# Digital Library Interoperability with the Open Archives Initiative (OAI)

Thomas Krichel

Palmer School of Library and Information Sciences
Long Island University

http://openlib.org/home/krichel

2001–09–11

## about this talk

- follows essentially a historical approach
- does not represent an official statement from anybody
- botches together various ideas from different people
- a couple of slides stolen from Michael L. Nelson
- please interrupt me at any time!

## personal introduction

- trained economist
- since 1992 activist for free online scholarship
- 1993 founded NetEc
- 1997 founded RePEc, now the largest distributed library of free scientific papers in the world
- 1999 co-founder of Open Archive Initiative (OAI)

## UPS call 1999–07

- Ginsparg, Luce and Van de Sompel
  "The purpose of this call is the mobilization 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

## 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
- Open e-print archive means that there is a machine interface
- "record" can be metadata or metadata & full text
- archive may be partitioned

## 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.

## Open Archives business model

- copied from 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.

## requirements & realizations

| | |
|---|---|
| metadata harvesting (not distributed database) | OA Dienst subset |
| namespace | fullId=archive:record |
| mandatory metadata & parallel sets | OAMS & XML transport |
| acceptable use | gentleperson's agreement |
| registration | primitive templates |

## OA technical model

- Subset of Dienst protocol used by NCSTRL
- Compatible archive respond to 4 requests
  - List-Partitions
  - List-Meta-Formats
  - List-Contents (partitionspec, file-after, meta-format)
  - Disseminate (fullID, meta-format, content-type)

## Dublin Core-ish Minimal Metadata for selective harvesting

| *mandatory* | *optional* |
|---|---|
| title | display ID [R] |
| date of accession | abstract |
| FullID | subject [R] |
| author [R] | comment [R] |
| | date for discovery [R] |

With questionable semantics :–)

## critique of the Santa Fe convention (SFc)

- Why OAMS, not Dublin Core?
- Dienst subset carries a lot of legacy from the full Dienst protocol that make it cumbersome to implement.

## developments in DL community

Interest in interoperability for a long time, stated interest of the digital library federation. there are two approaches

- union catalog
  - causes friction
- distributed search
  - high entry requirement
  - problematic to implement

## Harvard meeting 2000–05

- vision statement: SFc a new way forward for interoperability
- could the OAI develop in a more general fashion such that it can be used by different communities?
- political agenda of OAI (free access) perceived as problem

## San Antonio meeting 2000–06

- 45 people show broad range of interest leads to problem of not getting lost.
- view that SFc is a technical support infrastructure
- communities in different business and contents models can adopt the framework for interoperability

## Ithaca meeting 2000–09

- experience gained with implementing & discussing the current SFc specs
- aimed for new spec by the start of 2001
- stable for experimentation (12 to 18 months) but not definite
- hoped to minimize risks for implementors maximize chances for interoperability
- SFc+ to translate from eprint domain interoperability towards general domain interoperability

## Summary of an extraordinarily productive meeting

- OAI Dienst replaced by OA metadata harvesting protocol
- OAI ID revised
- OAMS replaced by wrapped DC
- introduction of the concept of native metadata
- generalized and marginalized partitions
- revisited registrations

## New OAI metadata

- Accession date to be renamed datestamp and stripped of semantic link to the records
- OAMS scrapped, Krichel and Warner to lead an EPMS discussion
- unqualified DC is mandatory, but empty DC may be returned
- introduction of the idea of native metadata

## Identifiers

- Identifiers point to metadata records
- OAI-specific identifiers concatenate
  - case-sensitive archive name
  - delimiter is a colon
  - anything internal to the archive appearing after that
- prefixed by OAI as a pointer to a resolution mechanism

## Sets

- replace partitions
- 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.

## OAI verbs

- Identify
- ListSets
- ListMetadataFormats ([identifier])
- GetRecord (identifer,metadataPrefix)
- ListRecords (metadataPrefix,[set],[before],[after])
- ListIdentifiers (metadataPrefix,[set],[before],[after])

## Request encoding

*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

## XML Schema usage

All responses to OAI requests are controlled by XML Schema instances.

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

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

## Registration of archives

- Central metadata format registration, names use alphanumeric and underscore
- Self-description introduced through the identify verb
- Fields of data provider templates
  - natural language name
  - description URL
  - archive id
  - maintainer (of OA interface) email
  - version of OA protocol used
  - OA base URL

## Flow Control

ListSets, ListIdentifiers, ListRecords are all allowed to return partial responses, via a combination of:

- resumptionToken: an opaque, archive-defined data string that when passed back to the archive allows the response to begin where it left off
- each archive defines their own resumptionToken syntax; it may have visible semantics or not
- 503 http status code "retry after"
- up to the harvester to understand this code and respect it, and up to the archive to enforce it

## 30+ Data providers

- http://www.openarchives.org
- arXiv, RePEc to follow soon...
- many being used for internal purposes; not registered

## 3+ service providers

- Repository Explorer at http://purl.org/net/oai_explorer
- ARC at http://arc.cs.odu.edu/

## Example data provider

- http://oai.repec.openlib.org
- RePEc archive
- AMF as “native metadata”
- mirrored on this laptop

## Future work

A technical committee of 16 people—operating on a closed mailing list—has been formed to develop a version 2 of the specification. This version is expected to be available in April 2000.

Version 2.0 will not expand the scope of the protocol but only improve on the current specification within the current scope.

## timeframe

| | |
|---|---|
| 2001–09–01 | circulate initial issues to be address |
| 2001–09–21 | prioritize issue list |
| 2001–10–01 | divide issue list among committee members |
| 2001–11–01 | post initial work to implementors for feedback |
| 2001–12–03/07 | meeting of technical committee |
| 2001–12–10 | redrafting starts |
| 2002–02–01 | alpha testing of new protocol starts |
| 2002–04–01 | release of version 2 |

## issue list draft

- error handling
- SOAP and wsdl
- harvesting granularity
- mandatory and non-qualified DC
- result set filtering
- set semantics
- XML Schema
- flow control
- result set cardinality
- multiple metadata instances returned

## Issues that are not on the official issues list draft

- terms and conditions of usage
- distinction between data provider and service provider becomes blurred when digital library interoperability becomes complex (e.g. RePEc)
- updates, additions and deletion of records not always clearly handled at the data provider level
- Distinction between data and metadata

Thank you for your attention