RePEc and OAI and AMF

Thomas Krichel

Palmer School of Library and Information Sciences
Long Island University

http://openlib.org/home/krichel

2001–09–17

about this talk

- assumes some level of familiarity with RePEc
- please interupt me at any time!
- does not represent an official statement from anybody

RePEc

- largest distributed academic digital library in the world
- amateur effort
- no new money coming in

RePEc strengths

- pioneered distinction between provision and usage of digital library data
- high degree of labour division
- relational features
  - separate personal registration
  - separate institutional registration
- separate log treatment
- starts to make inroads in the peer-review process

weaknesses of RePEc

- no central registration of series
- no relational checking
- template syntax of ReDIF
- iso-latin-1 character set of ReDIF
- no long-run funding for development work
- isolated system that is not taken up in other disciplines

"Open Archive Initiative" (OAI) and AMF "Academic Metadata Format" (AMF) work can help to overcome some of these problems.

Open Archive Initiative History

- Ginsparg, Luce and Van de Sompel (1999)
  "The purpose of this call is the mobilisation of a core group to work towards achieving a universal service for author-archived literature"
- emphasis on a pragmatic level of interoperability
- call for a meeting in Santa Fe, "protoproto" prepared to stimulate discussions

UPS protoproto

Van de Sompel, Krichel, Nelson and others build a federated eprint service. They find that the main problems of interoperability between eprint initiative are

- poor metadata
- no uniform identifier structure
- unclear legal terms and conditions
- lack of incremental harvesting

This work was done to prepare the Santa Fe meeting, and published in D-Lib magazine.

Santa Fe meeting 1999–10

- Representatives of arXiv, cogprints, Highwire, NCSTRL, NDLTD, RePEc, SLAC/SPIRES and others
- chaired by Lynch and Waters
- sponsored by CLIR, LANL and SPARC

basic concepts

- "Managed" or formal e-print archive; not papers on the web like a RePEc archive
- Open e-print archive means that there is a machine interface like the one provided by Guildford protocol
- "record" can be metadata or metadata & full text like ReDIF with full-text link.

Open Archives business model

- Inspired by RePEc initiative
- Separation between data providers and service providers
  - Many archives (data providers)
  - One logical database (in the RePEc case)
  - Many services (service providers)

RePEc now has over 200 archives and about 10 user services, OAI about 30 data providers and 3 service providers, but it is much younger.

Open Archives metadata model

The main extension that OAI offers over RePEc is the presence of any number of metadata formats.

- "Native" metadata format
- Unqualified DC required metadata for each record
- Not all records need to exist in all formats

Records may be grouped into sets.

Each record has a datestamp.

OAI Technical model: OAI Protocol for Metadata Harvesting

Compatible archive responds to five http "verbs", encoded *baseurl*`?verb=`*verbname*`&`[*argname*`=`*argval*]*... where

- *baseurl* is the location of the OAI v1.1 protocol as registered at openarchives.org
- *verbname* is the name of the verb
- *argname* is the name of the attribute
- *argval* is the value of the attribute

OAI PMH: verbs

verb	*arguments*	
	required	*optional*
Identify		
ListSets		
ListMetadataFormats		identifier
GetRecord	identifer metadataPrefix	
ListRecords	metadataPrefix	set,before,after

OAI PMH: Sets

- ONLY for a local community to implement selective harvesting
- there can be zero or more sets in an archive
- records can exist at interior nodes in the set hierarchy
- asking for records in a set returns records in the set and in all its subsets.

currently not used in RePEc OAI archive, but could be applied to archives and series

OAI PMH: XML Schema usage

All reponses to OA PMH requests are controled by XML Schema instances.

This includes the header of the reponse as well as the metadata instances that the header may include.

Note that version 1 of the OA PMH protocol had to be revised to version 1.1 in July 2001 following a change in the XML Schema specification.

RePEc OAI PMD implementation

- spare-time job by Thomas Krichel
- runs on arcano, the first machine that WoPEc bought with JISC money
- move to carnation.ier.hit-u.ac.jp delayed by hardware problems

Morton protocol

A protocol to store files on a disk that come from RePEc but are stored for delivery of OAI PMH data provider. Idea: make everything available via ftp, too, for us oldies.

`etc/` files for deliver of general metadata information
`lib/` files of metadata records
`var/` files that have the identifier per datestamp
`bin/` implementation software

There is a written-down version of this, but not up-to-date.

directory `etc/`

- `Identify`
- `ListMetadataFormats`
- `ListSets`

the latter one will need moving once we represent RePEc series and archives as sets.

directory **var/**

File name of type *yyyy-mm-dd*, where *yyyy* is the year, *mm* is the month, and *dd* is the day.

These contain identfier lines, e.g.
<identifier>RePEc:rus</identifier>

These used in the ListRecords and ListIdentifier verbs.

directory **lib/**

has a directory **lib/amf/** for AMF, and a directory **lib/oai_dc/** for OAI DC.

Each of those has a directory **RePEc**, followed by a subdirectory *arc*, where *arc* is the archive code, followed by a subdirectory *series*, where *series* is the six-character series code, if available. The subdirectories contain one XML file per ReDIF record.

That is a lot of little files, be careful about inodes capacity!

directory **bin/**

Software to run all this

- **red2amf** a perl script that translates ReDIF to AMF and OAI_DC, the unqualified DC as required by the OAI. It build the contents of **var/** anew. For files in **lib/** anew. For files in **lib/**, it creates a temporary file, and if the temporary file differs from the installed file, it itstalls the temporary file. Very I/O intensive, verrry slow.

- **morton.php** a PHP script that prepares OAI responses. It is used as the root document of the apache virtual server that implements the OAI PMH responses.

OAI vs Guildford protocol

OAI PMH much more involved than Guildford protocol. In the near future, it will be prohibitively costly to replace the Guildford protocol with OAI PMH on all RePEc archives. That is a non-starter.

Gateway will be maintained centrally. That gateway however, may also be useful to feed other AMF-based data into RePEc.

OAI critique

OAI set out, at the beginning, to work on the interoperability of eprint archives as a means to promote their universal use.

Now all it has produced is a "DC and maybe other metadata format transporter". This is of little use for us.

This critique overlooks that the main players are still related to the eprint or the electronic theses and dissertation (ETD) movements. On the http://www.openarchives.org

> The Open Archives Initiative has its roots in an effort to enhance access to e-print archives as a means of increasing the availability of scholarly communication. Continued support of this work remains a cornerstone of the Open Archives program.

Thus we need to use the OAI to put our adgenda of academic self-organization to the forefront.

AMF, the Academic Metadata Format

Aim: to create a successor to ReDIF in XML.

- simpler to create
- more general
- use standard XML validation and API tools

Creation follows decision at workshop of the OAI Technical Committee in 2000–09. Lone leaders of AMF have been Thomas Krichel and Simeon M. Warner of arXiv.

Design rationale at http://openlib.org/home/krichel/kanda.html

More formal sematics at http://amf.openlib.org/doc/ebisu.html

AMF is simpler to create than ReDIF

There are only four record types in AMF
text – collection – person – organization.
called "nouns". Each noun is represented by an XML element.

Each noun admits XML child elments called "adjective" that qualify the noun.

There are "verbs" that allow to relate nouns.

The documentation is only 7 pages long.

AMF is more general than ReDIF

- It uses UTF–8 encoding of Unicode.
- It allows for arbitrary grouping of resource records in the collection noun.
- Currently only on resource type "text" but software or others could be added.

AMF use standard XML validation

- XML Schema based
- Schema file at `http://amf.openlib.org/2001/amf.xsd` still in need of a bit of TLC

Thank you for your attention!