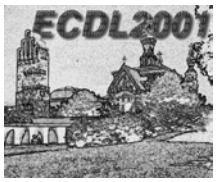
Digital Library Interoperability via Metadata Harvesting

Applying the Open Archives Initiative Protocol



Carl Lagoze, Cornell University
lagoze@cs.cornell.edu
Edward A. Fox, Virginia Tech
fox@vt.edu

Acknowledgements

- People
  - Herbert Van de Sompel
  - Dan Greenstein
  - Clifford Lynch
  - Hussein Suleman
  - Members of the OAI community
- Funding Organizations
  - Digital Library Federation
  - Coalition for Networked Information
  - National Science Foundation

Agenda

- Goal: to produce communities of OAI implementers and supporters
- Process:
  - History and context of the OAI
  - Definitions and concepts of the technology
  - Protocol details
  - Working with the OAI community
    - Tools
    - Mailing lists
    - Projects
  - Future Plans

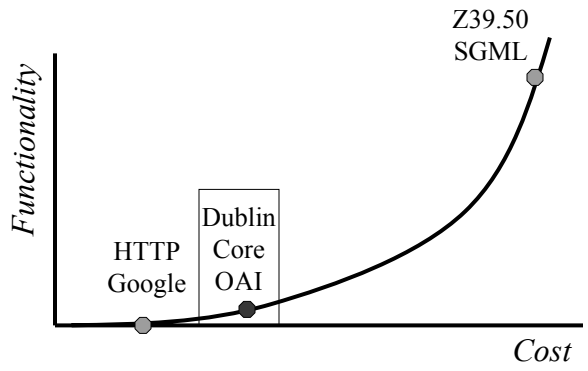
Digital Library Interoperability

Paepcke, A., C.-C. Chang, et al. (1998). "Interoperability for Digital Libraries Worldwide." Communications of the ACM **41**(4): 33-42.

A Short History of Interoperability

- Naming: URNs, Handles, DOIs
- Metadata: Dublin Core, IMS, MARC
- Search and Discovery: Z39.50, Harvest, Dienst, STARTS, SDLIP
- Object Models: Kahn/Wilensky, FEDORA, Buckets
- Encoding: SGML, HTML, XML, RDF

Interoperability Trade-offs



OAI's Location in a Broader Interoperability Fabric

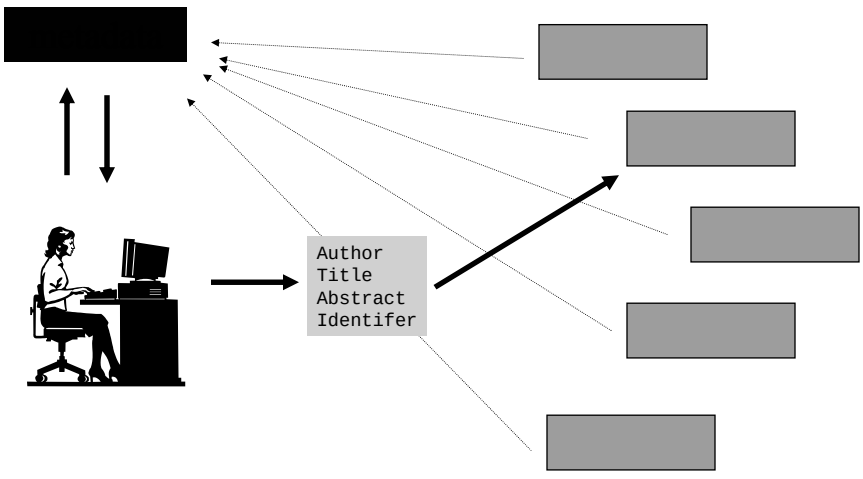
Data Structuring (XML, XML Schema)

Data Semantics (Dublin Core, other metadata)

Exchange of Structured Information

Object Access

Yes, it's about resource discovery over distributed collections



Beyond resource discovery to distributed custodianship

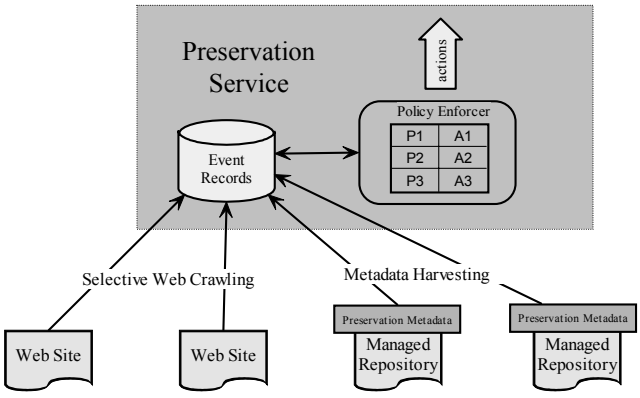
- Traditional portal (e.g., Yahoo!)
  - linkage with limited responsibility
- Hybrid Portal
  - Goal: assertion of (some semblance) of curatorial role over linked objects
  - Mechanism: sharing structured information (metadata) amongst distributed content providers

Broadening the Goals of Interoperability

The Library should selectively adopt the portal model for targeted program areas. By creating links from the Library's Web site, this approach would make available the ever-increasing body of research materials distributed across the Internet. The Library would be responsible for carefully selecting and arranging for access to licensed commercial resources for its users, but it would not house local copies of materials or assume responsibility for long-term preservation.

LC21: Digital Strategy for the Library of Congress
page 5

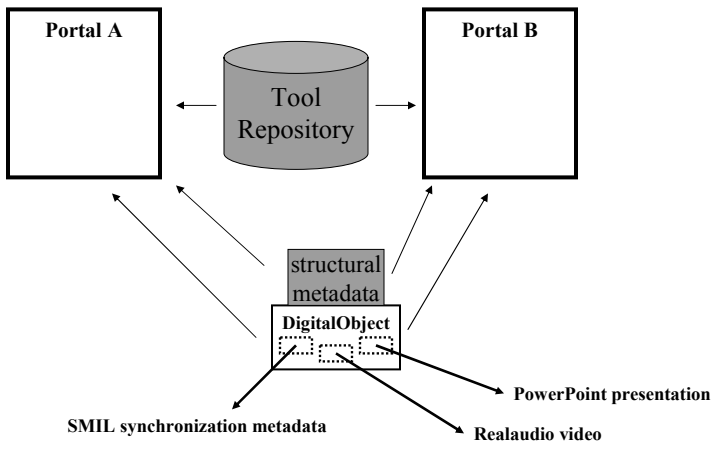
Facilitating/Monitoring Longevity of Distributed Content



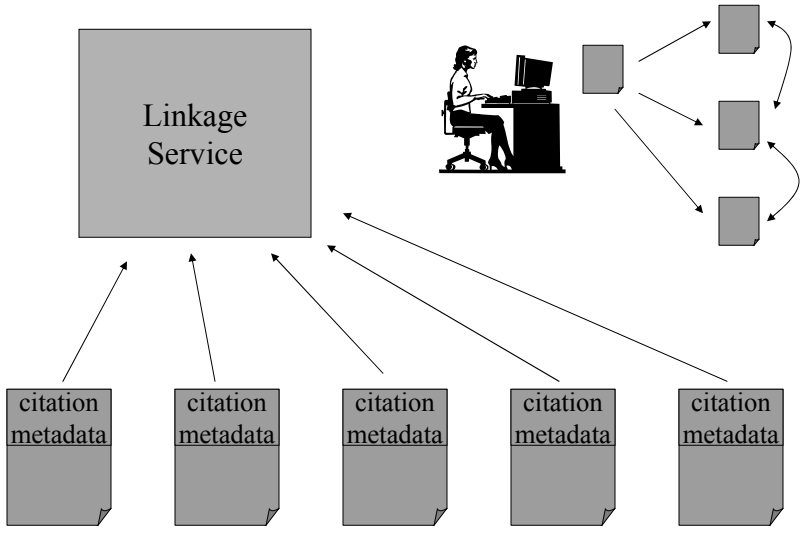
Personalization of Content

View A:
- **View slides**
- **View video**
- **View synchronized presentation using applet**

View B:
- **Get transcript of audio**
- **Search for keyword**
- **Get slides translated to French**



Cross-Repository Reference Linking



Origins of the OAI

- Increasing interest in alternative scholarly publishing solutions – e.g., LANL arXiv
- Increasing impact through federation
- UPS Mtg., Sante Fe, October 1999
  - Representatives of various E-Print, library, and publishing communities
  - Goal: definition of an interoperability framework among E-Print providers
  - Result: Santa Fe Convention, interoperability through metadata harvesting

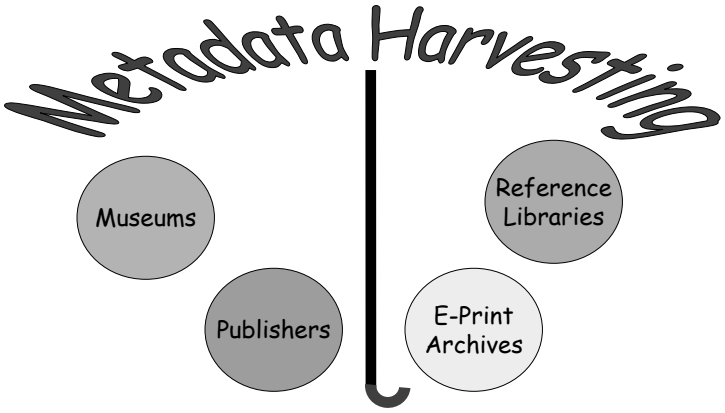
"Open" Archives

- Political Agenda?
  - Author self-archiving of E-Prints
  - "Mission" to reformulate scholarly publishing framework
- Technical?
  - Infrastructure to facilitate interoperability across multiple domains

Other Communities of Interest

- "Cambridge" Digital Library Federation meetings
  - research library community has many materials for which they'd like to 'expose' metadata
- OAI workshops
  - librarians, publishers (some), researchers, others
- Museum Community
  - Museums on the Web and CIMI

Technical Umbrella for Practical Interoperability…



…that can be exploited by different communities

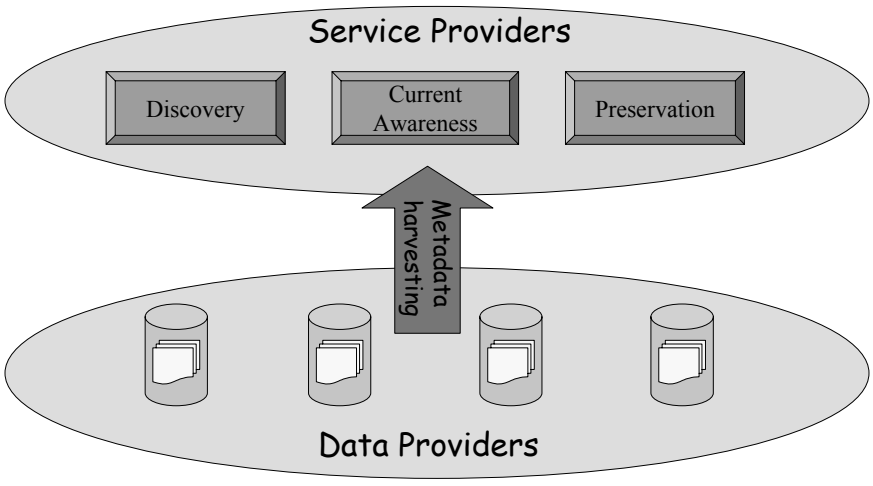
OAI Organizational Structure
Key Features

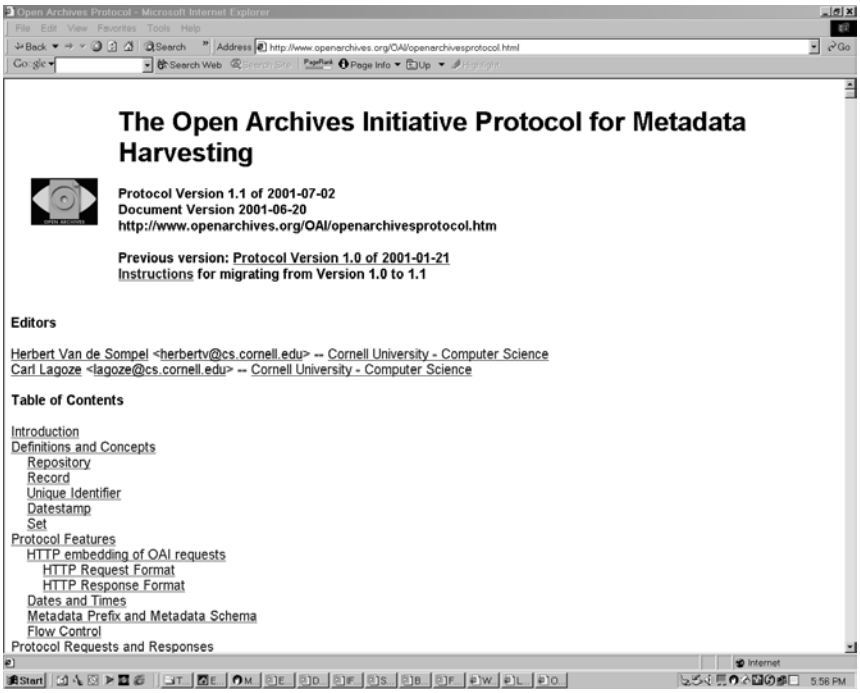
- Clear focus and scope
  - Developing and refining technical specification
  - Community building and evangelism limited to serving that goal and to encouraging widespread adoption
- Encouraging specialization and community-specific activities
- Division of responsibility
  - Executive (Van de Sompel and Lagoze)
  - Steering Committee
  - Technical Committee
  - Mailing Lists (community)

OAI Technical Infrastructure
Key Technical Features

- Deploy now technology – 80/20 rule
- Two-party model – providers (*data providers)* and consumers (*service providers*)
- Simple HTTP encoding
- XML schema for some degree of protocol conformance
- Extensibility
  - Multiple item-level metadata
  - Collection level metadata

The World According to OAI





The Open Archives Initiative Protocol for Metadata Harvesting

Protocol Version 1.1 of 2001-07-02
Document Version 2001-06-20
http://www.openarchives.org/OAI/openarchivesprotocol.htm

Previous version: Protocol Version 1.0 of 2001-01-21
Instructions for migrating from Version 1.0 to 1.1

Editors

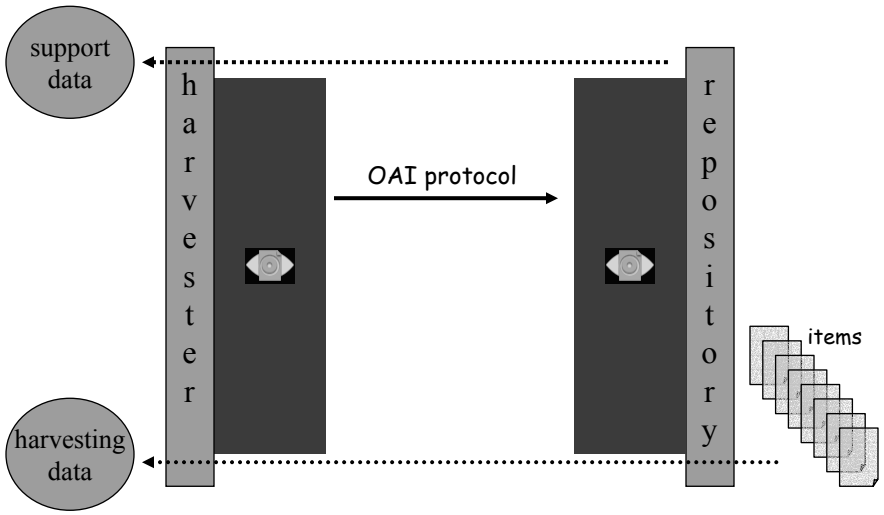
Herbert Van de Sompel <herbertv@cs.cornell.edu> -- Cornell University - Computer Science
Carl Lagoze <lagoze@cs.cornell.edu> -- Cornell University - Computer Science

Table of Contents

Key Features of the OAI Metadata Harvesting Protocol

- definitions & concepts
  - repository
  - record
  - identifier
  - datestamp
  - set
- protocol features
  - HTTP encoding
  - metadata prefix & schema
  - flow control
- protocol requests
  - supporting requests
  - harvesting requests

repository



record

```
<record>
    <header>
        <identifier>oai:eg:001</identifier>
        <datestamp>1999-01-01</datestamp>
    </header>
    <metadata>
        <dc xmlns="http://purl.org/dc">
            <title>My Example</title>
        </dc>
    </metadata>
    <about>
        <ea xmlns="http://www.arXiv.org/ea"
            <usage>No restrictions</usage>
        </ea>
    </about>
</record>
```

protocol support

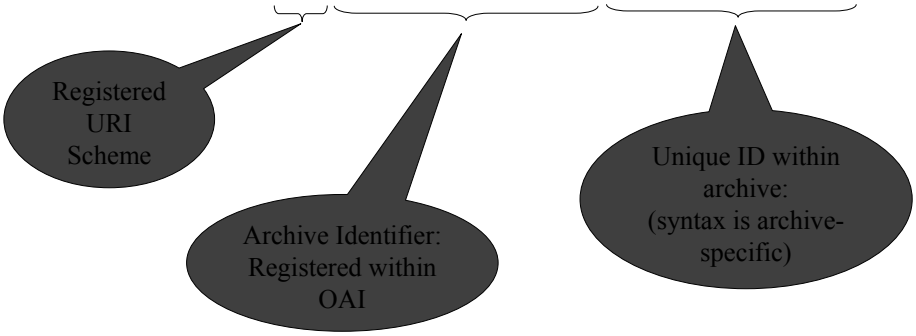
format-specific metadata

community-specific record data

identifiers

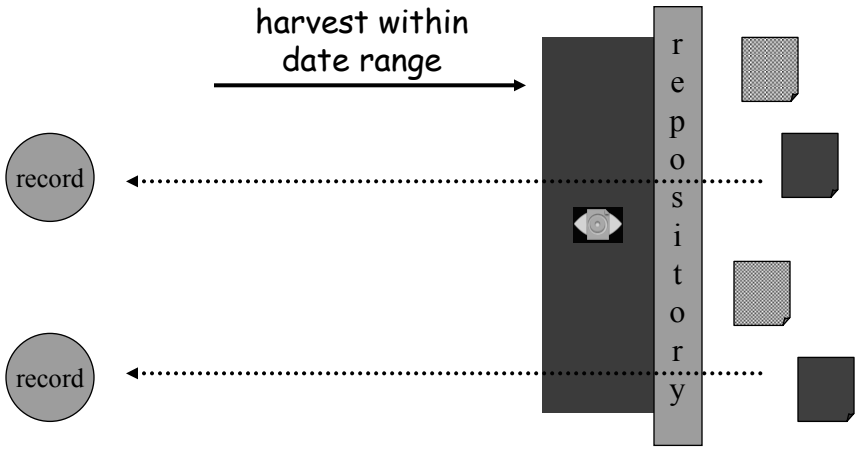
locally unique key for extracting a record from a repository

oai-identifier = oai:archive-identifier:record-identifier

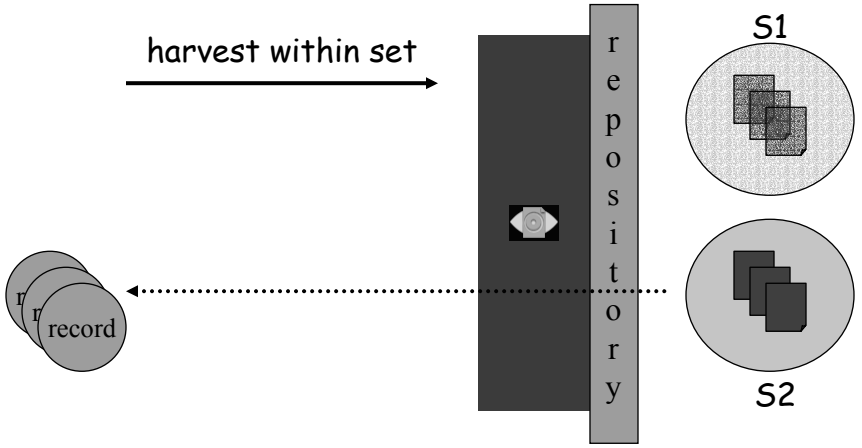


example = oai:ncstrl:ncstrl.cornellcs/TR94-1418

selective harvesting - datestamps



selective harvesting - sets



set specifics

- repositories define hierarchical organization
- each item in a repository may be organized in one set, several sets, or no sets at all
- meaning of sets or of set hierarchy is not defined in protocol
- individual communities may formulate common set configurations

HTTP encoding - requests

BASE-URL ----------> an.oa.org/OAI-script
keyword arguments --> verb=ListIdentifers&set=S1

GET
http://an.oa.org/OAI-script?verb=ListIdentifers&set=S1

POST
POST http://an.oa.org/OAI-script HTTP/1.0
Content-Length: 78
Content-Type: application/x-www-form-urlencoded
verb=ListIdentifers&set=S1

HTTP encoding - responses

```
<xml version=1.0 encoding="UTF-9" ?>
<GetRecord
      xmlns="http://oai.namespace.uri"
      xmlns:xsi="http://w3.namespace.uri"
      xsi:schemaLocation="http://oai.namespace.uri
         http://oai.schemaURL">
   <responseDate>2000-19-01T19:30:30-04:00</responseDate>
   <requestURL>http://an.oa.org/OAI-script?verb=GetRecord
         &amp;amp;identifier=oai%3AarXiv%3A0001
         &amp;amp;metadataPrefix=oai_dc</requestURL>
   <record>
      record contents
   </record
      additional records
</GetRecord>
```

xml namespaces

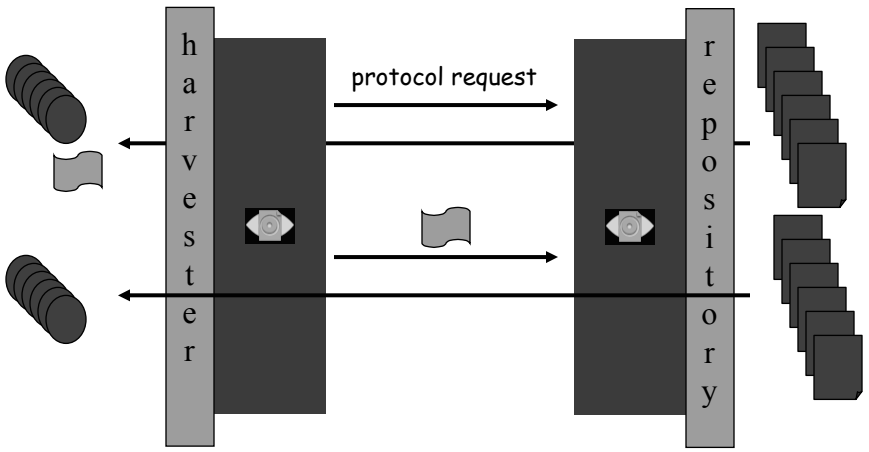
response header

response data

metadata prefix and schema

- support for harvesting multiple metadata formats
  - *metadata schema*: each format must have a validating XML schema at a publicly accessible URL (communities may define shared formats and schema).
  - *metadata prefix*: each repository maps a prefix to the schema it supports, which is used in protocol requests.
- support for unqualified Dublin Core mandatory
  - reserved schema URL at http://www.openarchives.org/OAI/dc.xsd
  - reserved prefix *oai_dc*.

flow control



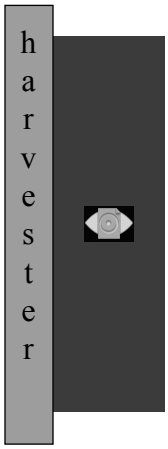
flow control specifics

- applies to all protocol requests that return lists: *ListRecords*, *ListIdentifiers*, *ListSets*
- resumptionToken is opaque
- semantics of partitioning of responses within resumption requests is undefined
- time-to-live of resumptionToken is not defined by the protocol

OAI Protocol

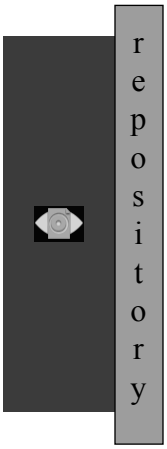
service provider

data provider



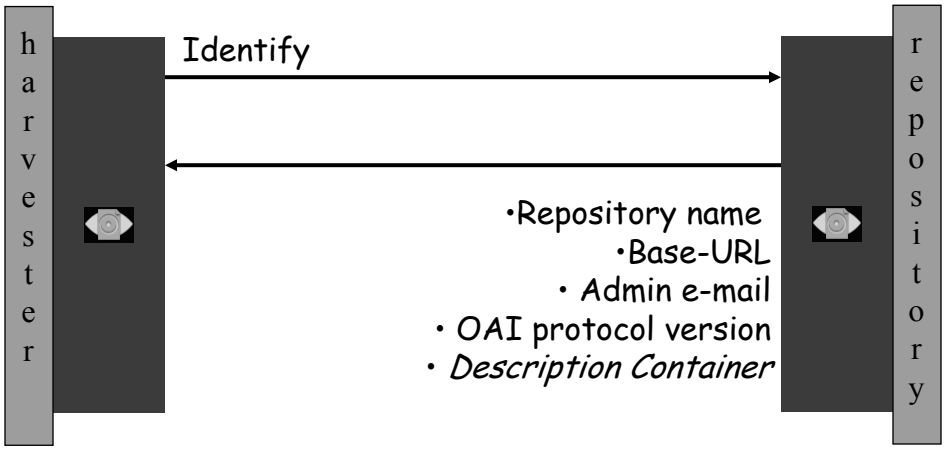
Supporting protocol requests:
- Identify
- ListMetadataFormats
- ListSets

Harvesting protocol requests:
- ListRecords
- ListIdentifiers
- GetRecord



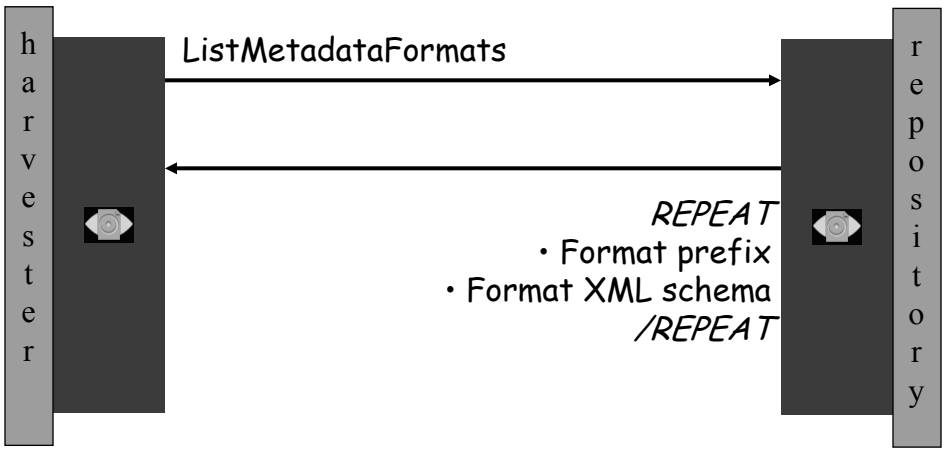
Supporting Protocol Requests

service provider — data provider

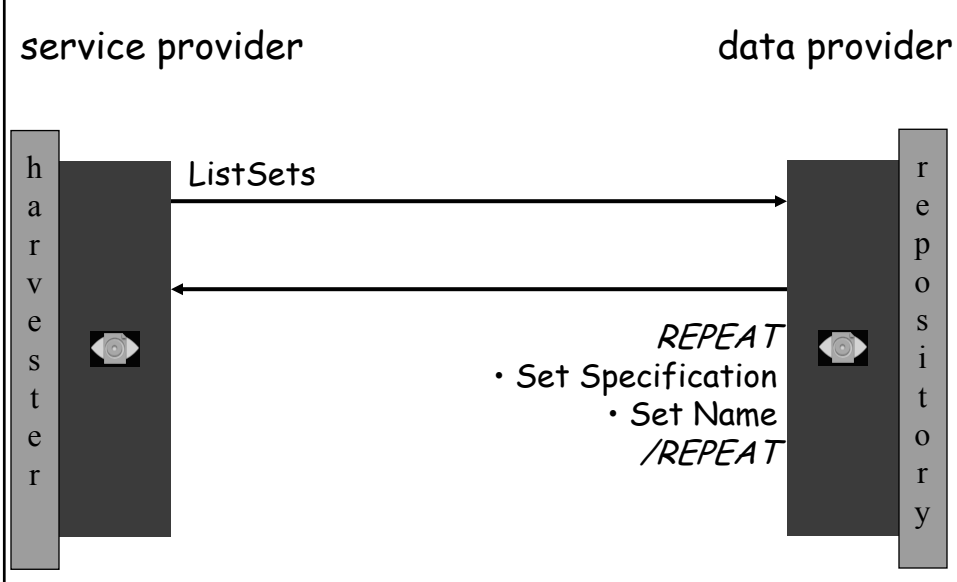


Supporting Protocol Requests

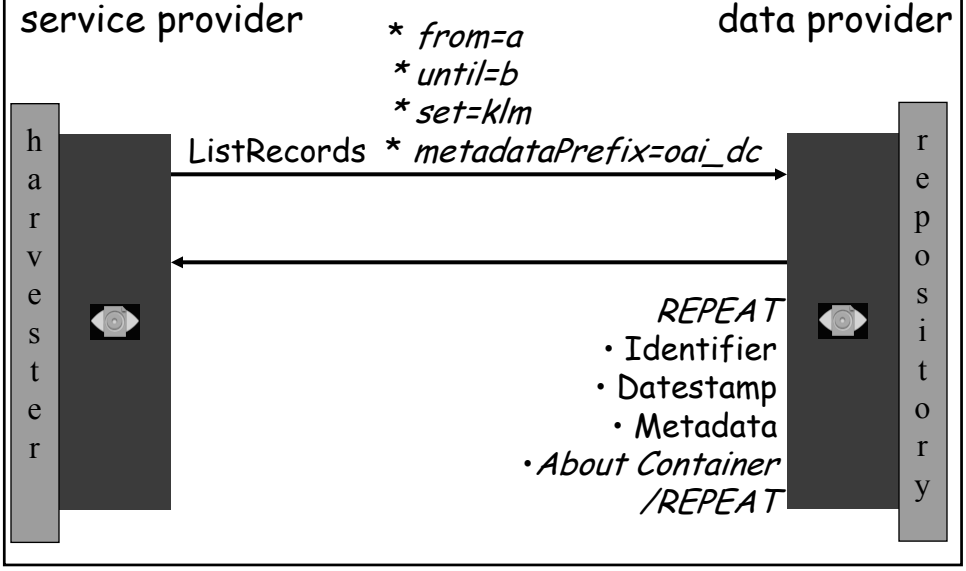
service provider — data provider



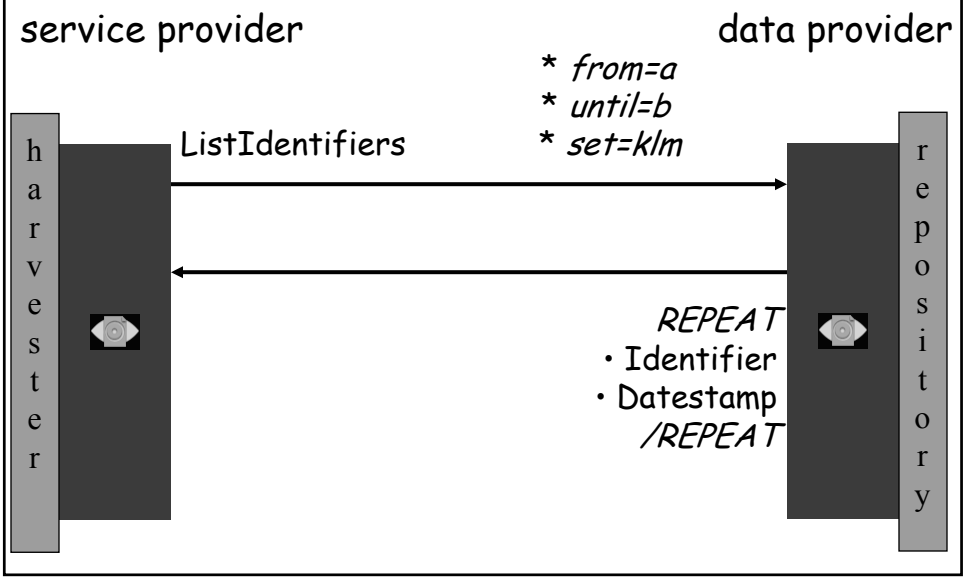
Supporting Protocol Requests



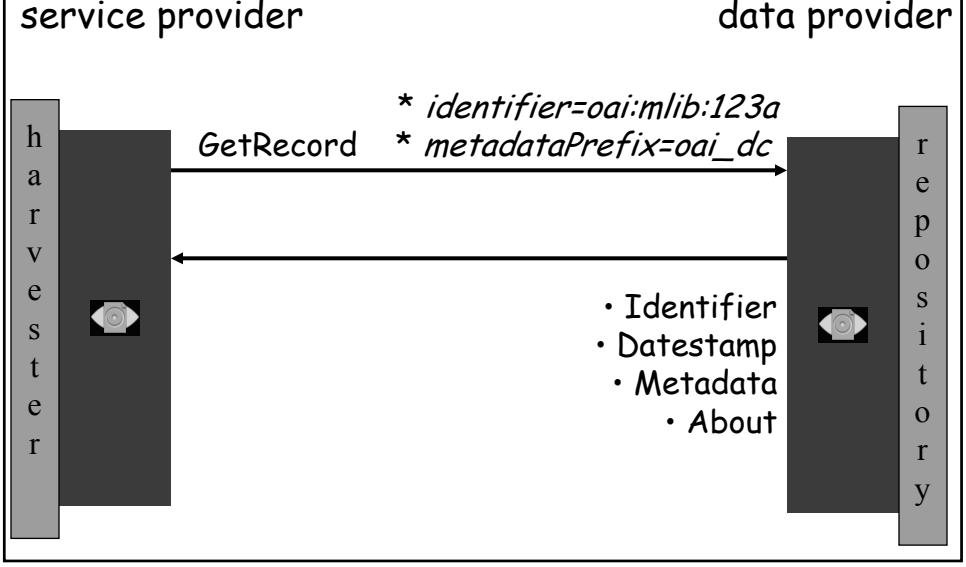
Harvesting Protocol Requests



Harvesting Protocol Requests

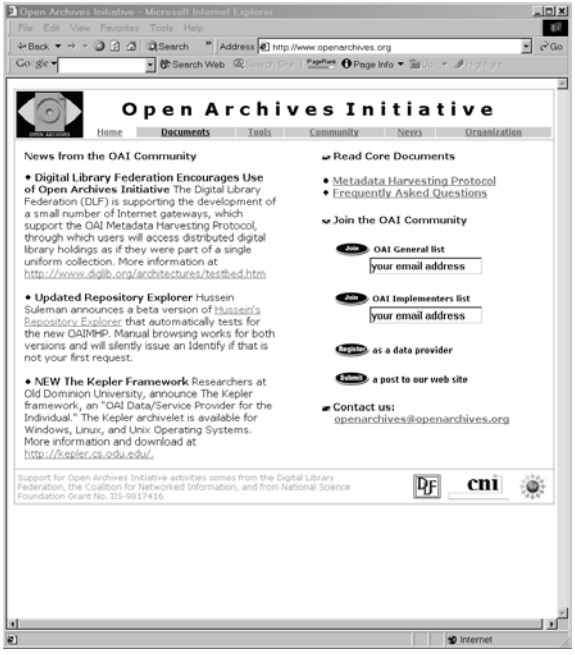


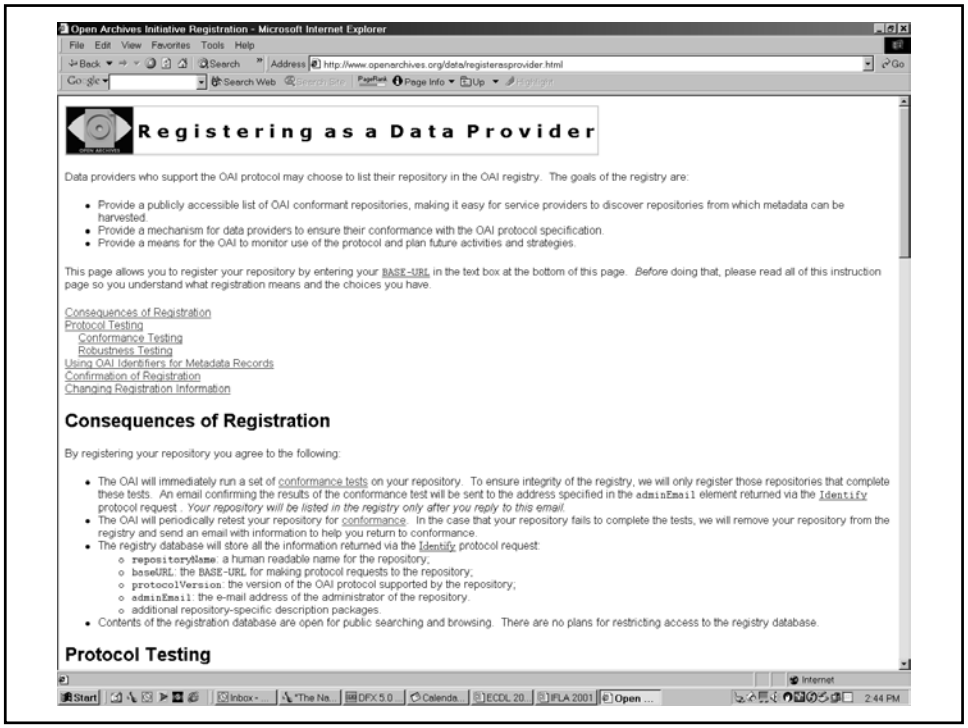
Harvesting Protocol Requests

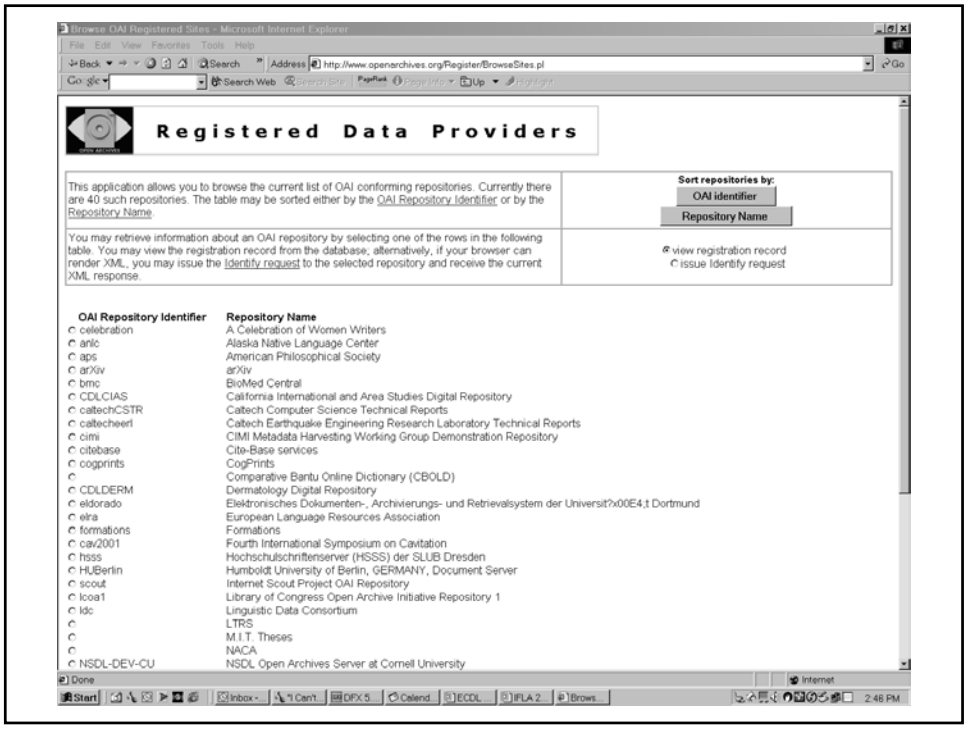


Other OAI Functions

- Registry of data and service providers
- Tool registry
- Community communication

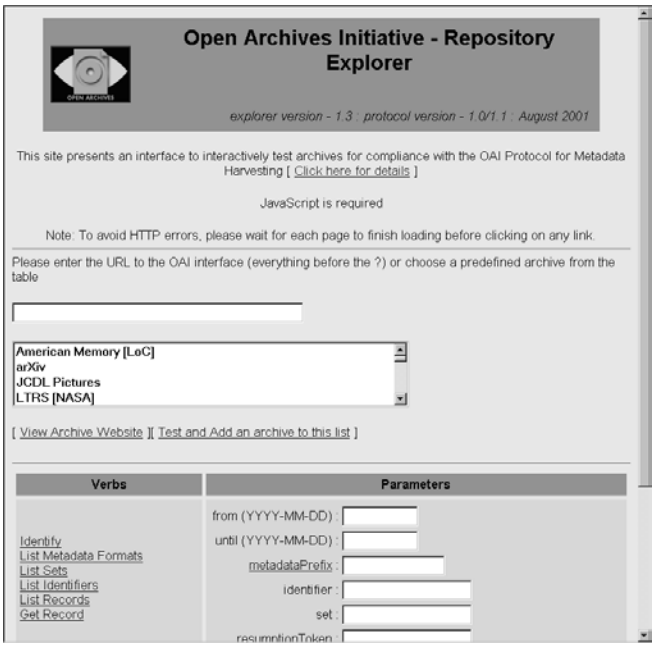



Open Archives Initiative Registration - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://www.openarchives.org/data/registerasprovider.html
Registering as a Data Provider
Data providers who support the OAI protocol may choose to list their repository in the OAI registry. The goals of the registry are:
Provide a publicly accessible list of OAI conformant repositories, making it easy for service providers to discover repositories from which metadata can be harvested.
Provide a mechanism for data providers to ensure their conformance with the OAI protocol specification.
Provide a means for the OAI to monitor use of the protocol and plan future activities and strategies.
This page allows you to register your repository by entering your BASE-URL in the text box at the bottom of this page. Before doing that, please read all of this instruction page so you understand what registration means and the choices you have.
Consequences of Registration
Protocol Testing
Conformance Testing
Robustness Testing
Using OAI Identifiers for Metadata Records
Confirmation of Registration
Changing Registration Information
Consequences of Registration
By registering your repository you agree to the following:
The OAI will immediately run a set of conformance tests on your repository. To ensure integrity of the registry, we will only register those repositories that complete these tests. An email confirming the results of the conformance test will be sent to the address specified in the adminEmail element returned via the Identify protocol request . Your repository will be listed in the registry only after you reply to this email.
The OAI will periodically retest your repository for conformance. In the case that your repository fails to complete the tests, we will remove your repository from the registry and send an email with information to help you return to conformance.
The registry database will store all the information returned via the Identify protocol request:
repositoryName: a human readable name for the repository;
baseURL: the BASE-URL for making protocol requests to the repository;
protocolVersion: the version of the OAI protocol supported by the repository;
adminEmail: the e-mail address of the administrator of the repository.
additional repository-specific description packages.
Contents of the registration database are open for public searching and browsing. There are no plans for restricting access to the registry database.
Protocol Testing
Internet
Start
2:44 PM


Browse OAI Registered Sites - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://www.openarchives.org/Register/BrowseSites.pl
Registered Data Providers
This application allows you to browse the current list of OAI conforming repositories. Currently there are 40 such repositories. The table may be sorted either by the OAI Repository Identifier or by the Repository Name.
Sort repositories by:
OAI identifier
Repository Name
You may retrieve information about an OAI repository by selecting one of the rows in the following table. You may view the registration record from the database; alternatively, if your browser can render XML, you may issue the Identify request to the selected repository and receive the current XML response.
view registration record
issue Identify request
OAI Repository Identifier
Repository Name
celebration
A Celebration of Women Writers
anlc
Alaska Native Language Center
aps
American Philosophical Society
arXiv
arXiv
bmc
BioMed Central
CDLCIAS
California International and Area Studies Digital Repository
caltechCSTR
Caltech Computer Science Technical Reports
caltecheerl
Caltech Earthquake Engineering Research Laboratory Technical Reports
cimi
CIMI Metadata Harvesting Working Group Demonstration Repository
citebase
Cite-Base services
cogprints
CogPrints
Comparative Bantu Online Dictionary (CBOLD)
CDLDERM
Dermatology Digital Repository
eldorado
Elektronisches Dokumenten-, Archivierungs- und Retrievalsystem der Universit?x00E4;t Dortmund
elra
European Language Resources Association
formations
Formations
cav2001
Fourth International Symposium on Cavitation
hsss
Hochschulschriftenserver (HSSS) der SLUB Dresden
HUBerlin
Humboldt University of Berlin, GERMANY, Document Server
scout
Internet Scout Project OAI Repository
lcoa1
Library of Congress Open Archive Initiative Repository 1
ldc
Linguistic Data Consortium
LTRS
M.I.T. Theses
NACA
NSDL-DEV-CU
NSDL Open Archives Server at Cornell University
Done
Internet
Start
2:48 PM

OAI Tools

- Repository Explorer
- Servers and utilities
- Related resources
  - XML
  - Unicode



Implementation Utilities

- Protocol handlers
  - OCLC
  - Virginia Tech
  - UIUC
- Metadata Utilities
  - MARC to DC (OCLC, Virginia Tech, …)
- eprints.org

Participating in the OAI Community

- Listservs
  - oai-general – discussion of OAI related issues
  - oai-implementers – sharing technical questions and agendas
- OAI website (www.openarchives.org)
  - Post news and links to OAI related activities
- Community-specific
  - How does OAI apply to your community?

Externally funded initiatives

- European Community
  - Open Archives Forum
  - Cyclades Project
- Andrew W. Mellon Foundation
  - Funding for 7 service providers
- Digital Library Federation
  - Gateways for access to member's digital collecitons
- National Science Foundation
  - NSDL (www.nsdl.nsf.gov) Core Infrastructure
  - Virginia Tech awards IIS-9986089, 0086227, 0080748 with joint funding by DFG (Germany), CONACyT (Mexico)

Where do we go from here 2001-2002

- Controlling the stampede
- Technical re-evaluation leading to "final" 2.0 specification
  - OAI Technical Committee
- Strategy for standardization
- Community building focused on verification and validation

Open Archives: Communities, Interoperability and Services (Workshop - Sep. 13, 2001 - New Orleans)

- http://purl.org/net/oaisept01
- Session 1: Intro to OAI
- Session 2: Technical Details
- Session 3: Concurrent Group Discussions
  - Applicability of OAI to distributed community building,; community support needed to leverage OAI standards
  - Evaluation of tech stds; current and future directions of stds and services (related to the OAI protocols)
  - See details on next slide
- Session 4: Presentations of Group Findings
- Session 5: Moving Forward

Open Archives: Communities, Interoperability and Services (Workshop - Sep. 13, 2001 - New Orleans)

Building Communities	**Technical Services**
Support for different types of communities	Protocol evaluation: experiences, efficiency, …
Developments aiding community building	Support for internationalization
Selective harvesting (sets)	Services enabled by OAI
Community building ex's	Support for full-text retrieval
Social aspects of OAI-based community projects	Support for protocol adoption

Open Archives: Communities, Interoperability and Services (Workshop - Sep. 13, 2001 - New Orleans)

- Attendees from various institutions

Caltech	U. of Illinois, U-C
CMIS, Carlton, Australia	U. of Oldenburg, GE
Dartmouth College	U. of Southampton
Emory University	U. of Tennessee
Los Alamos Nat'l Lab	US Dept. of Energy
Louisiana State Univ.	Virginia Tech
Michigan State Univ.	
NASA Center for Aerospace Information	

Case Study: NSDL

- National Science, mathematics, engineering and technology education Digital Library (NSF)
- Urgent need: "doors open" Oct. 2002
- Core integration track: building on collections
- Collections track: building portals and centralized repositories (metadata, learning objects / educational resources)
- Metadata: DC, LOMS/IMS
- Problems: will publishers share metadata? Will those with small repositories adopt OAI?

Case Study: NDLTD

- Metadata: MARC21 (coded in XML), ETDMS (see www.ndltd.org/standards)
- Protocols in use: Z39.50, Harvest, Dienst, OAI, as well as http (web sites)
- OCLC's LAF (authority control) to work with RDF implementation of ETDMS
- Union collection -> VTLS's Virtua, Virginia Tech's MARIAN
- Phased efforts for development and testing over more than a year

Case Study: NCSTRL

- CSTR and WATERS -> NCSTRL
  - Federated search of regular sites, harvesting of lite sites
- Changes: disinterest in central service, decline in interest in dept report series, increase in interest in personal web pages (ACM allowance)
- Kepler to support personal Open Archives
- Shift from Dienst-based service to OAI-based service underway in Fall 2001 (aided by Virginia's Internet Technology Innovation Center, through ODU, UVA, and Virginia Tech – along with others)

Case Study: SOLINET

- Mellon Foundation
- SOutheastern LIbrary NETwork (Atlanta)
- Deadline: February 2003
- 10+ univ. collections about American South
- Scholars to learn about OAI, decide how to apply, work toward controlled vocabulary
- Harvesting to central site
- New central DL services (to be developed)

Community Options

- Is DC sufficient, or is there a list of one or more metadata standards existing or that can be developed to suite community needs?
- Is there a natural set structure, or several?
  - Year? Topical areas? Location / institution?
- What are the social, economic, political issues regarding who will run an Open Archive?
- Will all share metadata or must there be federated search as well?

Community Assistance

- Awareness
- Training
- Tools
- Test and validation
- Operation
- Logging and analysis
- Sharing experiences and solutions

Conclusion

- Interoperability
- History / evolution of OAI
- Protocol for metadata harvesting
- Implementations and support
- Current situation / progress
- Community building and support by OAI