata primers on the OAI website.

2.4 Integration of OAI systems

Ed Fox describe how OAI is used as infrastructure in a number of projects and cited the draft static repository specification as an important development. He suggested the need for care to ensure the long-term viability of OAI and perhaps the need for a general OAI meeting to bring together many of the disparate efforts now using the OAI-PMH. Projects providing bridges between OAI-PMH and other standards, such as ZMarco [18] which implements a Z39.50↔OAI gateway, are also important for the integration of OAI with other systems.

There was some discussion of the use of OAI in secure and restricted situations. Sandia National Laboratory and the USAF are already using OAI in a secure setting and a number of projects are using password and IP-based authentication to restrict access to OAI servers. Some publishers also use OAI in restricted contexts; perhaps exposing more metadata privately than publicly. Many publishers consider detailed metadata to be of commercial value and currently sell it to libraries. Even the ACM will give away only title and author data – the rest is considered to be of commercial value.

2.5 Rights, restrictions and metadata

Kat Hagedorn drew on experience with OAIster [12] to talk about rights issues within OAI. As an illustration she reported that OAIster users are sometimes surprised to find that metadata records in OAIster refer to digital objects that are not freely available.

There is often rights information included in free-text form but this cannot be used by automated agents. Should there be protocol support for rights information? There was consensus that information about *resources* should be in the metadata and should not be part of the protocol. The situation is less clear for rights information about the metadata themselves.

Once again the notion of a separate service to support rights management was suggested. The complexity of digital rights management issues means that many proposed formats and architectures are complex (see, for example, [1, 3]). Such complexity is somewhat at odds with the OAI approach.

The possibility of set and repository level rights statements was considered. Overall it was felt that no single 'one size fits all' solution would be possible as different communities would need different languages and semantics for rights statements and that it would unwise for the OAI to invent any rights language. As a community exemplar it was noted that the OLAC [13] community is considering rights expressions in its metadata [14].

2.6 Automated repository discovery

Kat Hagedorn discussed issues related to automated repository discovery. Two existing discovery methods are the central OAI registry of sites and the (underused) friends mechanism. Having discovered repositories to harvest from, one problematic issue for service providers is how they should know when a repository is 'dead'? Also, what about repositories that are known to be temporary? There was some discussion of mechanisms to indicate when a repository moves, goes off-line or is only temporary but no consensus was reached. It was, however, agreed that temporary or 'test' repositories should not be registered with the central OAI repository or appear in friends lists.

The idea of an additional service that classifies and tracks repositories was suggested. It could predict performance and availability based on logs and present some sort of 'quality map'. Many other characterizations could also be implemented such as size, similarity or 'aboutness'.

2.7 Synchronization problems

Xiaoming Liu gave a summary of the issues he had presented in his full-paper [5]. The key issue is that different repositories use different update schedules and in order for a service-provider to maintain synchronization with a data-provider in an efficient manner, it needs to know the update schedule. Suggestions included adding an RSS synchronization container and investigation of 'subscribe' and 'push' models.

There was consensus that 'push' models weren't really in the spirit of the OAI-PMH. The idea of a third-party service to provide monitoring and characterization (updates frequency, historical record, and style: bursty, irregular or uniform) of repositories was popular and it was pointed out that some aggregators already provide much of this functionality with their log files (see, for example, [4, 2]).

2.8 OAI for resource exchange

Michael Nelson described ideas that might fall under the banner of "OAI-PRH – OAI Protocol for Resource Harvesting". He cited experience with the NACA↔MAGiC [9, 7] mirror system. In this system the OAI data-provider and service-provider understand how a URL for the PDF resource is encoded withing the metadata and the resource harvesting occurs outside OAI-PMH.

There was consensus that experiments with the use of METS [8] to provide XML encoding for both digital resources and metadata would be worthwhile. These METS objects could then be exchanged via OAI-PMH without modification and it would be interesting to see what problems were encountered. Discussion touched on the idea that OAI-PMH might be used in an archival architecture, perhaps in conjunction with LOCKSS [6] or other systems.

2.9 Metadata normalization

Heinrich Stamerjohanns reported experience working with PhysDoc [15] which is based around the Harvest system, and with the Max Planck EDoc [**?**] system. PhysDoc uses metadata tags incorporated within HTML pages and they have put considerable effort into normalizing metadata which has necessitated the development of many of heuristics. Naomi Dushay reported similar experience with metadata within the NSDL [10].

It was suggested that software to comment on metadata quality would be extremely useful. Since the workshop, Heinrich Stamerjohanns has produced DC-Checker [16], software that does just this. One enters an OAI baseURL into a web form and DC-Checker will harvest DC metadata and run several checks on it, providing a report back to the user.

2.10 Teaching OAI

Heinrich Stamerjohanns gave a breakdown of experience and issues related to teaching OAI. First, he pointed out that technical details are not important for most people. There are many existing implementations but there are problems deploying them. Second, he has found that discussions always lead to two issues: 1) organizational difficulties, and 2) Simple Dublin Core is not the only metadata format supported by the OAI-PMH. He concluded that teaching must focus on more than just harvesting and must include discussion of XML and encoding issues. Several participants mentioned education efforts as part of their work (see, for example, [11, 10]).

2.11 Definition and use of responseDate

Simeon Warner detailed problems with the current definition and use of the responseDate. There is currently a discrepancy between the schema and specification: the schema permits any valid XML datetime whereas the specification stipulates that it must be in UTC and be expressed using 'Zulu' form. There was agreement that this should be corrected but considerable discussion about how harvesters should deal with bad values. What about bad synchronization between responseDate and record datestamps? What about export of bad responseDate values in provenance containers? Much of this cannot be schema enforced and the conclusion was that better education and better validation services are necessary to reduce the problem.

A few other minor problems with the OAI-PMH specification were noted and there was agreement that these should be addressed. Use of the toolkit container [17] was mentioned as one way to include information about the toolkit or kits used at an OAI site. Knowledge of the toolkits used to build a site gives information about the facilities it will support and any known problems/issues.

3 Summary

Several recurring themes emerged as the workshop progressed. These were: the need for better documentation, issues of metadata quality, and ideas for additional services.

The need for additional and improved documentation was mentioned many times. OAI-PMH servers and services are now being deployed using various programs and toolkits which do not require detailed technical knowledge of OAI-PMH. It is apparent that we need documentation suitable for these users. Both the OAForum and the NSDL are working on OAI documentation and tutorials. These and other resources need to be linked from the OAI website. At an even higher level, some very basic education and dissemination is required to dispel a number of persistent misunderstandings about exactly what the OAI framework does and does not provide. In particular, the OAI must address the common misunderstanding that "OAI is just about Simple Dublin Core". It must instead promote the message that Dublin Core is mandated only to provide a baseline for OAI-wide interoperability and the use of other metadata formats is encouraged.

Representatives of harvesting projects, especially the NSDL and PhysDoc, reported widely varying metadata quality and said that metadata normalization is essential. Both have developed heuristics for data cleaning and find it necessary to hand-customize these algorithms on a per-repository basis. While these problems are not new to the OAI there is clearly considerable scope for development of tools and improved practices to support the creation of services on metadata from many sources and of varying quality.

There were several suggestions for additional infrastructure components to provide specialized services. In all cases, discussion lead us to believe that these could be provided at the service-provider level of OAI-PMH, without need to change the protocol. The suggestions included services to aid the identification of duplicate records; to create maps of repositories and proxies; to track deleted records in upstream repositories; and to classify repositories based on content, size, update schedule, availability, etc.

Participants mentioned numerous private and intra-net OAI-PMH implementations, making it apparent that the OAI-PMH is used more widely than the number of registered repositories and services suggests. The community building work of the Open Languages Archives Community (OLAC) was admired and it was agreed that more community-specific development is required within the e-prints community. Along with the development of higher level documentation, the draft static repository specification was considered an important way to encourage participation by further lowering of the barrier to OAI interoperability.

References

[1] Mark Bide. Open archives and intellectual property: incompatible world views? *Report prepared for the Open Archives Forum*, 2002. URL: `http://www.oaforum.org/otherfiles/oaf_d42_cser1_bide.pdf`.

[2] Tim Brody. Celestial status (OAI-PMH harvesting status and logs). URL: `http://celestial.eprints.org/cgi-bin/status`.

[3] Renato Iannella. Digital Rights Management (DRM) Architectures. *DLib*, 7, 2001. URL: `http://www.dlib.org/dlib/june01/iannella/06iannella.html`.

[4] Alan Kent. OAI Harvester Crawling Status (OAI-PMH harvesting status and logs). URL: `http://www.mds.rmit.edu.au/~ajk/oai/interop/summary.htm`.

[5] Xiaoming Liu, Kurt Maly, Mohammad Zubair, and Michael L. Nelson. Repository synchronization in the OAI framework. *Proceedings of the 2003 Joint Conference on Digital Libraries (JCDL 2003)*, 2003. URL: http://csdl.computer.org/comp/proceedings/jcdl/2003/1939/00/19390191abs.htm.

[6] LOCKSS: Lots of copies keeps stuff safe. URL: http://lockss.stanford.edu/.

[7] MAGiC: Managing Access to Grey Literature Collections. URL: http://www.magic.ac.uk/index1.html.

[8] METS: Metadata Encoding & Transmission Standard. URL: http://www.loc.gov/standards/mets/.

[9] National Advisory Committee for Aeronautics Report Server. URL: http://naca.larc.nasa.gov/.

[10] NSDL: The National Science Digital Library. URL: http://www.nsdl.org/.

[11] OAForum: The Open Archives Forum. URL: http://www.oaforum.org/.

[12] OAIster: An OAI service-provider. URL: http://www.oaister.org/o/oaister/.

[13] OLAC: Open Langauge Archives Community. URL: http://www.language-archives.org/.

[14] Rights element in OLAC metadata format. URL: http://www.language-archives.org/OLAC/olacms.html#Rights.

[15] PhysDoc. URL: http://www.eps.org/PhysNet/physdoc.html.

[16] Heinrich Stammerjohanns. DC-Checker. URL: http://harvest.physik.uni-oldenburg.de/dc/index.html.

[17] Hussein Suleman. XML schema for the toolkit container. URL: http://oai.dlib.vt.edu/OAI/metadata/toolkit.xsd.

[18] Tom Habing *et al.* ZMarco – a Z39.50 to OAI gateway. URL: http://zmarco.sourceforge.net/.

[19] Simeon Warner. Website for the "OAI Metadata Harvesting Workshop" at JCDL 2003. URL: http://www.cs.cornell.edu/people/simeon/workshops/JCDL2003/index.html.

[20] Simeon Warner. Report on the "OAI Metadata Harvesting Workshop" at JCDL 2003. *DLib*, 9, 2003. URL: http://www.dlib.org/dlib/july03/07inbrief.html#WARNER.