www.openarchives.org

Open Archives Initiative

OAI

openarchives@
openarchives.org

"Opening Remarks & Historical Overview" - ACM SIGIR'2001

Ed Fox (w. Lagoze & Suleman)

Acknowledgements

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

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)

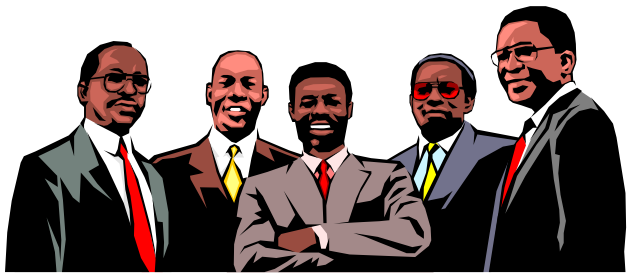
Building Communities	**Technical Services**
Support for different types of communities	Protocol evaluation: experiences, efficiency, …
Developments aiding community building	Support for internationalization
Selective harvesting (sets)	Services enabled by OAI
Community building ex's	Support for full-text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

Ex.: NDLTD Access Possibilities



Web search engines	www.theses.org	www.openarchives.org	library catalog clients	3rd Party Services (e.g., UMI)

Virginia Tech	MIT	National Library of Portugal	CBUC (Spain)	Ohio Link	National Projects: AU, GE, …

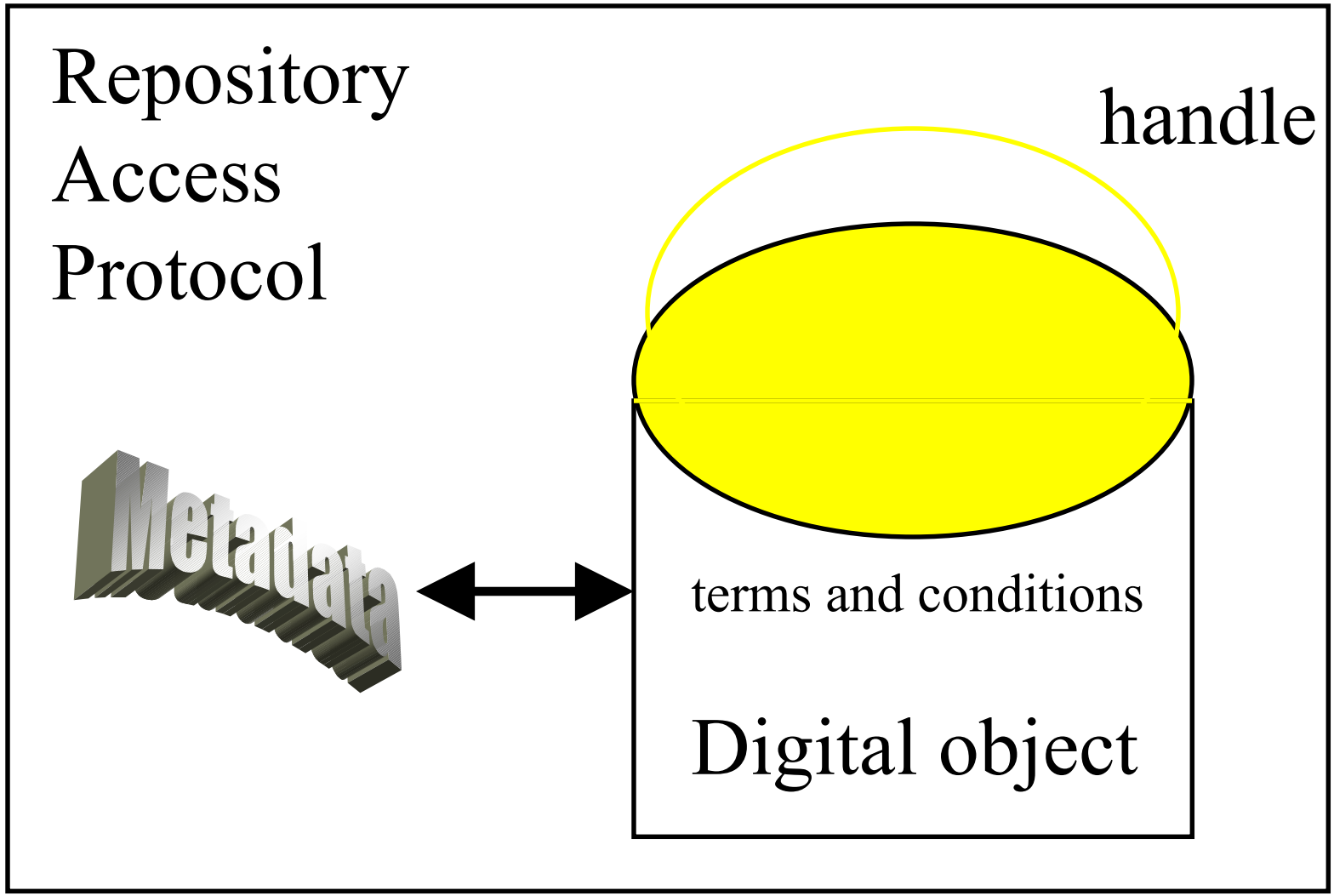
Open Archives Initiative (OAI)

- xxx@LANL, high-energy physics (Ginsparg, 1991)
- CSTR + WATERS = NCSTRL (Lagoze,1994)
- xxx + NCSTRL = CoRR collaboration (1998)
- Universal Preprint Service protoproto, Oct. 21-22, 1999, Santa Fe – led by LANL, CNI, DLF, Mellon --> OAi
- Santa Fe Convention (see Feb. D-Lib Magazine article)
- Follow-on mtgs: 6/3@San Antonio, 9/21@Lisbon (ECDL)
- Archives -> Open Archives
  - Support unique archive identifiers
  - Implement Open Archives metadata set (DC, using XML)
  - Implement OA harvesting protocol (derived from Dienst protocol)
  - Register the archive
- Build tools, layer other services: linking, searching, …

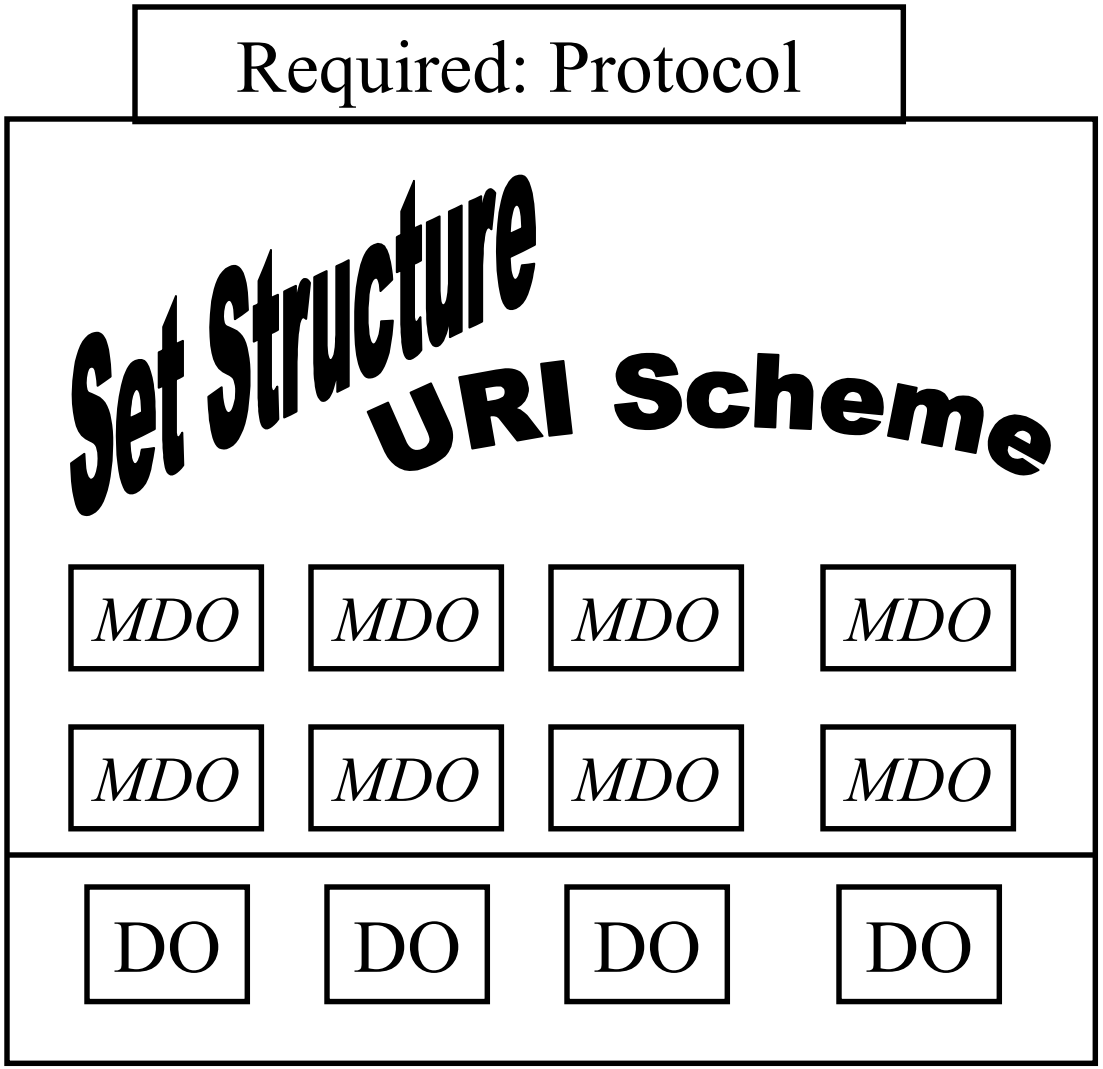
OAi Philosophy

- Self-archiving = submission mechanism
- Long-term storage system = archive
- Open interface = harvesting mechanism
- Data provider + service provider
- Start with "gray literature"
  - e-prints/pre-prints, reports, dissertations, …

Repository of Digital Objects

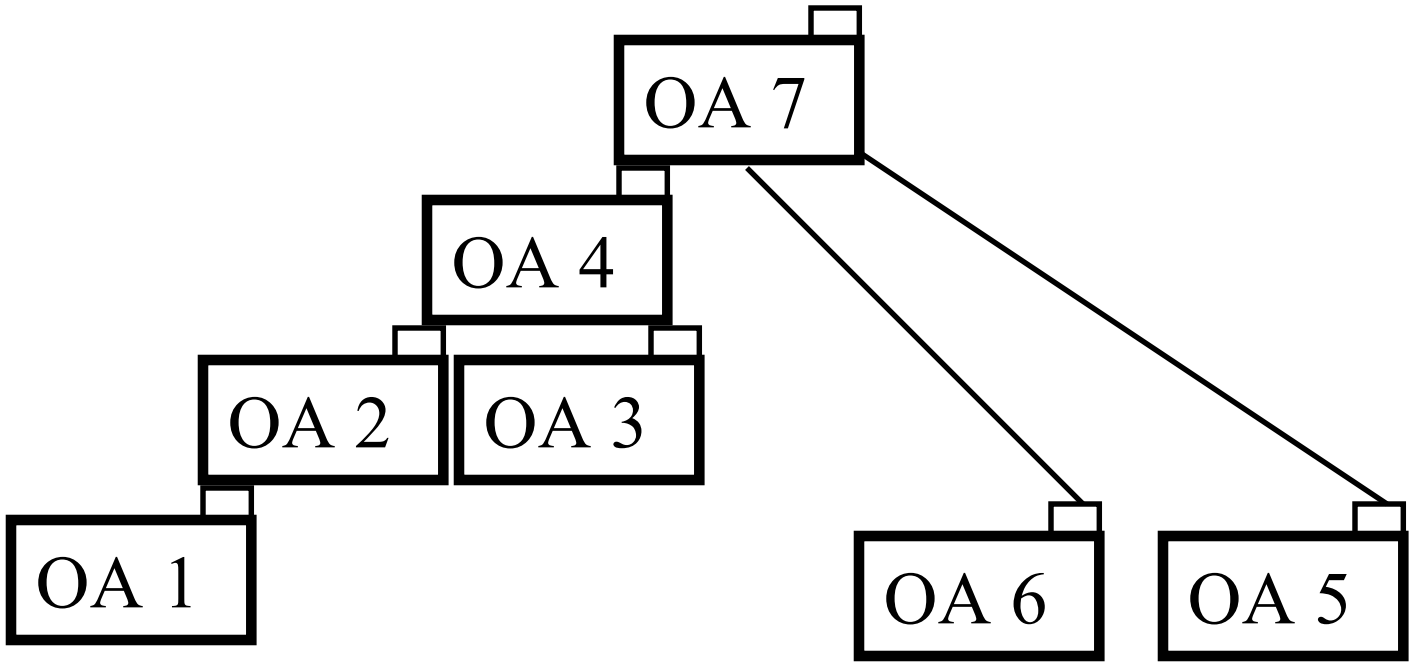


OAI – Repository Perspective

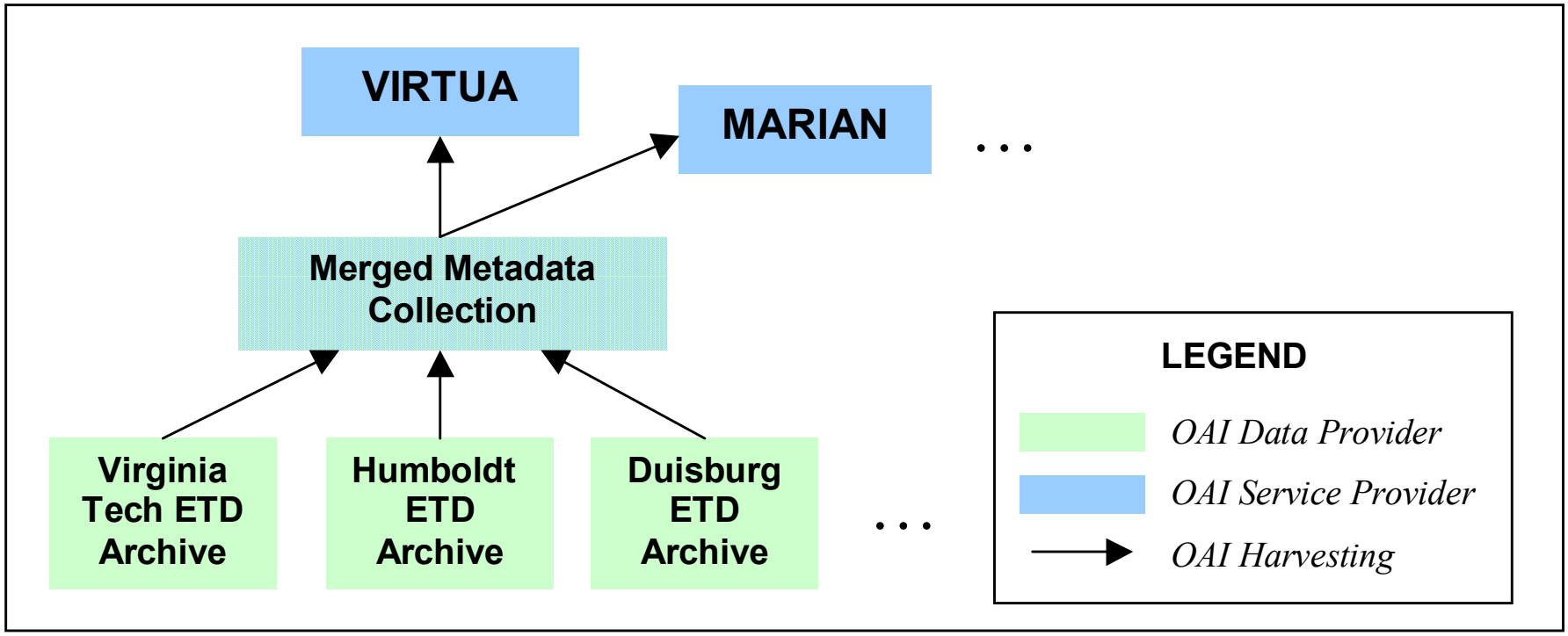


Required: DC

OAI – Black Box Perspective



ETD Union Collection (OAI)



Open Archives (protoproto)

- **ArXiv** & Los Alamos National Lab
- **CogPrints** & U. Southampton
- **NACA** & NASA (reports)
- **NCSTRL** & Cornell U.
- **NDLTD** & Virginia Tech
- **RePEc** & U. Surrey
- Total of around 200K records

Original Open Archives Members

- American Physical Society
- California Digital Library
- Caltech
- Coalition for Networked Info.
- Cornell University
- Harvard University
- Library of Congress
- Los Alamos Nat'l Lab
- Mellon Foundation
- NASA Langley Research Cntr
- Old Dominion University
- Stanford University
- U. of Ghent
- U. of Surrey
- U. of Southampton
- Vanderbilt University
- Virginia Tech
- Washington University

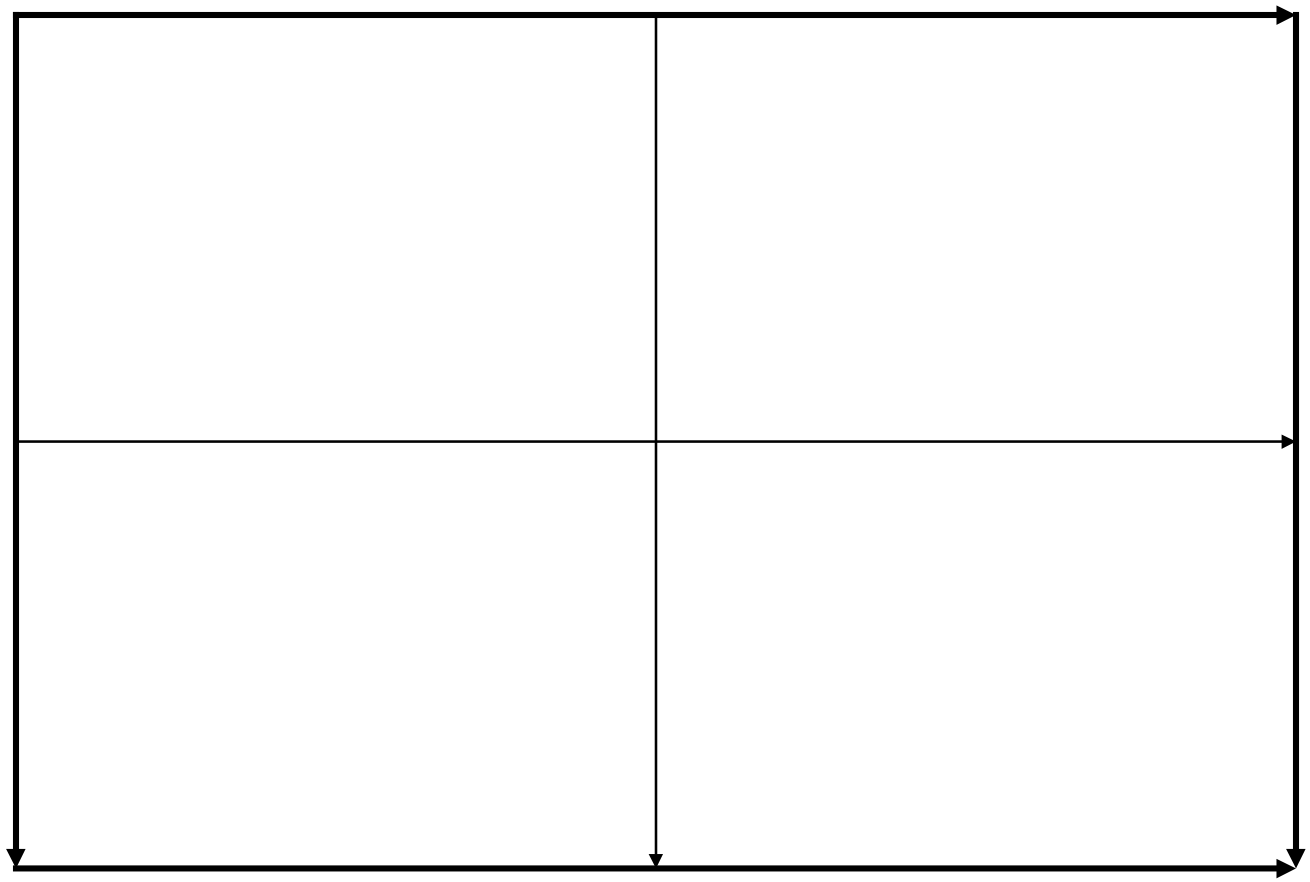
Open Archives Future

- EconWPA (U. Washington)
- e-biomed -> PubMed Central (NIH)
- PubScience (DOE)
- Clinical Medicine Netprints (+ other HighWire Press holdings)
- University ePub (California Digital Library)
- All public e-prints (MIT)
- Scholar's Forum (Caltech)
- Int'l: CERN, Germany, India, Mexico, …
- **Goal: millions of books/articles/reports / yr**

Approaches to Open Archives

Build By Institution

Build By Discipline



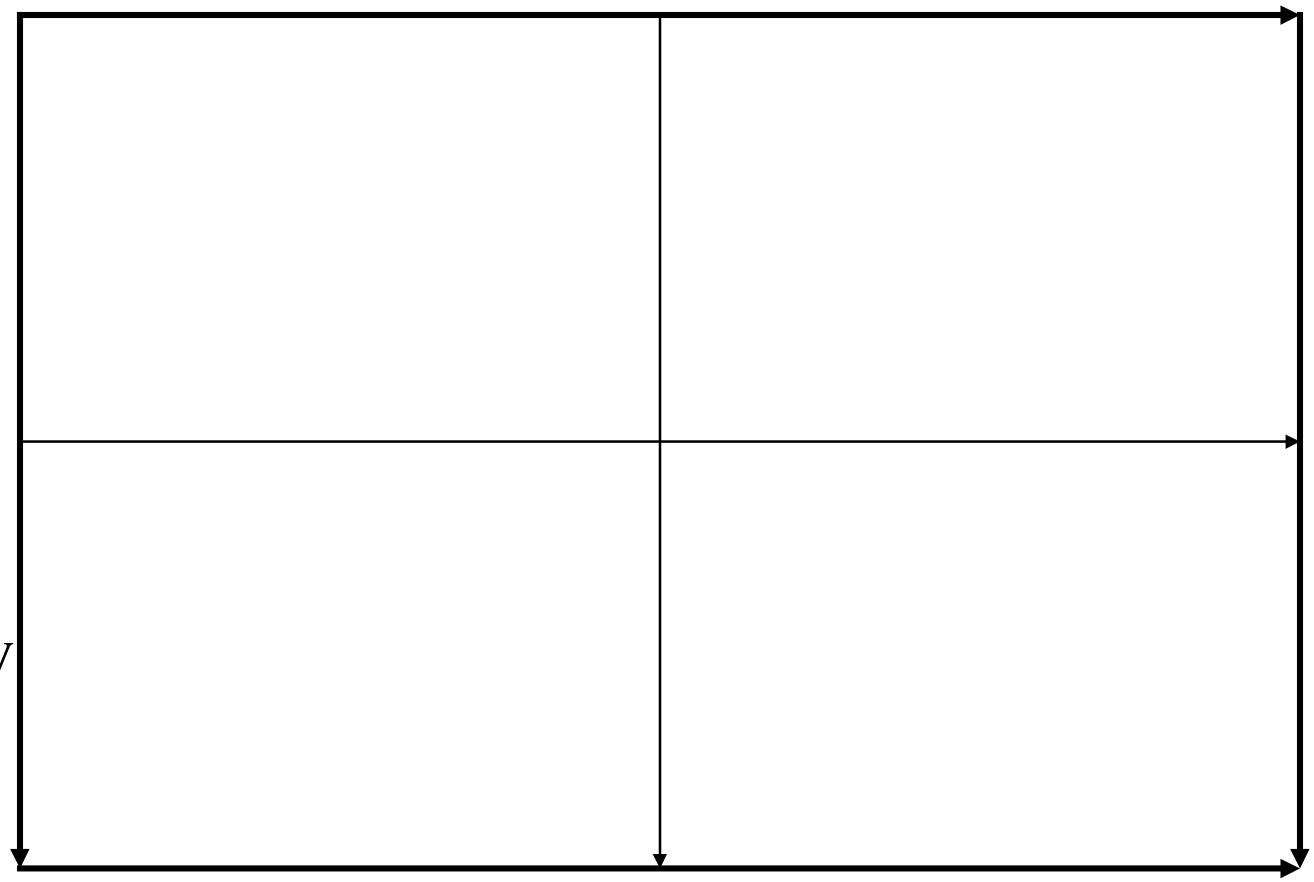
Approaches to Open Archives

Build By Institution

Build By Discipline

Access by

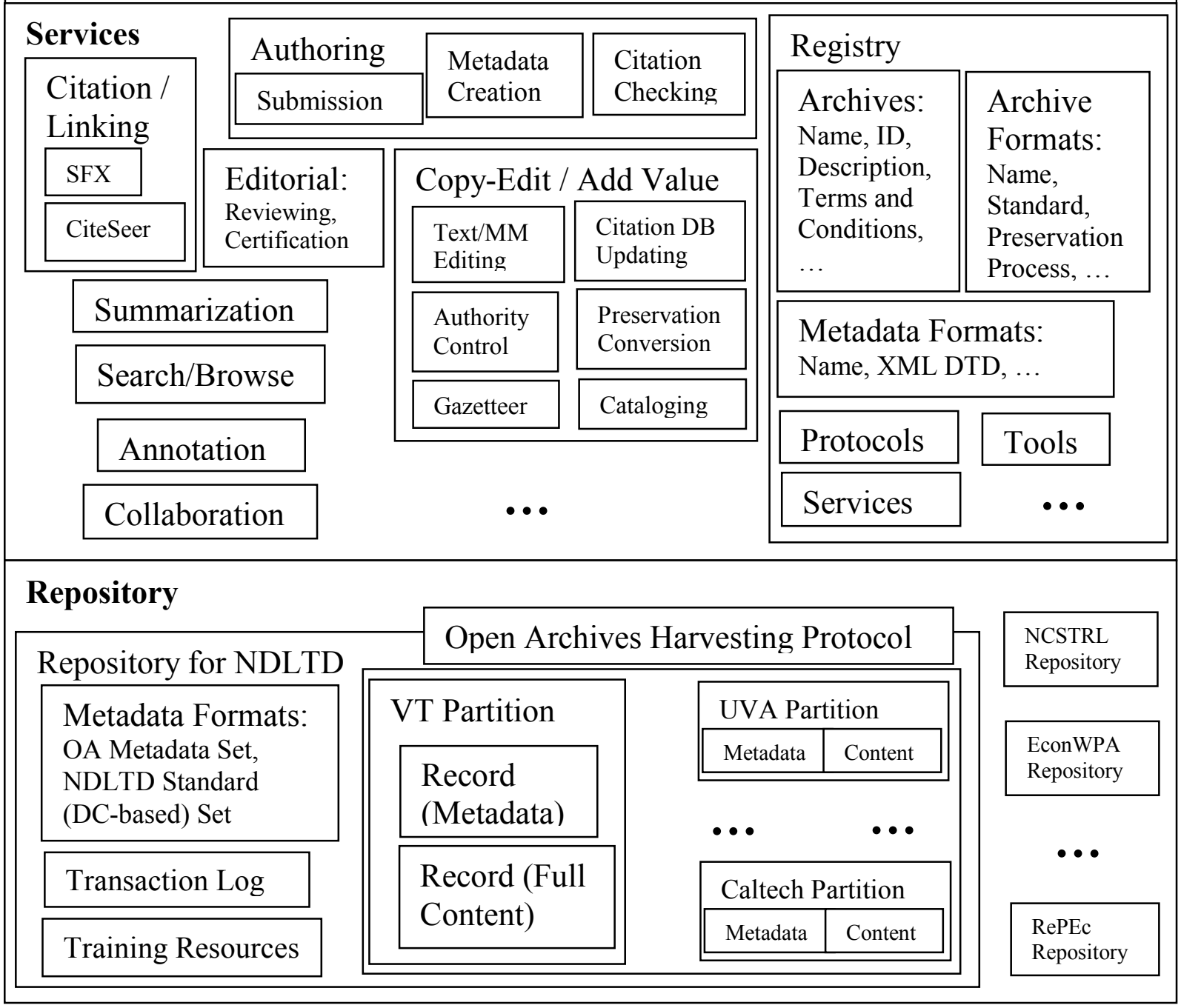
Author
Category
Interdisciplinary
Year
Language
Query …



Mechanisms

- **Sharing**
  - **Join federation, run software**
  - **Make metadata and archive available**
- **Aggregating**
  - **By discipline**
  - **By institution**
  - **By genre**
- **Automating**
  - **Workflow**
  - **Harvesting and providing services**
  - **Federated searching**
  - **Dynamic linking (e.g., with SFX (OpenURLs))**

Figure 1. Layers Related to Open Archives Initiative



VT View of the Open Archives Initiative (OAI)

- Enable sharing of publication metadata and full-text by digital libraries
- Standardize low-level mechanisms to share contents of libraries
- Build higher-level user-centric and administrative services in meta-libraries
- Install organizational mechanisms to support the technical processes

Virginia Tech Projects

- MARC XML-DTD
- Computer Science Teaching Centre (CSTC)
- W3C Web Characterization Repository
- OAI Repository Explorer
- Networked Digital Library of Theses and Dissertations (NDLTD)

MARC XML-DTD

- XML Transport format for US-MARC records
- Standardized metadata exchange format for traditional library services joining OAI

OAI Repository Explorer

- Serves as a compliancy test
- Allows browsing of open archives using only OAI protocol
- Sends requests on behalf of user, parses and checks responses and displays browsable interface
- Will detect most discrepancies in protocol
- http://purl.org/net/explorer

Request, Response – OAI, VT ETDs

Request

http://scholar.lib.vt.edu/theses/OAI/cgi-bin/index.pl?
verb=GetRecord&metadataPrefix=oai_etdms&identifier=oai:VTETD:etd-520112859651791

Response

```
<?xml version="1.0" encoding="UTF-8" ?>
- <GetRecord xmlns="http://www.openarchives.org/OAI/1.1/OAI_GetRecord"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.openarchives.org/OAI/1.1/OAI_GetRecord
    http://www.openarchives.org/OAI/1.1/OAI_GetRecord.xsd">
    <responseDate>2001-08-18T18:03:28-05:00</responseDate>
    <requestURL>http://scholar.lib.vt.edu:80/theses/OAI/cgi-bin/index.pl?
      verb=GetRecord&metadataPrefix=oai_etdms&identifier=oai:VTETD:etd-
      520112859651791</requestURL>
  - <record>
    - <header>
        <identifier>oai:VTETD:etd-520112859651791</identifier>
        <datestamp>1996-06-05</datestamp>
      </header>
    - <metadata>
      - <thesis xmlns="http://www.ndltd.org/standards/metadata/etdms/1.0/"
          xsi:schemaLocation="http://www.ndltd.org/standards/metadata/etdms/1.0/
          http://www.ndltd.org/standards/metadata/etdms/1.0/etdms.xsd">
          <title>Analysis of Tow-Placed, Variable-Stiffness Laminates</title>
          <creator>Waldhart, Chris</creator>
          <subject>variable-stiffness laminates</subject>
          <subject>curvilinear fibers</subject>
          <subject>tow placement machine</subject>
          <subject>buckling</subject>
          <description>It is possible to create laminae that have spatially varying fiber
            orientation with a tow placement machine. A laminate which is composed of
            such plies will have stiffness properties which vary as a function of position.
            Previous work had modelled such variable-stiffness laminae by taking a
```

Motivation

- Existence of some established but independent archives
- Need for cross-archive services (like search engines)
- Lack of low-cost interoperability technology
- Experience from past projects such as Dienst

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

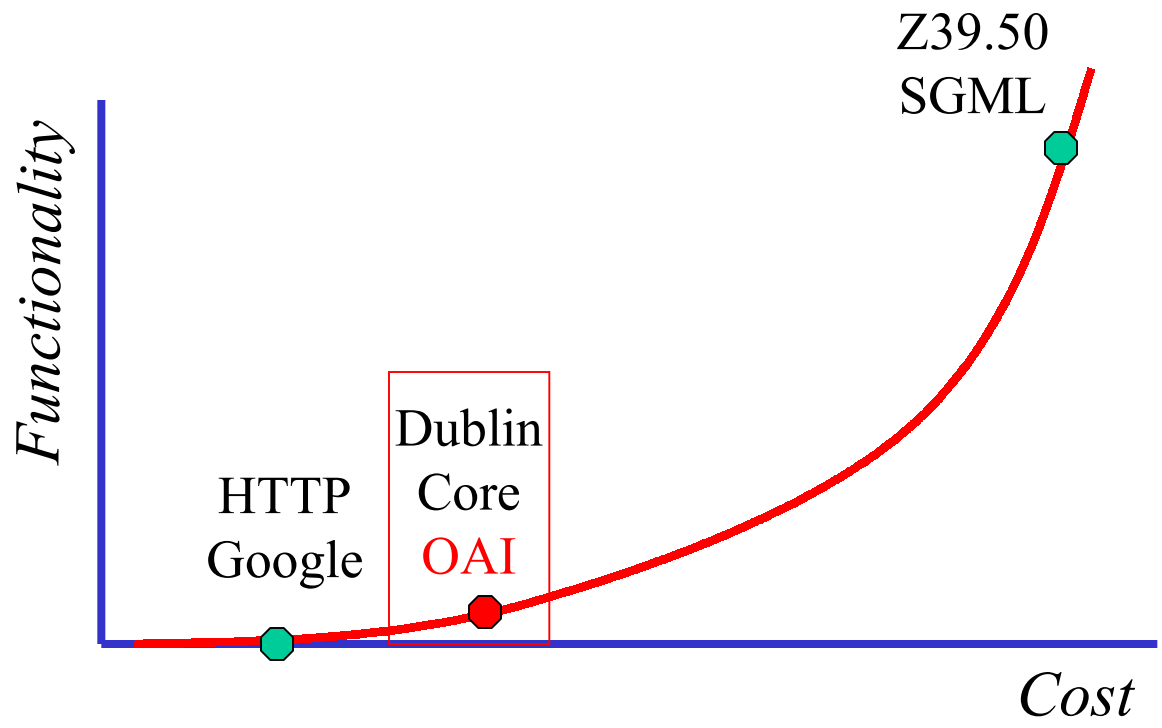
Digital Library Interoperability

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

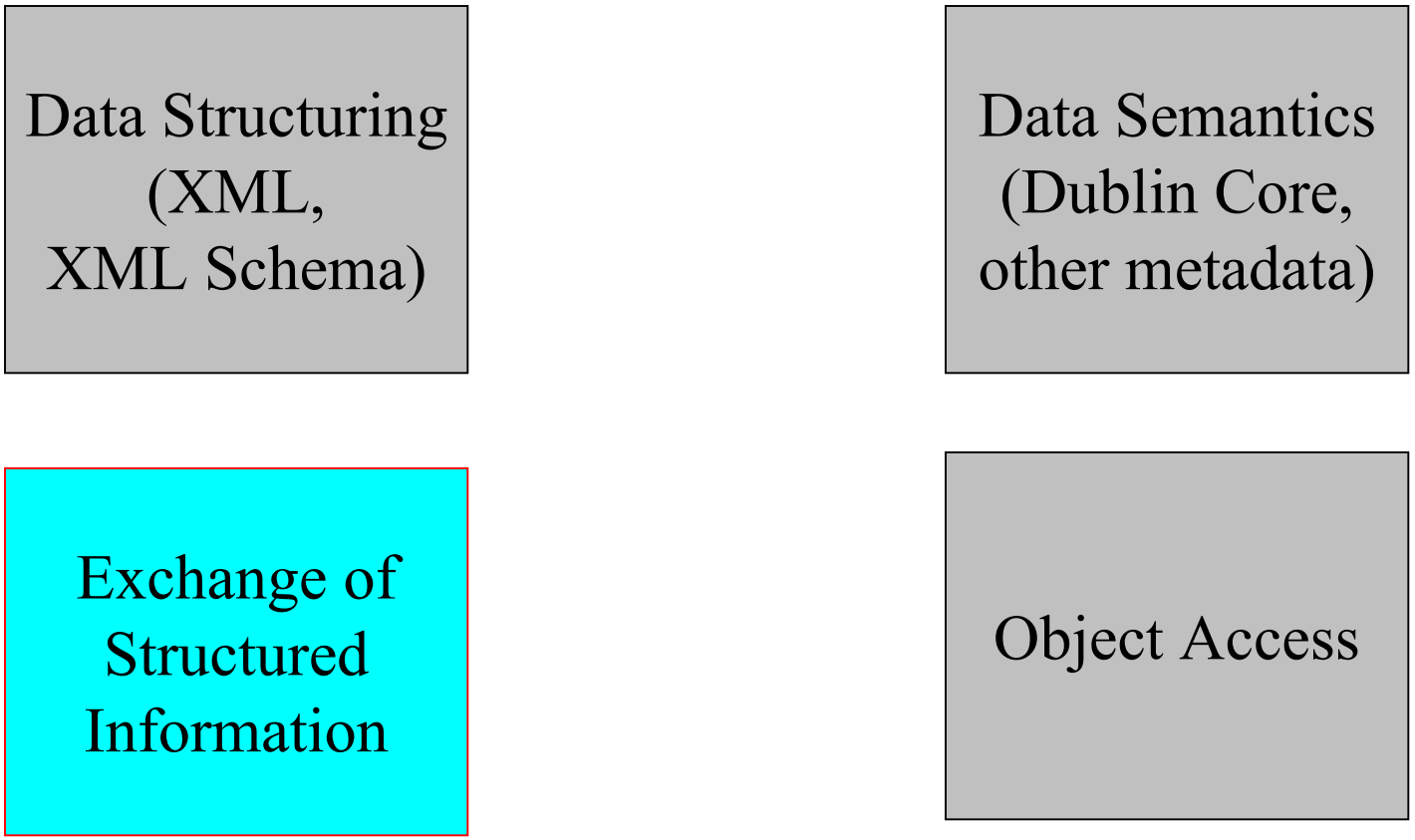
A Short History of Interoperability

- 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

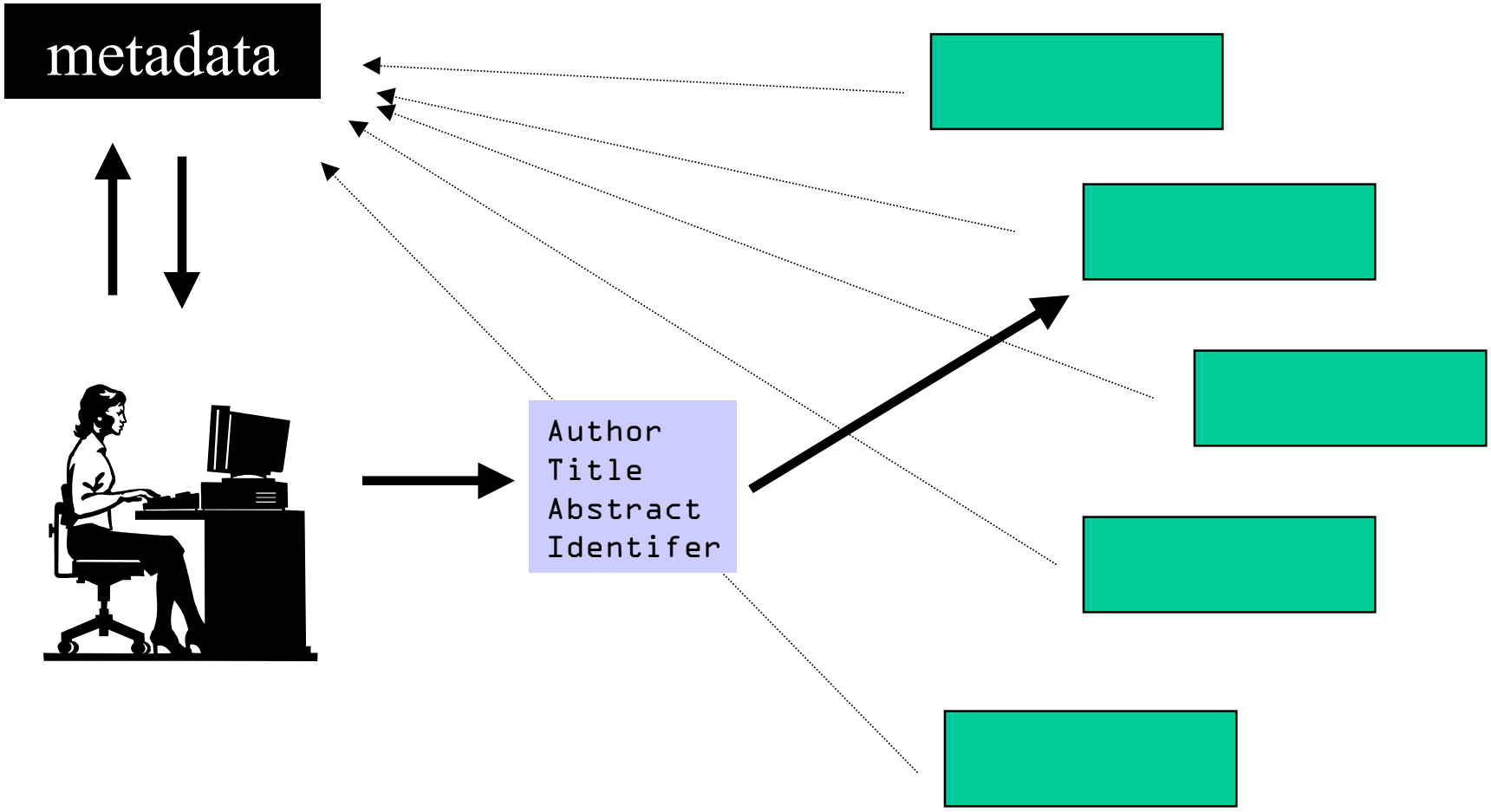
Interoperability Trade-offs



OAI's Location in a Broader Interoperability Fabric



Yes, it's about resource discovery over distributed collections



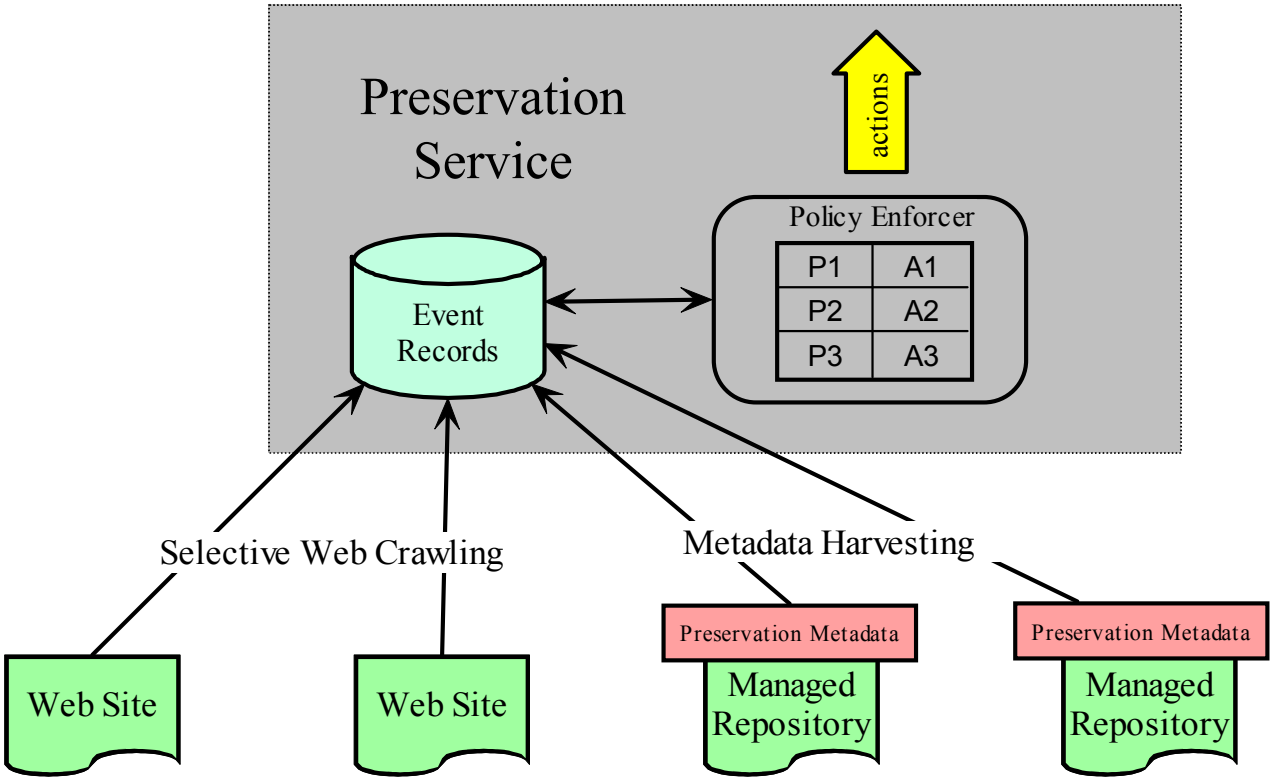
Beyond resource discovery to distributed custodianship

- 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.

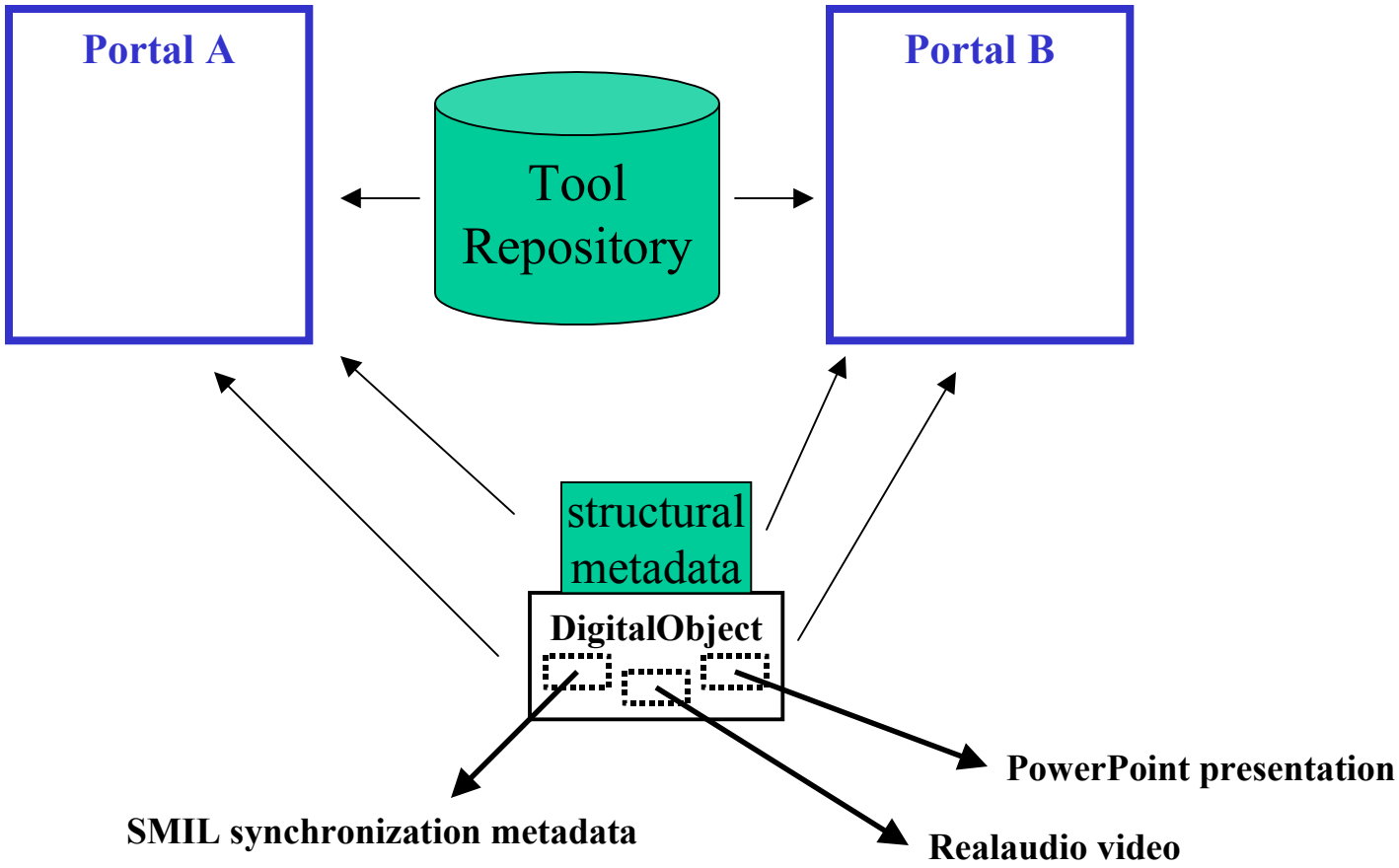
Facilitating/Monitoring Longevity of Distributed Content



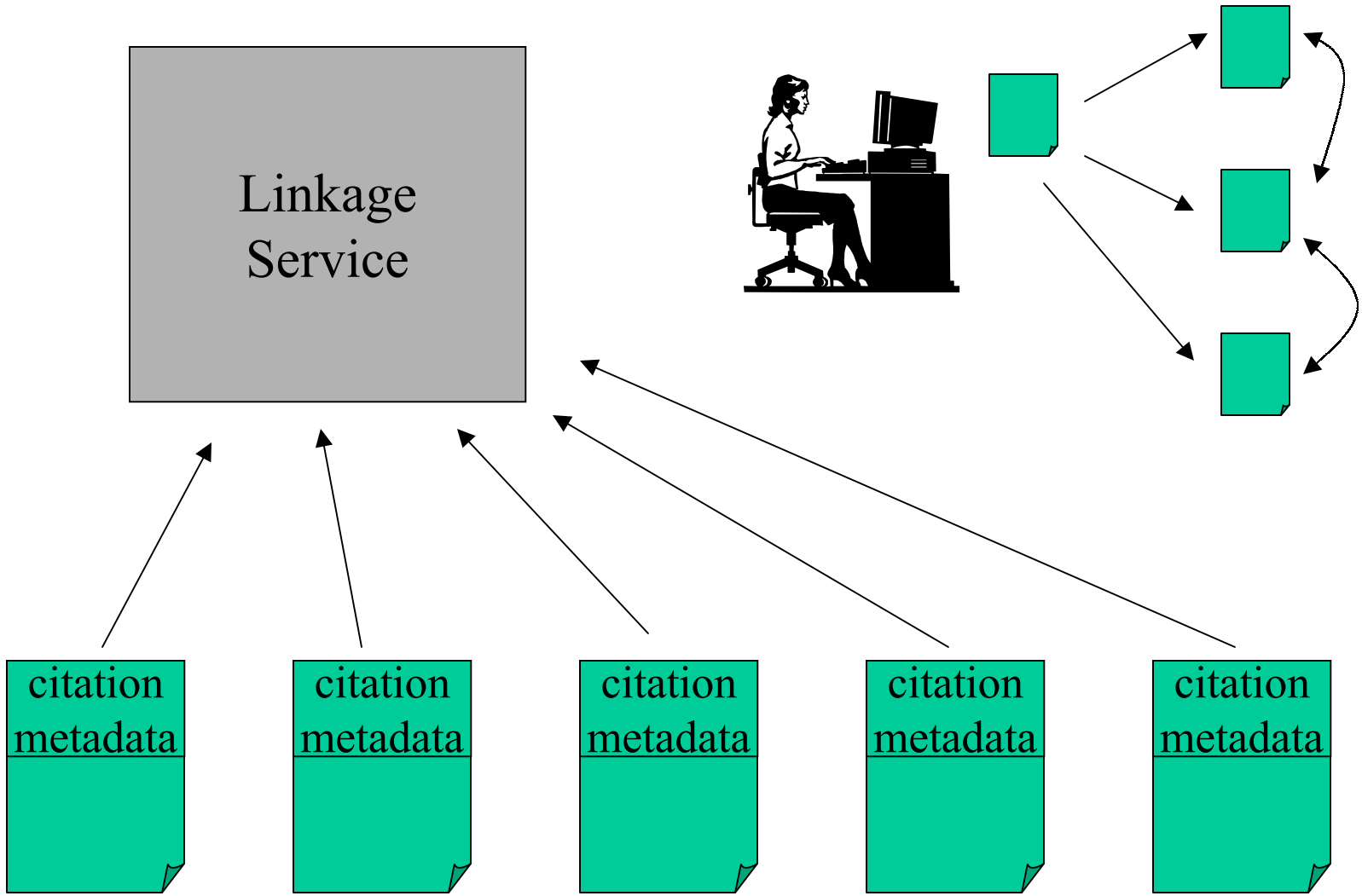
Personalization of Content

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

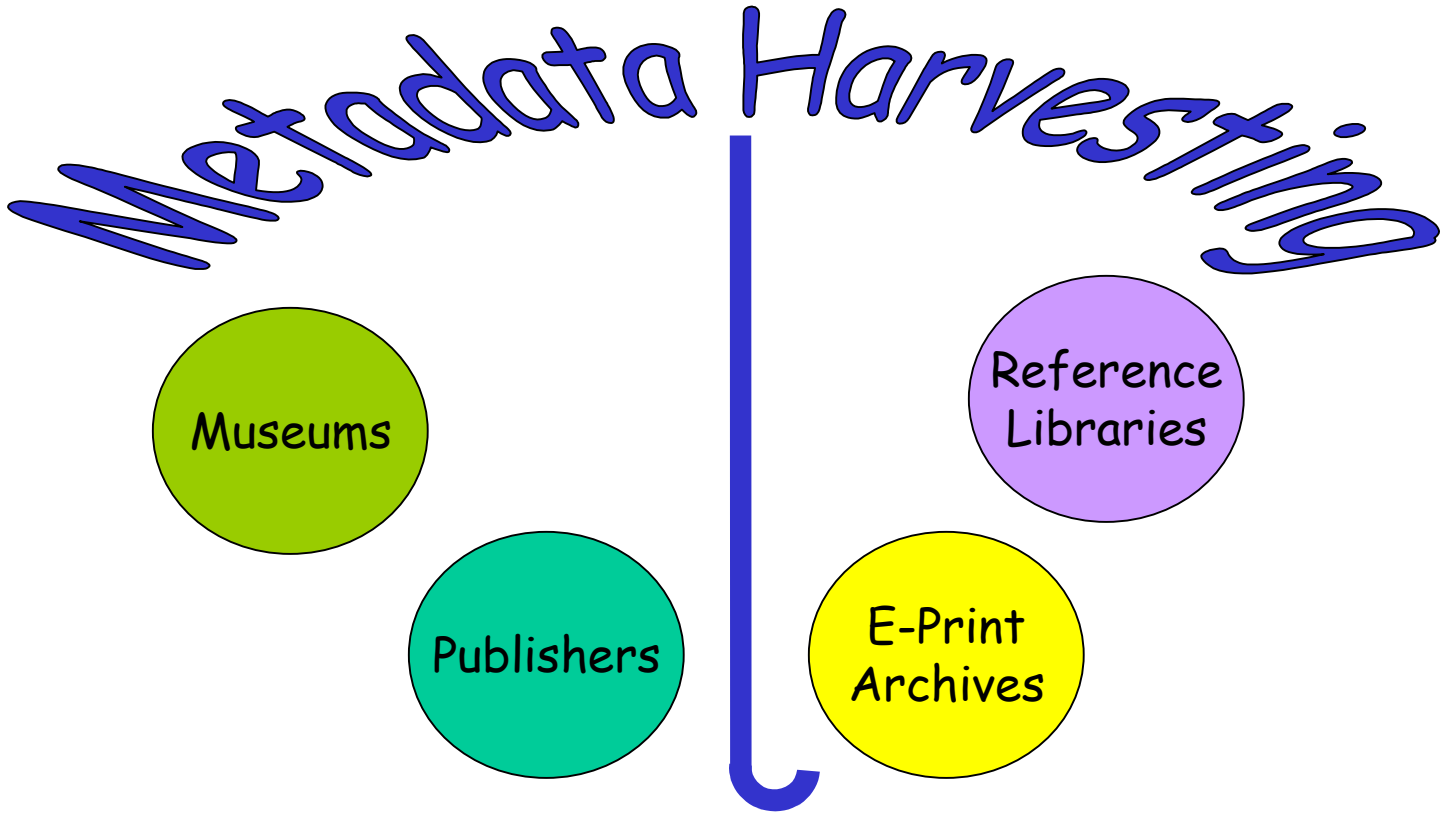
"Open" Archives

- 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

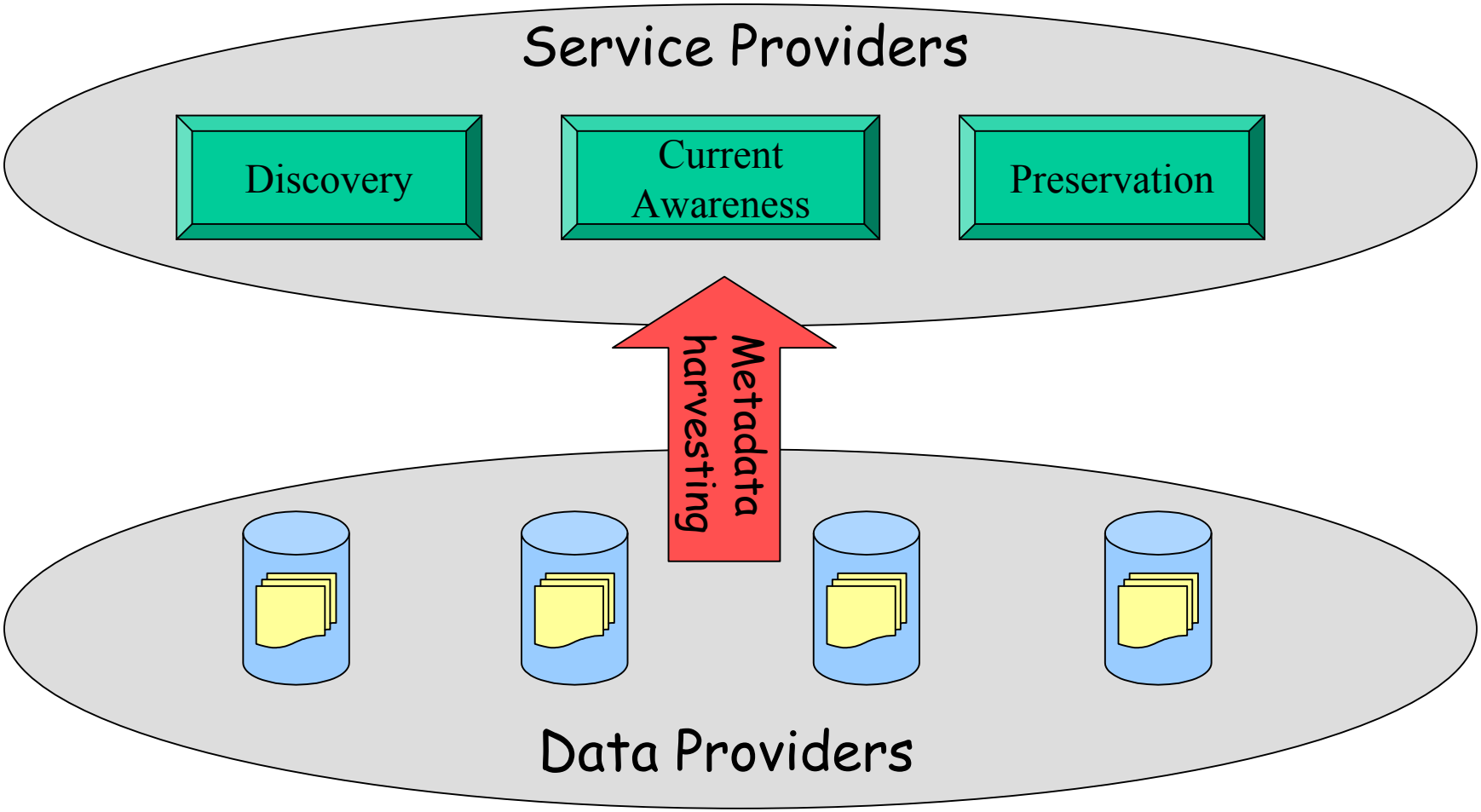
OAI Organizational Structure
Key Features

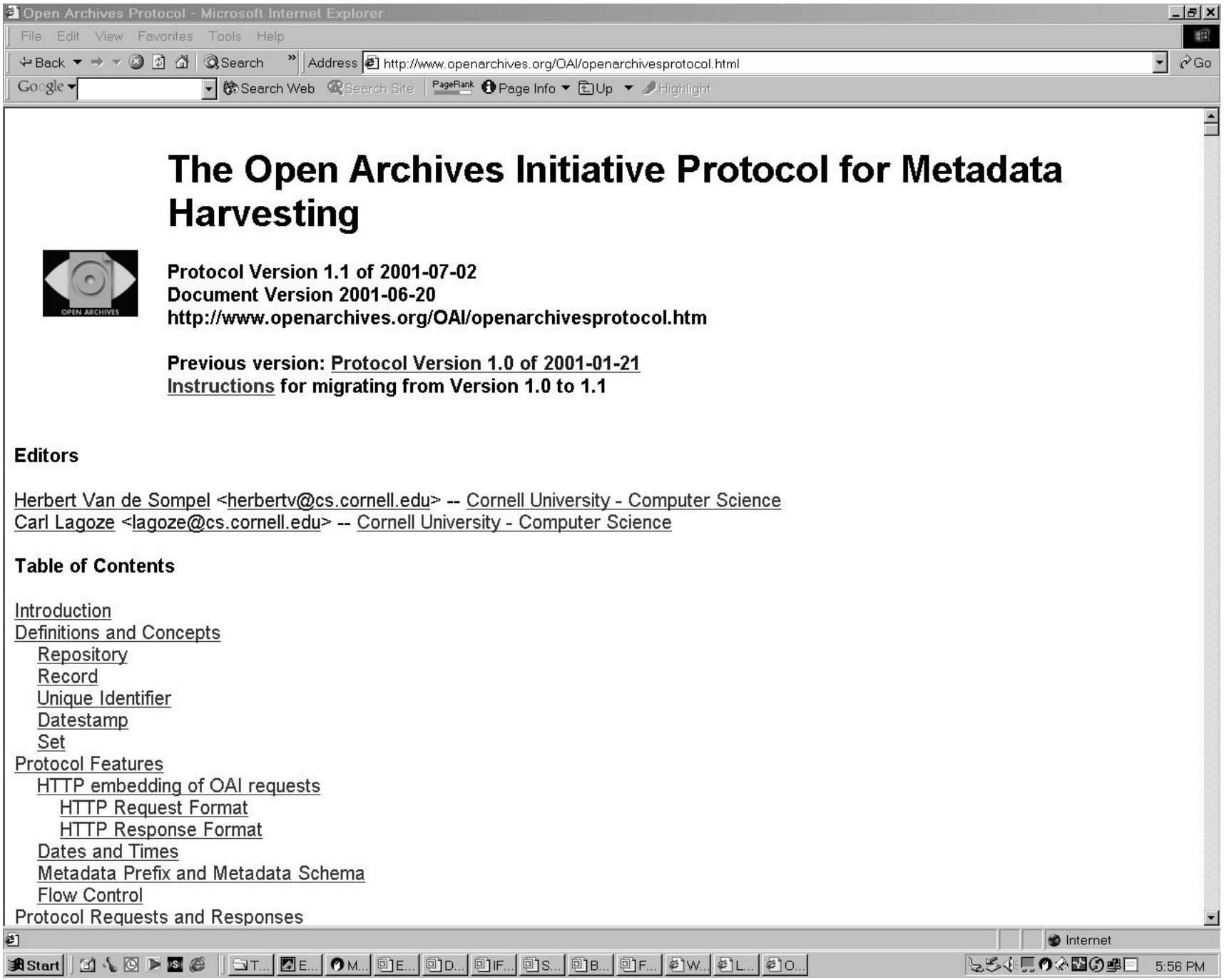
- Clear focus and scope
  - Developing and refining technical specification
  - Community building and evangelism limited to serving that goal and to encouraging widespread adoption
- Encouraging specialization and community-specific activities
- Division of responsibility
  - Executive (Van de Sompel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




Open Archives Protocol - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Address http://www.openarchives.org/OAI/openarchivesprotocol.html Go
Google Search Web Search Site PageRank Page Info Up Highlight
The Open Archives Initiative Protocol for Metadata Harvesting
OPEN ARCHIVES
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
Introduction
Definitions and Concepts
Repository
Record
Unique Identifier
Datestamp
Set
Protocol Features
HTTP embedding of OAI requests
HTTP Request Format
HTTP Response Format
Dates and Times
Metadata Prefix and Metadata Schema
Flow Control
Protocol Requests and Responses
Internet
Start
5:56 PM

What is the OAI-MHP ?

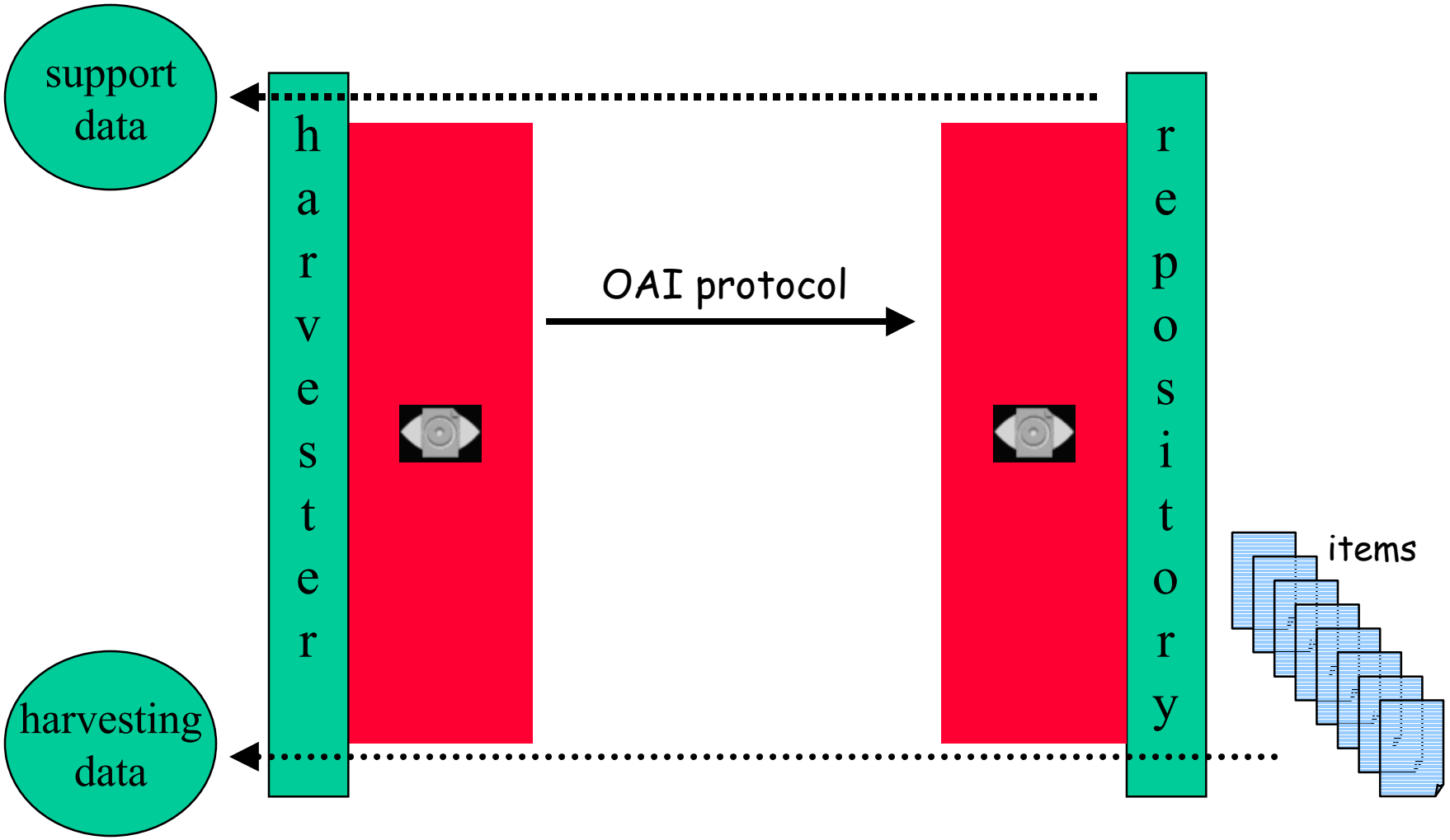
- What is the Metadata Harvesting Protocol?
  - Protocol to transfer metadata from a source archive to a destination archive

    - Any metadata
    - In a continuous stream
    - As simply as possible

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


support data
harvester
OAI protocol
repository
items
harvesting data

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

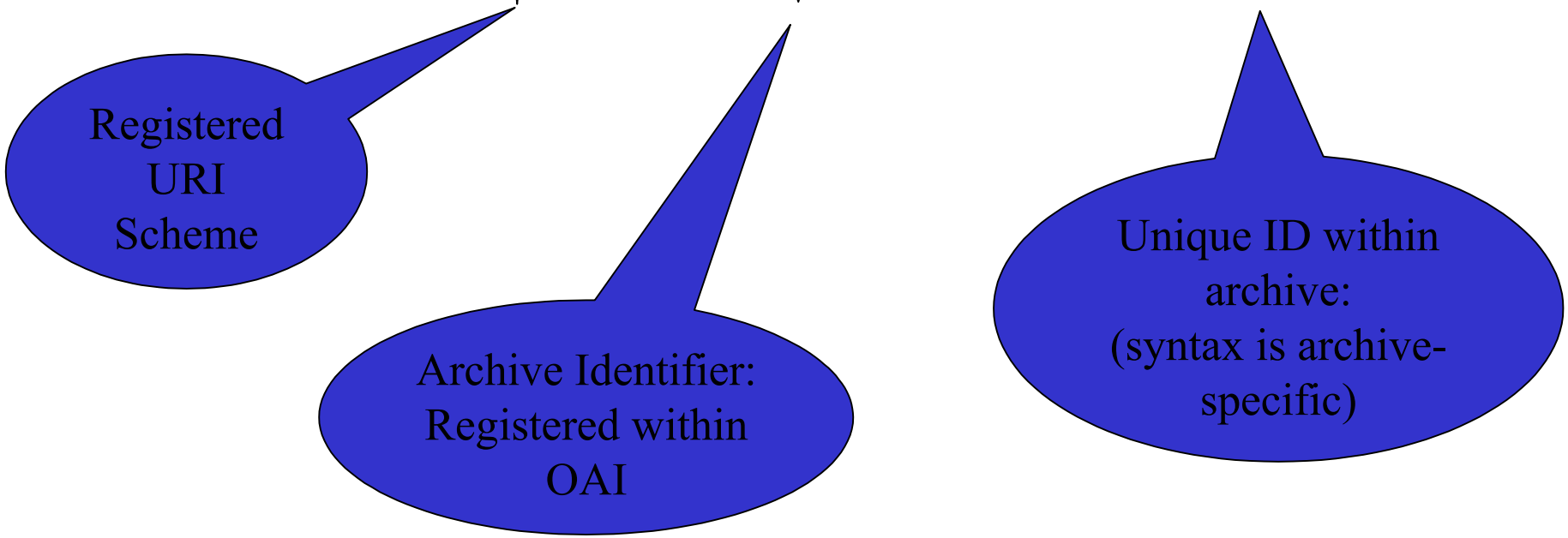
community-specific record data

identifiers

locally unique key for extracting a record from a repository

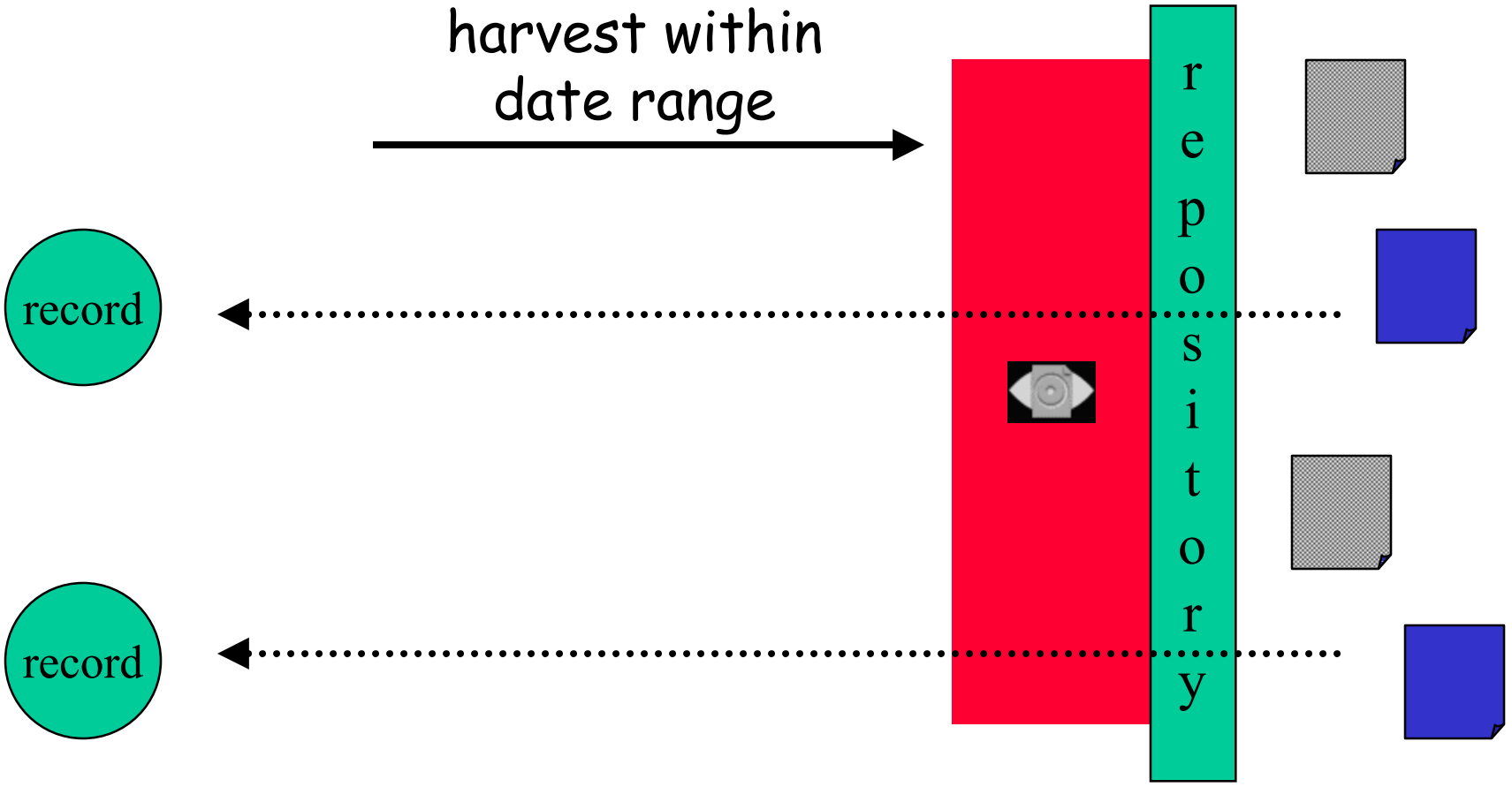
---

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

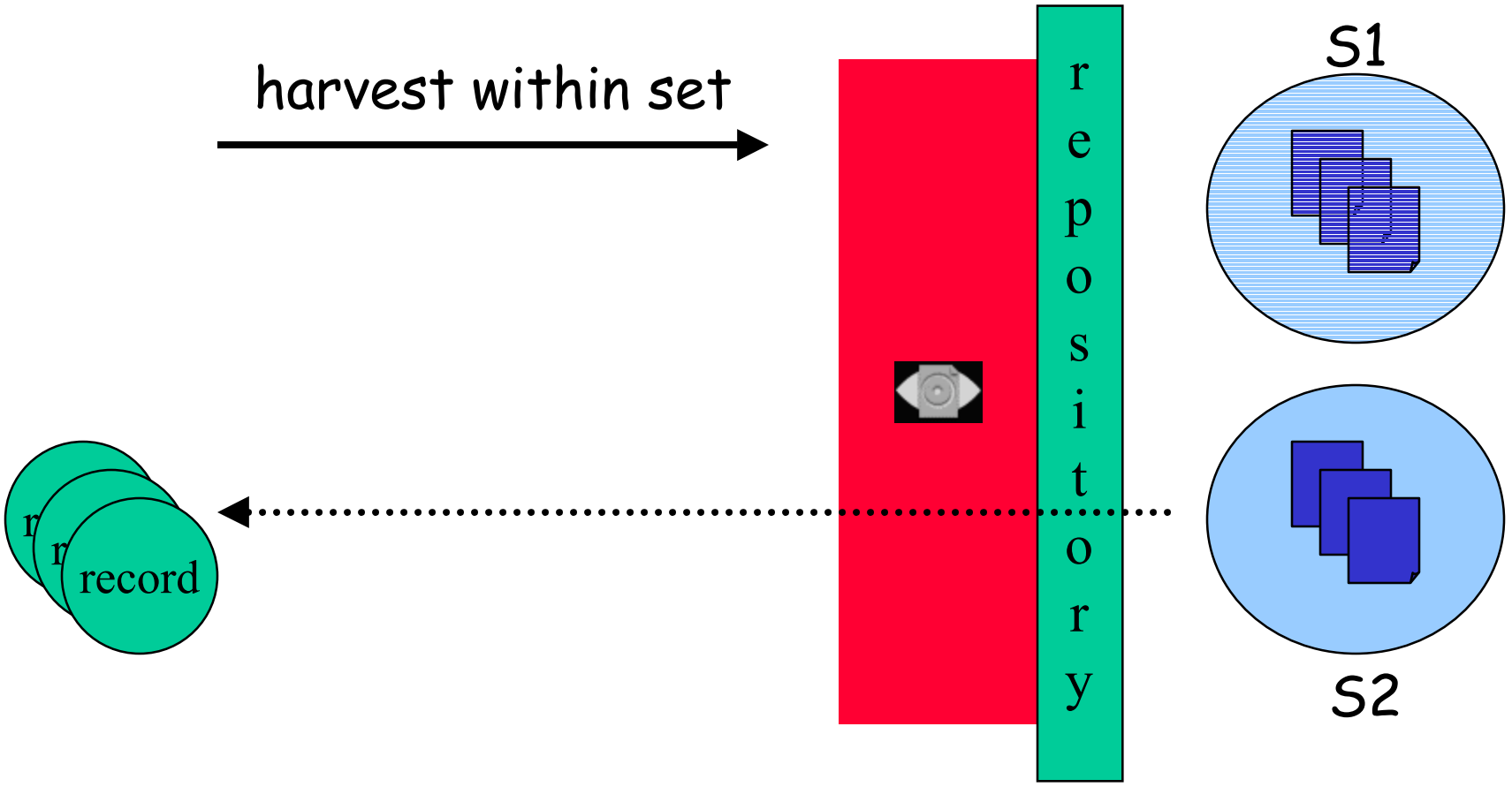


example = oai:ncstrl:ncstrl.cornellcs/TR94-1418

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;identifier=oai%3AarXiv%3A0001
            &amp;metadataPrefix=oai_dc</requestURL>
   <record>
      record contents
   </record
      additional records
</GetRecord>
```

xml namespaces

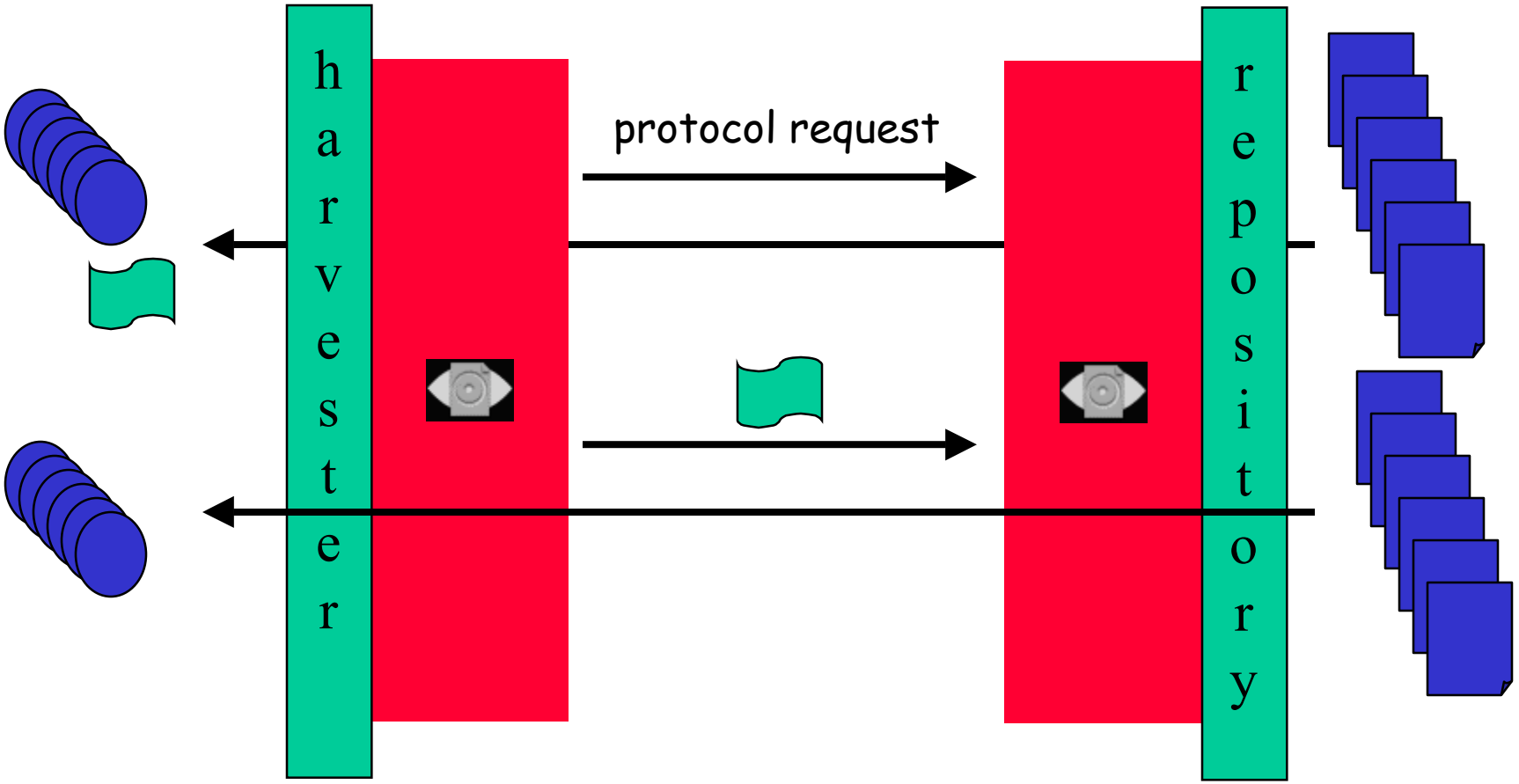
response header

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



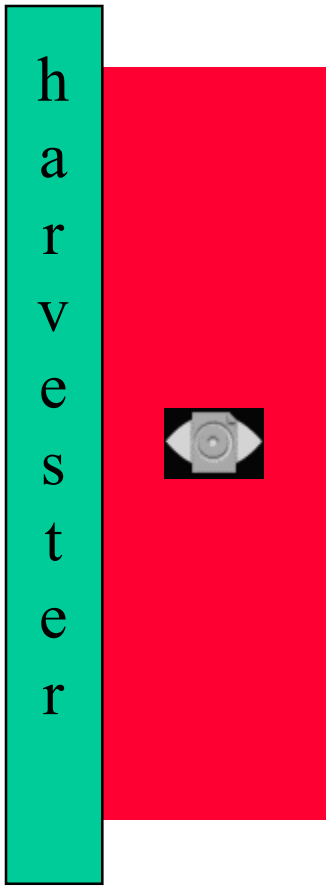
flow control specifics

- 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

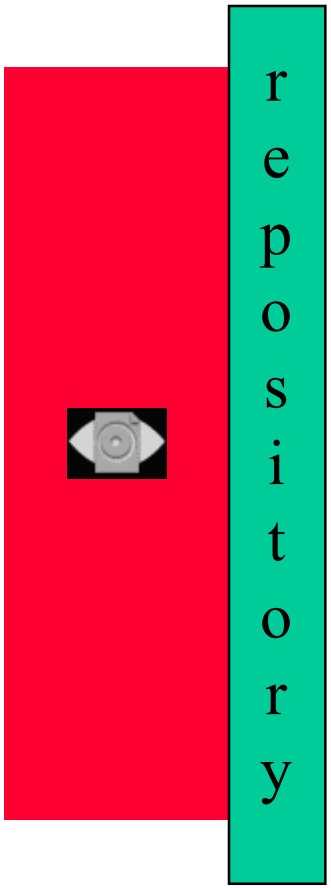


Supporting protocol requests:

- Identify
- ListMetadataFormats
- ListSets

Harvesting protocol requests:

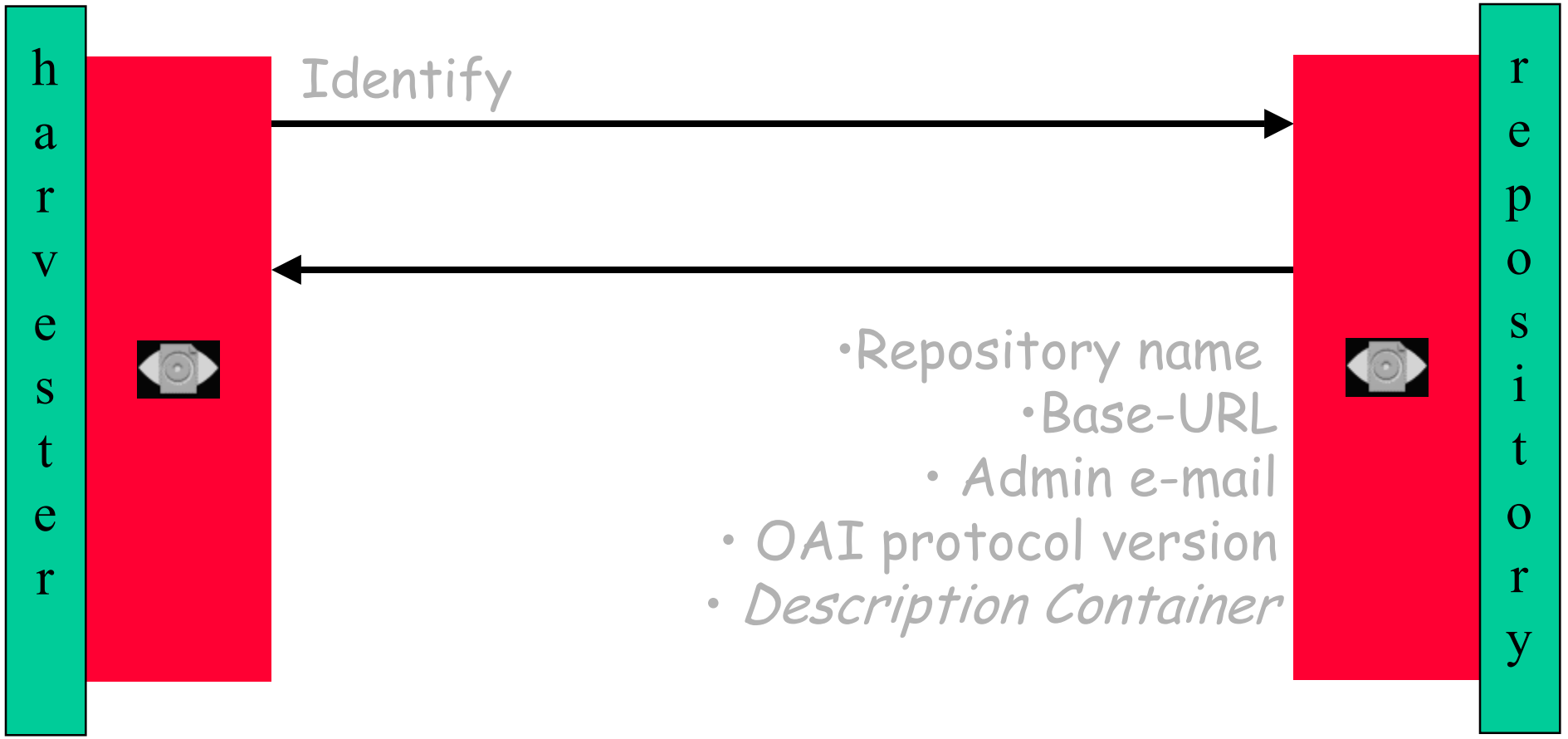
- ListRecords
- ListIdentifiers
- GetRecord



Supporting Protocol Requests

service provider

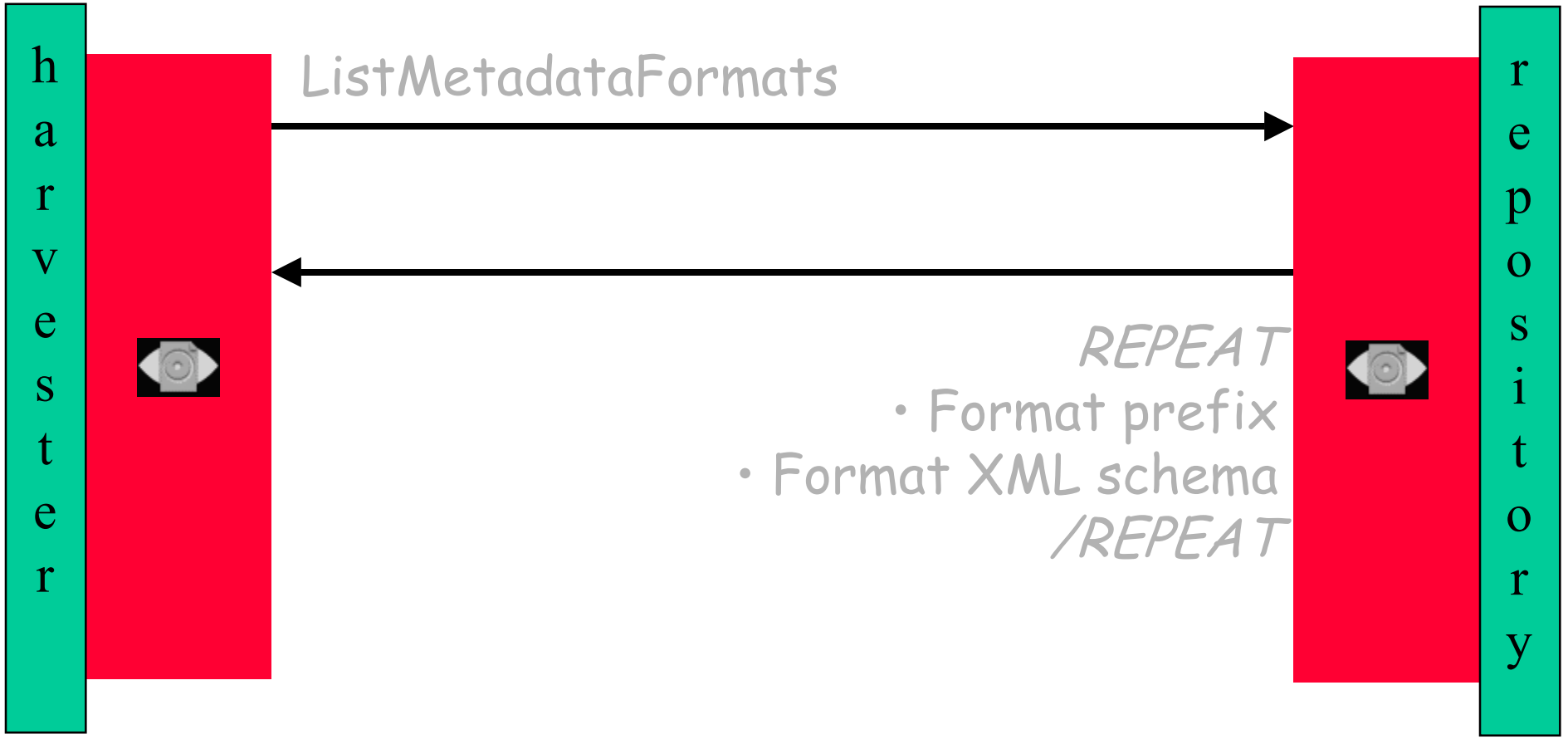
data provider



Supporting Protocol Requests

service provider

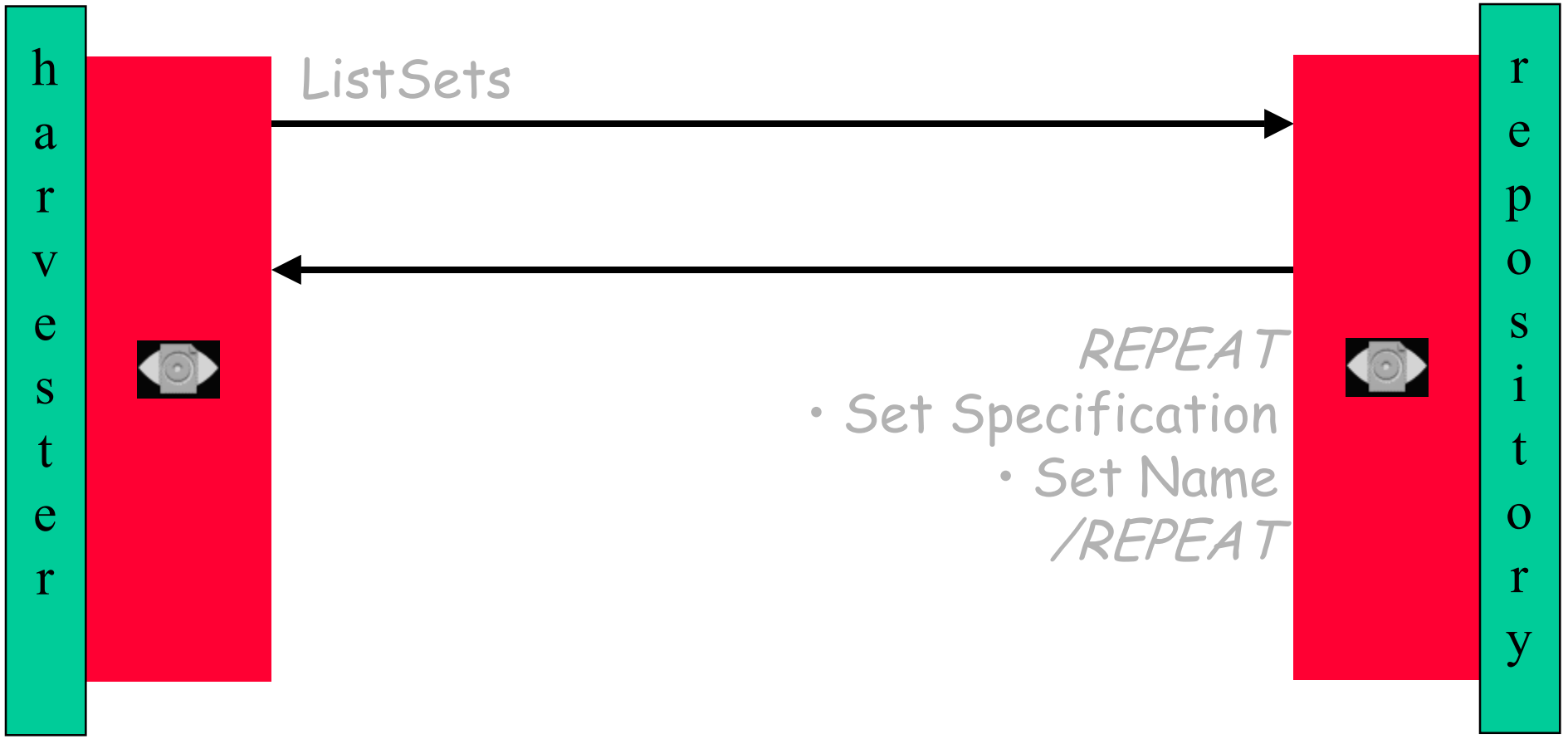
data provider



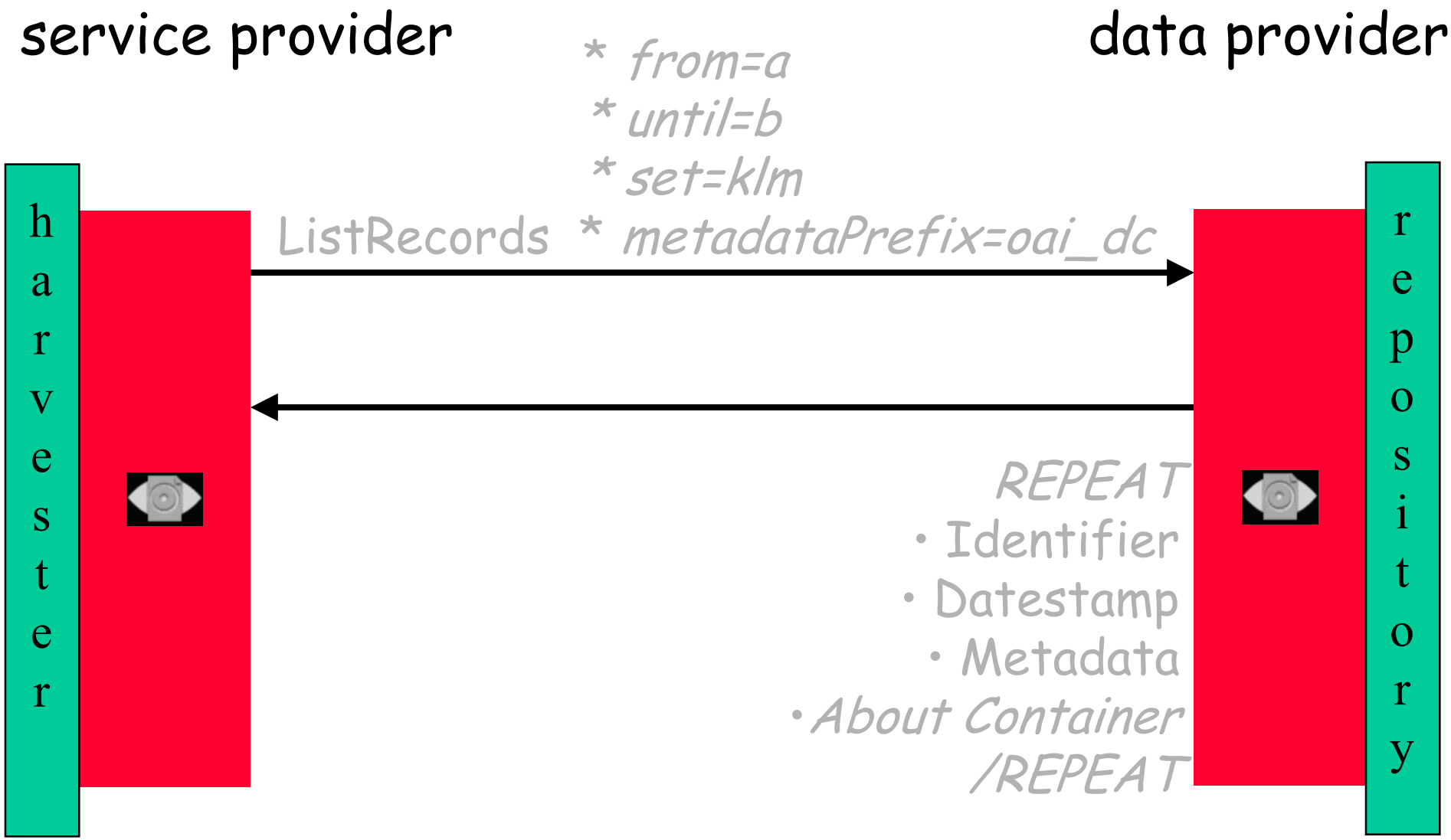
Supporting Protocol Requests

service provider

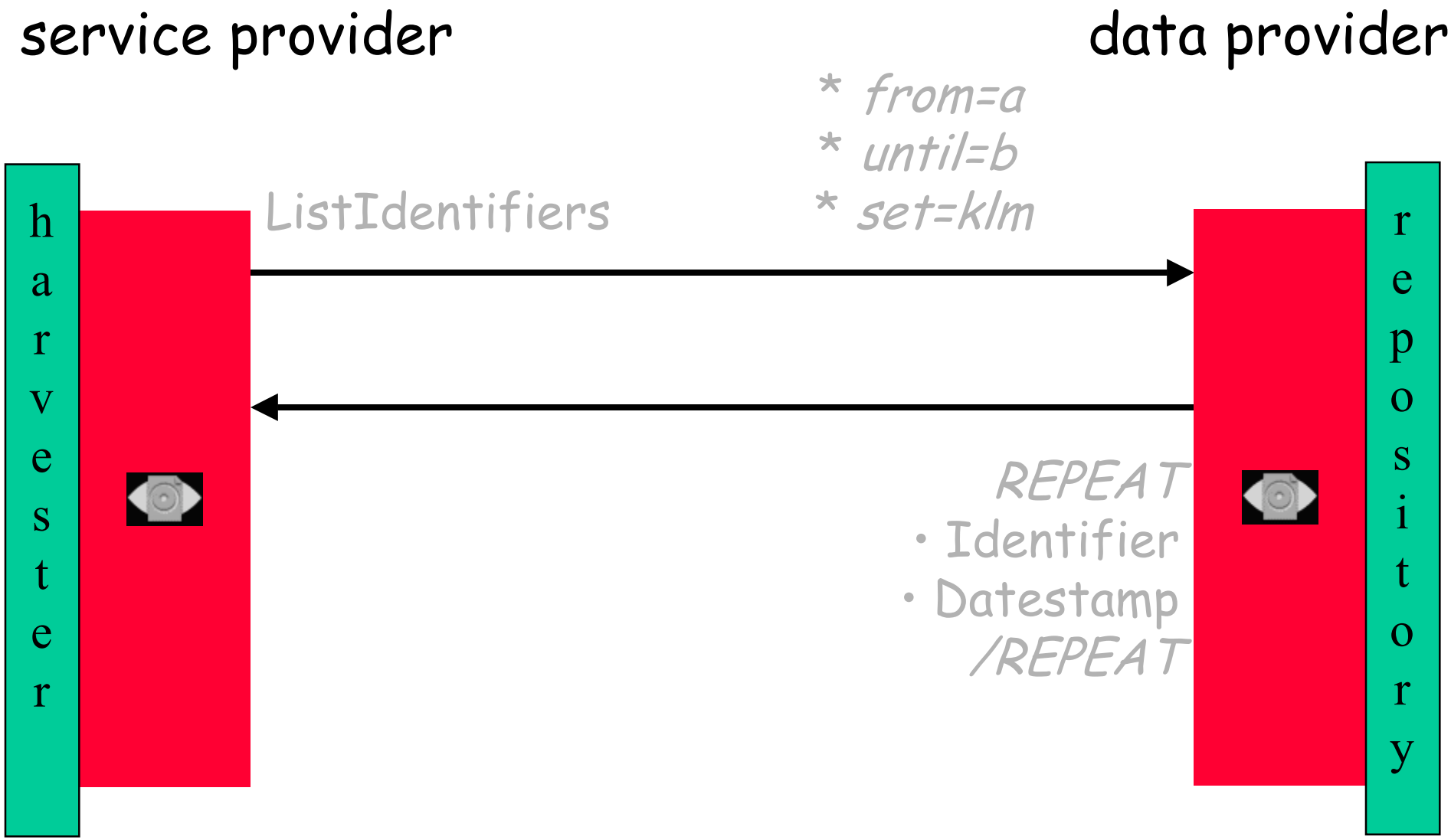
data provider



Harvesting Protocol Requests



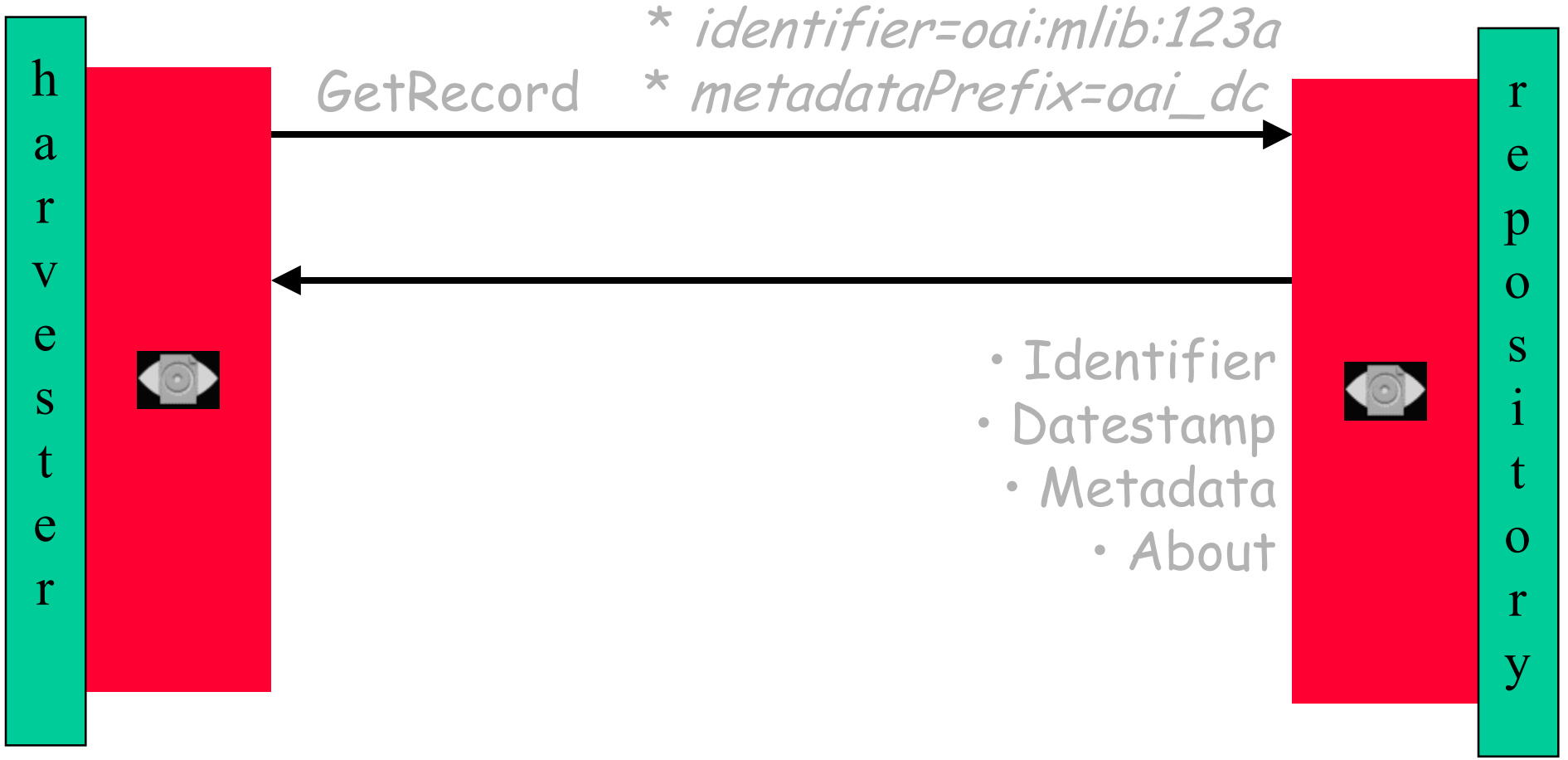
Harvesting Protocol Requests



Harvesting Protocol Requests

service provider

data provider



www.openarchives.org

Open Archives Initiative

OAI

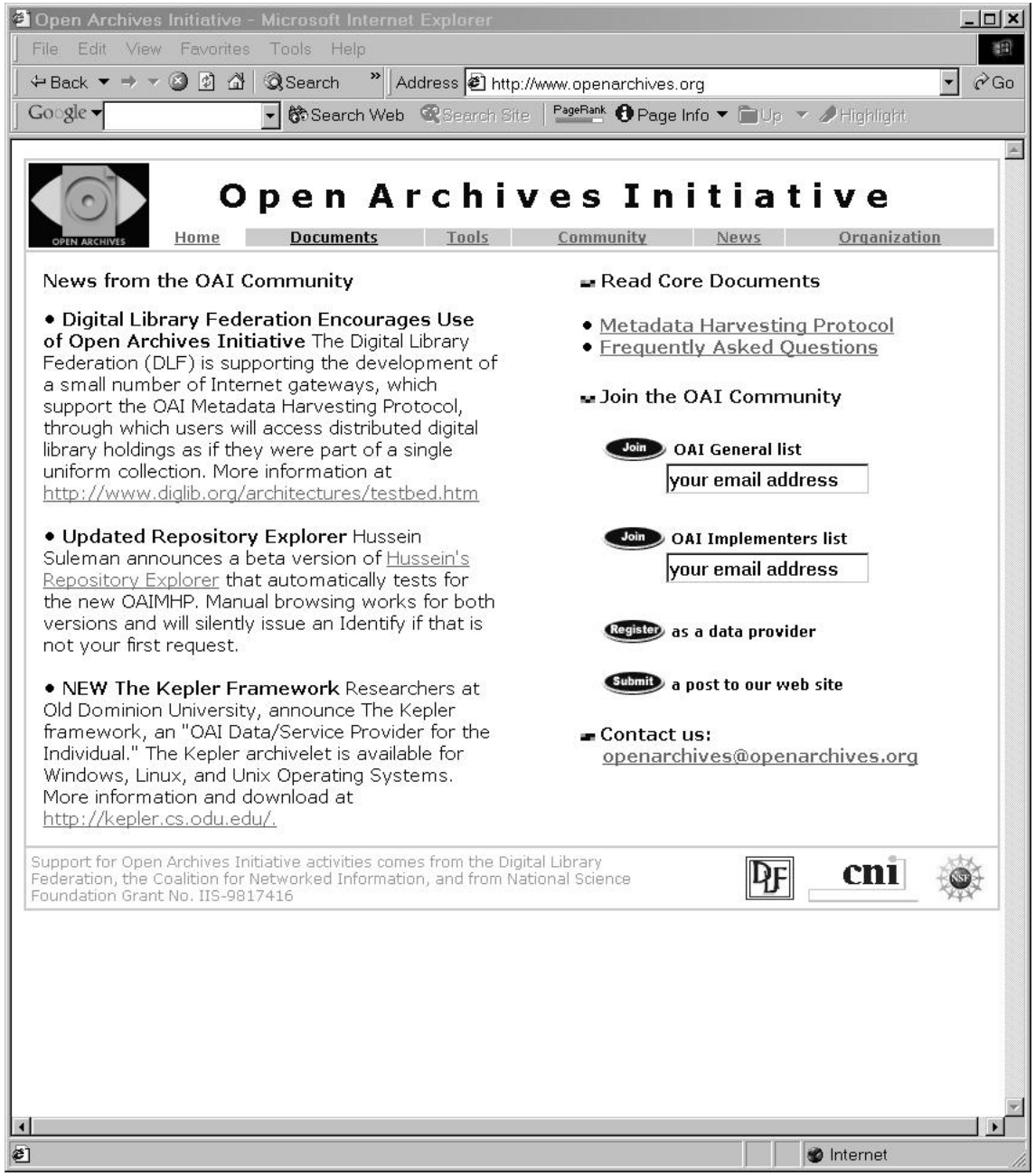
openarchives@
openarchives.org

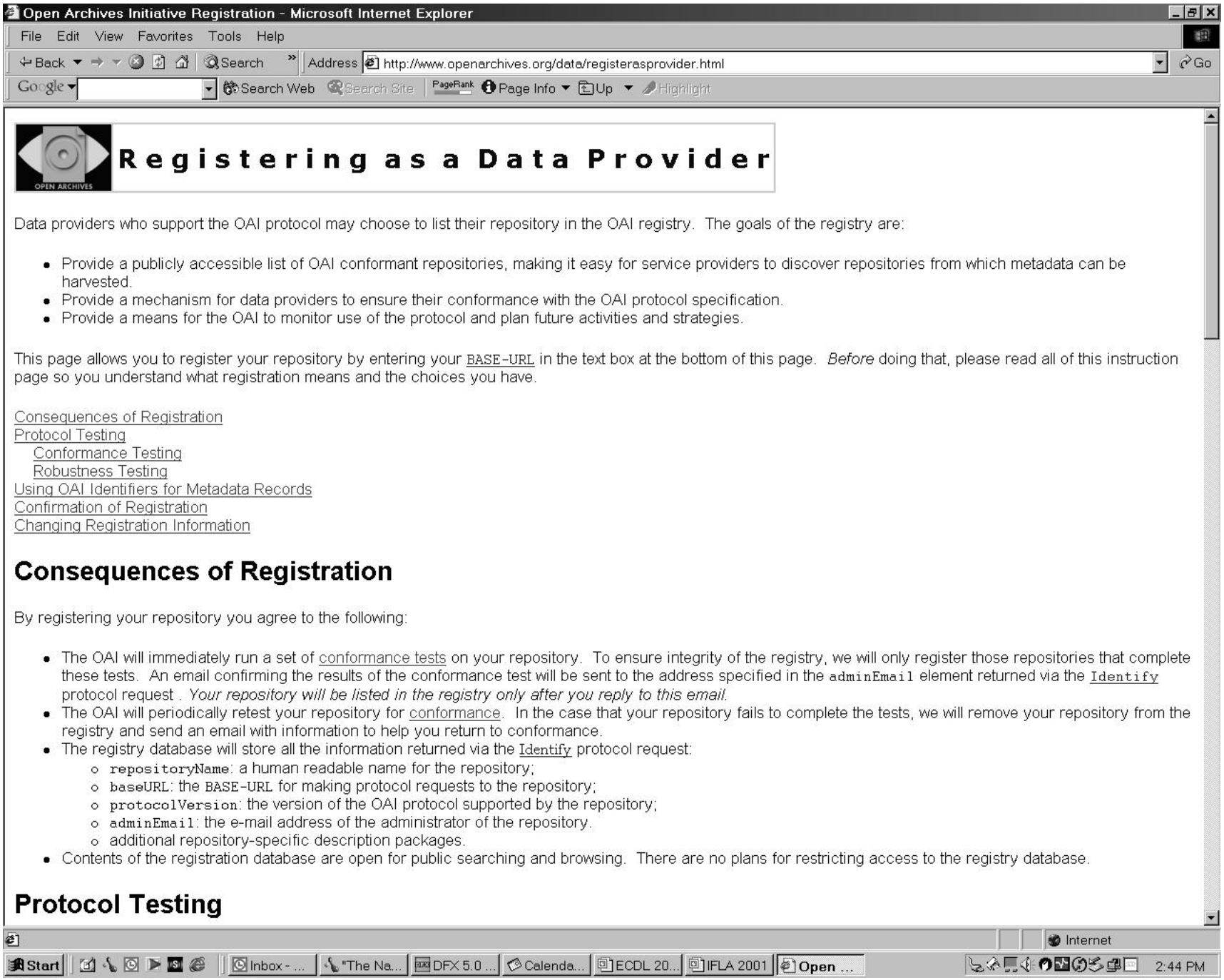
"Opening Remarks & Historical Overview" - ACM SIGIR'2001

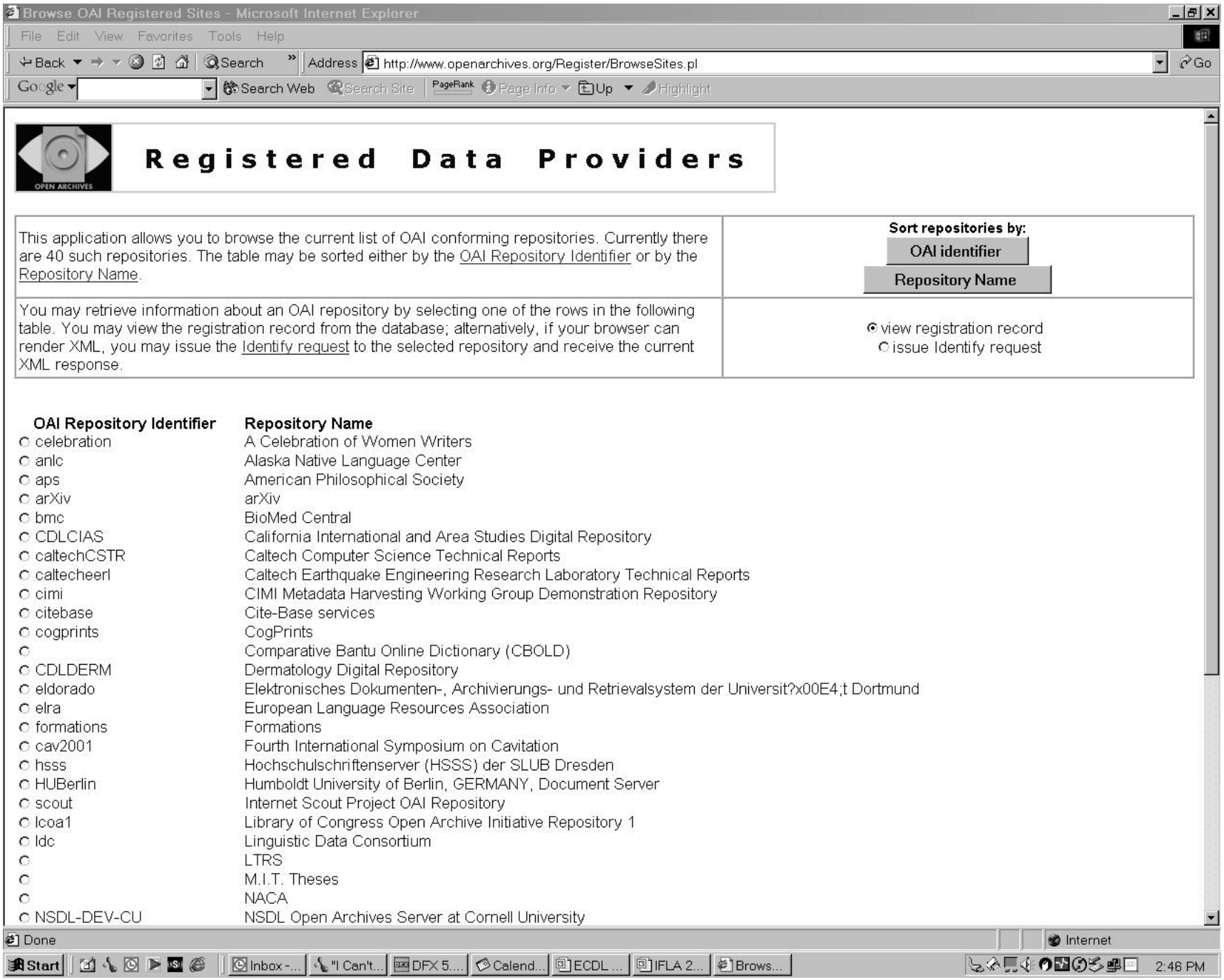
Ed Fox (w. Lagoze & Suleman):B

Other OAI Functions

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


Open Archives Initiative - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Address http://www.openarchives.org Go
Google Search Web Search Site PageRank Page Info Up Highlight
Open Archives Initiative
Home Documents Tools Community News Organization
News from the OAI Community
• Digital Library Federation Encourages Use of Open Archives Initiative The Digital Library Federation (DLF) is supporting the development of a small number of Internet gateways, which support the OAI Metadata Harvesting Protocol, through which users will access distributed digital library holdings as if they were part of a single uniform collection. More information at http://www.diglib.org/architectures/testbed.htm
• Updated Repository Explorer Hussein Suleman announces a beta version of Hussein's Repository Explorer that automatically tests for the new OAIMHP. Manual browsing works for both versions and will silently issue an Identify if that is not your first request.
• NEW The Kepler Framework Researchers at Old Dominion University, announce The Kepler framework, an "OAI Data/Service Provider for the Individual." The Kepler archivelet is available for Windows, Linux, and Unix Operating Systems. More information and download at http://kepler.cs.odu.edu/.
Read Core Documents
• Metadata Harvesting Protocol
• Frequently Asked Questions
Join the OAI Community
Join OAI General list
your email address
Join OAI Implementers list
your email address
Register as a data provider
Submit a post to our web site
Contact us:
openarchives@openarchives.org
Support for Open Archives Initiative activities comes from the Digital Library Federation, the Coalition for Networked Information, and from National Science Foundation Grant No. IIS-9817416
cni
Internet


Open Archives Initiative Registration - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Address http://www.openarchives.org/data/registerasprovider.html Go
Google Search Web Search Site PageRank Page Info Up Highlight
OPEN ARCHIVES
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 Inbox - ... "The Na... DFX 5.0 ... Calenda... ECDL 20... IFLA 2001 Open ... 2:44 PM


Browse OAI Registered Sites - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Address http://www.openarchives.org/Register/BrowseSites.pl Go
Google Search Web Search Site PageRank Page Info Up Highlight
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
Inbox -...
"I Can't...
DFX 5....
Calend...
ECDL ...
IFLA 2...
Brows...
2:46 PM