1999 co-founder of Open Archive Initiative (OAI) •

Free scientific papers in the world •
1997 founded RepRFP, now the largest distributed library of •
1993 founded Netec •
Since 1992 active for free online scholarship •
Trained econometric •

Personal introduction

Please interrupt me at any time •
A couple of slides stolen from Michael L. Nelson •
Bouches together various ideas from different people •
Does not represent an official statement from anybody •
Follows essentially a historical approach •

About this talk

2001-09-11
http://opendc.org/thome/kitchen

Long Island University
Palmer School of Library and Information Sciences

Thomas Kitchen
Digital Library Interoperability with the Open Archives Initiative
(0AI)
archive may be partitioned

“record” can be metadata or metadata & full text

Open e-print archive means there is a machine interface

“Managing” or formal e-print archive: not papers on the web

basic concepts

Sponsored by CIR, LANL and SPARC

chaired by Lynne and Waters

NDLTD, RePEc, SLAC/Spires and others
Representatives of archivists, archivists, Librarians, NCSFRL

Santa Fe meeting 1999-10

call for a meeting in Santa Fe

emphasizes on a pragmatic level of interoperability

archived literature
to work towards achieving a universal service for author

"The purpose of this call is the mobilization of a core group

Ginsberg, Luce and van de Somerd
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: absent subset
mandatories metadata & parallel sets
acceptable use
transport agreement
protocol that make it cumbersome to implement.

- Dense subject carries a lot of legacy from the full Dublin

- Why OAIS, not Dublin Core?

Critique of the Semantic Text Enhancement (SfC)

With questionable semantics :(

Dublin Core-Enhanced Metadata for Selective Harvesting

- Disseminates (full-text, meta-formats, content-type)
- Lists-Contents (partitions, file-sets, meta-formats)
- Lists-Meta-Formats
- Lists-Partitions

- Compatible archive response to 4 failures

- Subset of Dublin Core protocol used by NCSFRL

OA technical model
adopt the framework for interoperability

communities in different business and contexts models can

view that SFC is a technical support infrastructure

not getting lost.

45 people show broad range of interest leads to problem of

San Antonio meeting 2000–06

can be used by different communities

could the Gov develop in a more general fashion such that it

vision statement: SFC a new way forward for interoperability

Harvard meeting 2000–05

— problematic to implement

— high entry requirement

— distributed search

— causes friction

union catalog

digital library federation. there are two approaches

interest in interoperability. for a long time, stated interest of the

developments in DL communities

political agenda of Gov (free access) perceived as problem
Introduction of the idea of native metadata

unqualified DC is mandatory, but empty DC may be returned

custom

OAMS scrapped, Kitchen and Warner to lead on EPSD db-

semantic link to the records

accession date to be renamed datetaken and stripped of

New OAI metadata

revised registrations

generalized and marginalized partitions

Introduction of the concept of native metadata

OAMS replaced by wrapped DC

OAI ID reverted

OAI Direct replaced by OA Metadat Harvesting protocol

Summary of an extraordinarily productive meeting

words general domain interoperability

SFC to translate from other domain interoperability to-

to interoperability

hoped to minimize risks for implementations maximize chances

stable for experimentation (12 to 18 months) but not definite

aimed for new spec by the start of 2001

real spec specs

experience gained with implementing & discussing the cur-

Inecha meeting 2000–09
List of entities (metadata\prefix\set\before\after)

List of records (metadata\prefix\set\before\after)

get record (identifiers\metadata\prefix)

list of identifiers (identifiers)

Identity

OAI verbs

all its subtypes

asking for records in a set returns records in the set and in

records can exist at interior nodes in the set hierarchy

there can be zero or more sets in an archive

ing

only for a local community to implement selective harvesting

replace partitions

seri

preixed by OAI as a pointer to a resolution mechanism

anything internal to the archive appearing after that

delimiter is a colon

case-sensitive archive name

OAI-specific identifiers concatenate

Identifiers

Identifiers point to metadata records
- OA base URL
- version of OA protocol used
- maintainer (of OA Interface) email
- archive id
- description URL
- natural language name

Fields of data provider templates
Self-description introduced through the identity verb
Central metadata format registration, name use allow-
Registration of archives

XML Schema usage

argval is the value of the attribute
argname is the name of the attribute
verbname is the name of the verb

at attribute

baseurl=value[baseurl=argname=argval]...where

Request encoding

Specification
Note that version 1 of the OA API protocol had to be revised to instances that the reader may include.
This includes the header of the response as well as the metadata

All responses to OA requests are controlled by XML Schema


3+ service providers

30+ data providers

and up to the archive to approve it
dep the hanlader to understand this code and respect it.

503 http status code "Retry after"

may have visible semantics or not
each archive defines their own resumptionToken syntax: it

to begin with, if it is set off
partial responses, via a combination of:

lists, listeners, listeners are all allowed to return

Flow Control
<table>
<thead>
<tr>
<th>Timeframe</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2002-04-01</td>
<td>release of version 2</td>
</tr>
<tr>
<td>2002-02-01</td>
<td>debugging testing of new protocol</td>
</tr>
<tr>
<td>2001-07-07</td>
<td>stars</td>
</tr>
<tr>
<td>2001-12-10</td>
<td>meeting of technical committee</td>
</tr>
<tr>
<td>2001-11-01</td>
<td>post initial work to implementations</td>
</tr>
<tr>
<td>2001-10-01</td>
<td>for feedback</td>
</tr>
<tr>
<td>2001-09-21</td>
<td>provide issue list</td>
</tr>
<tr>
<td>2001-09-01</td>
<td>circulating initial issue to address</td>
</tr>
</tbody>
</table>

**Future work**

- mirrored on this laptop
- AMF as "native metadata"
- RFP of archive
- http://example.provider.org
Thank you for your attention

- Distinction between data and metadata
- handled at the data provider level
- Updates, additions and deletion of records not always cleanly

Issue: (c. Report)
- Complex when digital library interoperability becomes
- Distinction between data provider and service provider be-
- Terms and conditions of usage

Issues that are not on the official Issues List draft

- Multiple metadata instances returned
- Result set cardinality
  - How controlled
  - XML schema
  - Set semantics
- Not set filtering
- Mandatory and non-qualified DC
- Harvesting granularity
- SOAP and WSDL
- Error handling