Workshop on OAI and peer review journals in Europe
Geneva, Switzerland – March 22nd to 24th 2001

Herbert Van de Sompel
Cornell University
Computer Science – Digital Library Research Group

OAI metadata harvesting specifications
a brief history of the OAI

herbert van de sompel
The Open Archives Initiative has been set up to create a forum to discuss and solve matters of interoperability between preprint solutions, as a way to promote their global acceptance.

Paul Ginsparg, Rick Luce & Herbert Van de Sompel

=> Santa Fe Convention: preprint metadata harvesting

herbert van de sompel
interest from other communities

• Digital Library Federation meetings
  ~ research library community has many materials for which they would like to ‘expose’ metadata

• OAI San Antonio meeting:
  ~ interest from librarians, publishers, others, ...
resulting actions: organizational

- establish organizational stability for the OAI:
  - institutional backing from CNI & DLF
  - steering committee: policy guidance
  - technical committee: technical specifications
  - executive group: day to day coordination
  - workshops: public dissemination, feedback
resulting actions: technical

- [09/2000] revise specifications to allow adoption beyond preprints: technical committee
- [09/2000-01/2001] compile new specifications: editing by Carl and Herbert
- [01/2001] discontinue the Santa Fe Convention
- [01/2001] release version 1.0 of the OAI protocol
the OAI METADATA HARVESTING PROTOCOL
The OAMH protocol is a low-barrier interoperability specification for the recurrent exchange of metadata between systems.
federated services

- e-print
- FTXT
- A&I
- OPAC
- image

herbert van de sompel
metadata harvesting via OAMH

harvester

metadata

user

e-print

FTXT

A&I

OPAC

image
federated services via OAMH

metadata

Author
Title
Abstract
Identifier

e-print

FTXT

A&I

OPAC

image

herbert van de sompel
core concepts in OAMH

• low-barrier interoperability
• data-provider & service-provider model
• metadata harvesting model
• shared metadata format and parallel, community-specific metadata formats

OAMH protocol
HTTP based
Reply
XML Schema
Self contained
Dublin Core
OAI harvesting tools

service provider

Datastamp
Identifier
Set

Records

data provider

herbert van de sompel
OAI harvesting tools

service provider

Supporting protocol requests:
• Identify
• ListMetadataFormats
• ListSets

Harvesting protocol requests:
• ListRecords
• ListIdentifiers
• GetRecord

data provider

herbert van de sompel
supporting protocol requests

service provider  data provider

ListM etadataFormats

service provider

data provider

ListM etadataFormats / Time / Request

• Format prefix
• Format XML schema

REPEAT

REPEAT
harvesting requests

service provider

\* from = a
\* until = b
\* set = k l m

ListRecords * metadataPrefix = dc

Repository

Data provider

ListRecords / Time / Request

\* REPEAT
- Identifier
- Datestamp
- Metadata

/REPEAT
Applications of the OAMH protocol?

- federated services [S&R, SDI, alerting, linking, ...]
- database synchronization
- harvesting the deep Web
- ...

herbert van de sompel
OAI - status
revision of specifications

• freeze specifications for 12 -18 months:
  • stable for experimentation; not definitive
  • minimize risk for early adopters
  • maximize chances for future interoperability across communities
software to run OAI repository

- eprints.org - U. Southampton
- open source metadata server - OCLC
- NT OAI server - U. Illinois
- Aleph 500 - Ex Libris
- Z39.50 ↔ OAI gateway - Virginia Tech (ongoing)
- MARC to DC convertor - OCLC
- we expect a lot more ...
- listed on OAI site

herbert van de sompel
tools to support OAI implementation

• Hussein’s Repository explorer
• W3C XSV Schema Validator
• XML Spy
• the OAI conformance tester:
  • part of OAI registration service for repositories
• listed on OAI site
modes of running OAI 1.0 repository

• mode 0:
  • no registration of repository in the OAI registry
modes of running OAI 1.0 repository

• mode 1:  
  • registration of repository in public OAI registry  
    [includes validation of replies]

  existence of the repository is visible
modes of running OAI 1.0 repository

• mode 2:
  • registration of repository in public OAI registry
  • usage of the OAI format for identifiers

existence of the repository is visible
resolver for OAI formatted identifiers
implementation status

• early adoption by preprint community
• but also by others
implementation status

• data providers:
  • 20 registered repositories (US and Europe)

  arXiv
  OCLC Thesis and Dissertations
  Perseus Digital Library
  PhysNet
  Oxford Text Archive
  Library of Congress -- American Memory
  CogPrints
  Humboldt University
  MIT Thesis
  Linguistic Data Consortium
  Resource Discovery Network
implementation status

• service providers:
  • ARC
  • Open Language Archives
  • soon to be listed on OAI site
implementation status

• Mellon Foundation funding for OAI-based projects: data providers and service providers

• NSF Digital Library interest in OAI-related projects

• Close contacts with SPARC, DLF, CN1
communication re OAI

- lists: subscribe via http://www.openarchives.org
  - oai-general list
  - oai-implementers list
- web: http://www.openarchives.org
- FAQ: http://www.openarchives.org/faq.htm
- mail: openarchives@openarchives.org

herbert van de sompel
http://www.openarchives.org

openarchives@openarchives.org