



# The OAI and OAI-PMH: where to go from here?

Carl Lagoze - Cornell Information Science  
lagoze@cs.cornell.edu

Herbert Van de Sompel - LANL  
herbertv@lanl.gov

OAI3 - CERN - February 12, 2004



# Building on the base

- New infrastructure
- Protocol extensions
- Non-traditional uses
- Research contexts



# New Infrastructure

Building blocks for cross-  
repository federation

# Experimental OAI Registry at UIUC

**Grainger Engineering Library Information Center at  
University of Illinois at Urbana-Champaign**

## Information About This Registry

Search for repositories containing these words in their Identify or ListSets responses or sample records.

Words

OAI Protocol Version:  Any  2.0  1.1  1.0

## **Miscellaneous Reports**

- [All Repositories](#) = 518
- [Repositories Responding](#) = 424
- [Repositories Not Responding](#) = 94
  
- [2.0+ Repositories](#) = 285
- [Pre-2.0 Repositories](#) = 138
  
- [Distinct Metadata Schemas](#)

<http://gita.grainger.uiuc.edu/registry/searchform.asp>

# Extensible Repository Resource Locators (ERRoLs) for OAI Identifiers

---

## Table of Contents

- [1. Introduction](#)
- [2. Supported OAI Repositories](#)
- [3. Item ERRoLs with oai-identifiers](#)
  - [3.1. Examples](#)
- [4. Item ERRoLs with Other Identifiers](#)
  - [4.1. Examples](#)
- [5. Repository ERRoLs](#)
  - [5.1. Examples](#)
- [6. Coordinating Content in OAI Repositories](#)
- [7. OAI Viewer](#)
  - [7.1. Examples](#)
- [8. Caveats](#)
- [9. Credits](#)
- [10. Contact](#)

## 1. Introduction

An ERRoL is a "[Cool URL](#)" to metadata, content, and services related to [registered Open Archive Initiative](#) (OAI) repositories. Following the examples below, anyone can create/use a Cool URL to any metadata record or web resource related to supported OAI repositories.

## 2. Supported OAI Repositories

Any OAI repository can use the ERRoL service by registering a unique repository identifier with the [OAI Registry at UIUC](#).

<http://www.oclc.org/research/projects/oairesolver/default.htm>



# Protocol Extensions

New functionality on a stable base

# OAI Static Repository

- OAI-PMH is low-barrier protocol
- nevertheless, implementation is sometimes not trivial:
  - size of collection does not justify the investment
  - ISP does not allow 3rd party software
  - security considerations

# OAI Static Repository

- research on lowering barrier even further
  - make metadata available in XML files (not dbases)
  - put XML file on web-server
  - make XML file OAI-PMH harvestable
- 2 tracks:
  - autonomous data provider
  - dependent data provider



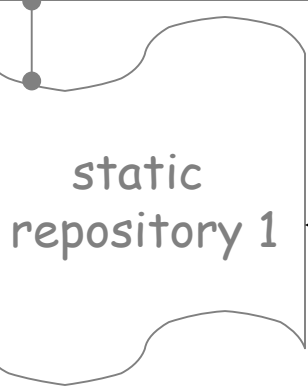
# OAI Static Repository

- autonomous data provider:
    - XML file on web-server
    - XSL style sheet to respond to OAI-PMH requests on web-server
    - requires:
      - native XSLT support in web server
      - XSL v.2 functionality
- => Not (yet) low barrier

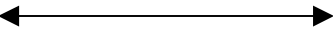
# OAI Static Repository

- dependent data provider:
  - XML file on web-server
  - depend on Gateway to respond to OAI-PMH requests
- requires:
  - *registration* with Gateway
  - Gateway implementation(s)

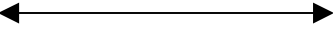
<http://an.oai.org/ma/mini.xml>



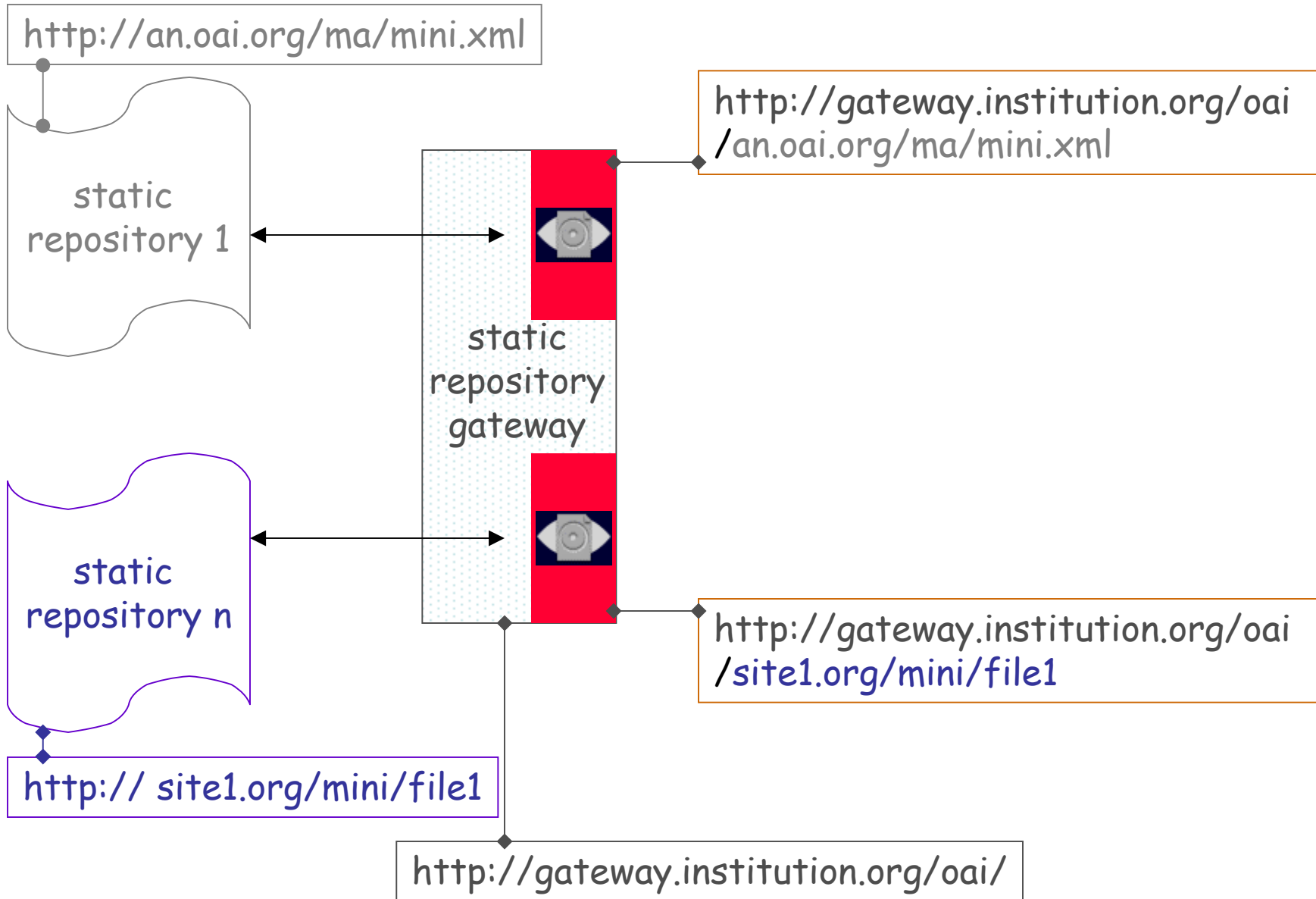
static  
repository 1

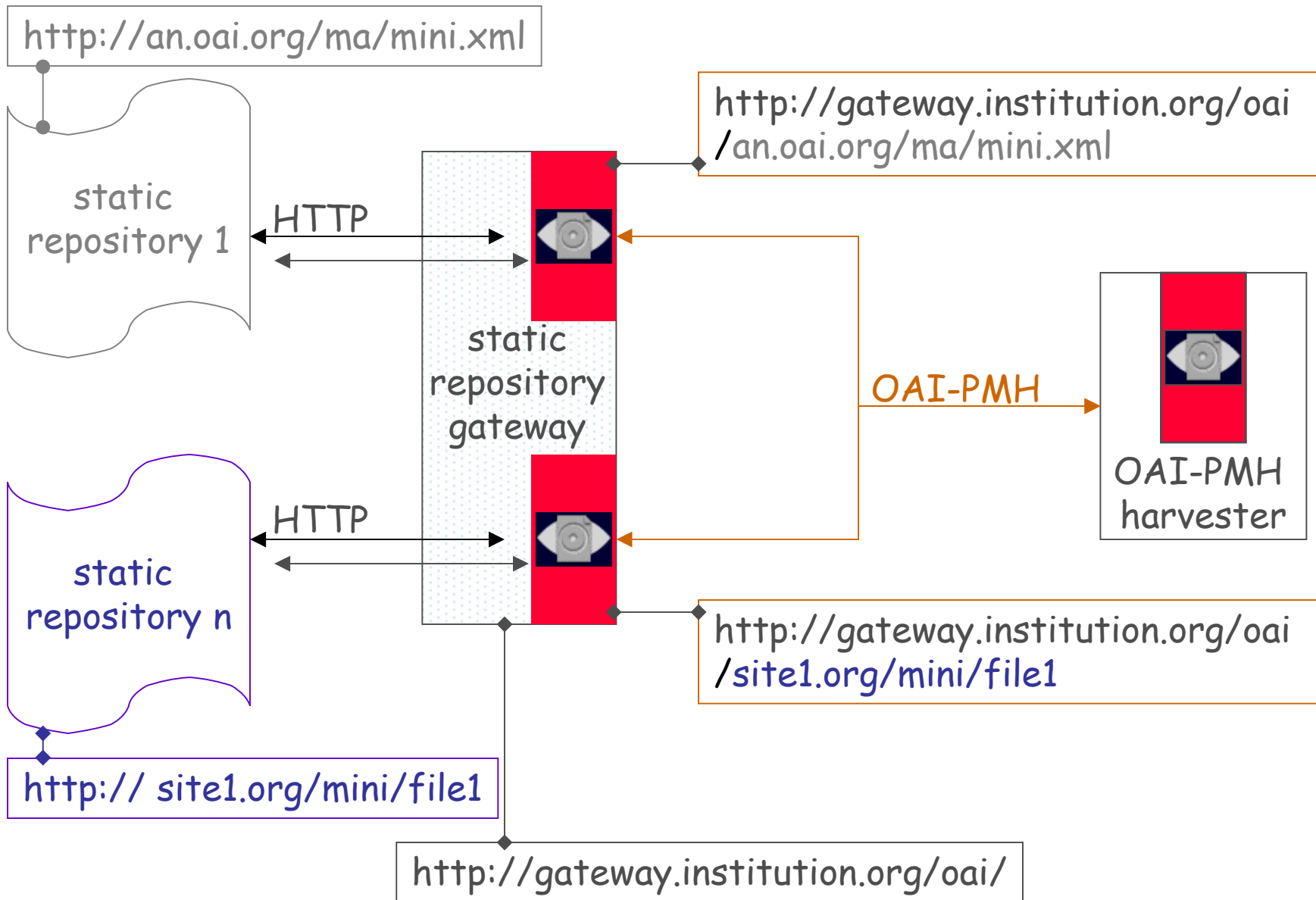


static  
repository n



[http:// site1.org/mini/file1](http://site1.org/mini/file1)





# LANL Static Repository Gateway

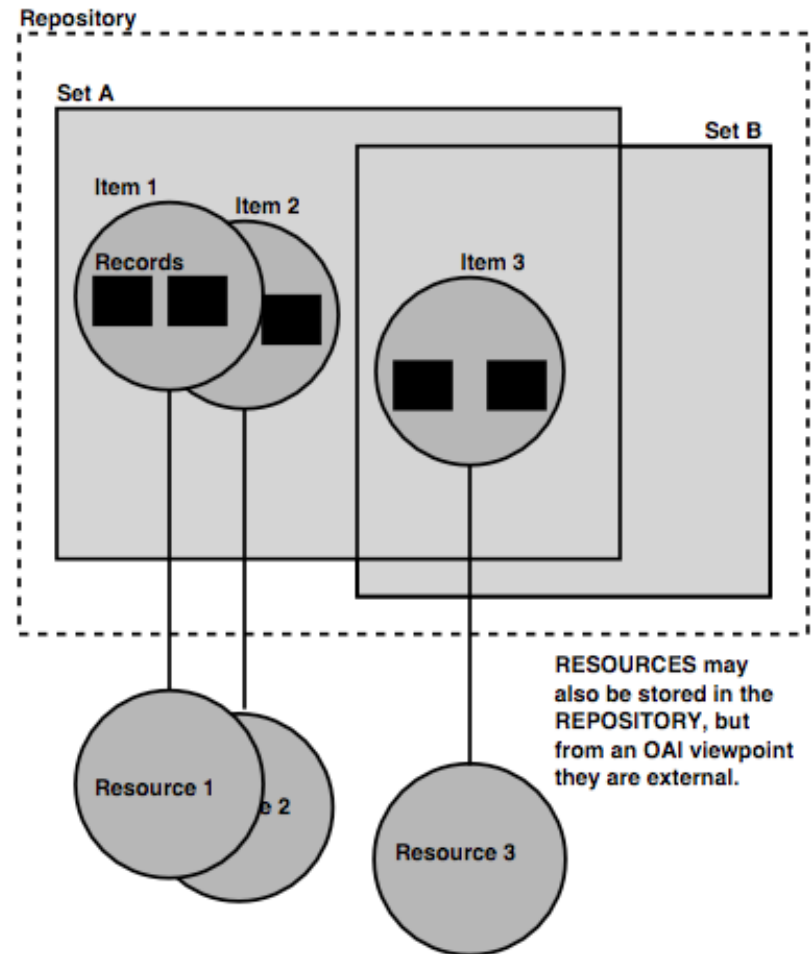
- *The OAI-PMH Static Repository and Static Repository Gateway* - Patrick Hochstenbach, Henry Jerez, Herbert Van de Sompel <http://lib-www.lanl.gov/~herbertv/papers/jcdl2003-submitted-draft.pdf>
- Experimental registration system - <http://libtest.lanl.gov/registry.htm>
- Sourceforge download site - <https://sourceforge.net/projects/srepod/>

# OAI Rights

- Motivations
  - Distinction between data and metadata fuzzy, especially regarding intellectual property
  - XML content already fits into protocol
  - Consumers of metadata are almost always interested in access to underlying resource
- Scope
  - No new definition of a rights expression language
  - Avoid restriction to any rights language
    - Initial prototypes with Creative Commons licenses

# OAI rights issues

- Entity Association
  - Focus on rights expressions for metadata and associated resources
- Aggregation association
  - OAI-PMH entities: repository, resource, item, record, set
- Binding
  - Use about container for metadata rights exp.
  - Designated metadata prefix to contain resource rights exp.







# Non-traditional usage

Beyond metadata for resource  
discovery



# OAI-PMH-based access to DL usage logs

<http://www.dlib.org/dlib/july03/young/07young.html>

# OAI-PMH access to DL usage logs

- usage logs filtered and stored in MySQL db
- accessible as 2 OAI-PMH repositories:
  - document oriented
  - agent oriented (user-proxy)
  - interlinked
- recommender system:
  - harvests logs
  - interpretes logs
  - exposes relationships (OpenURL access)

resource



about

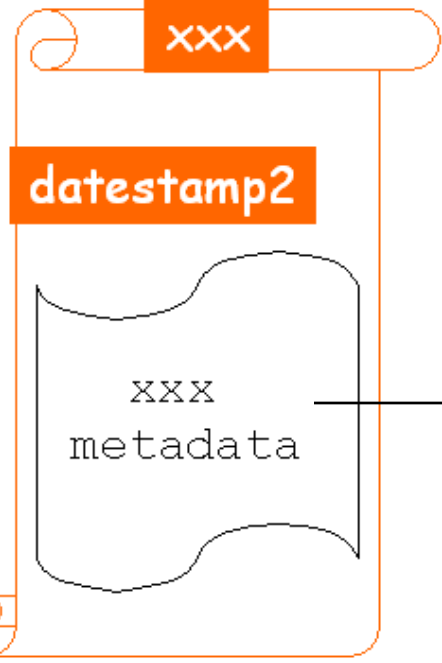
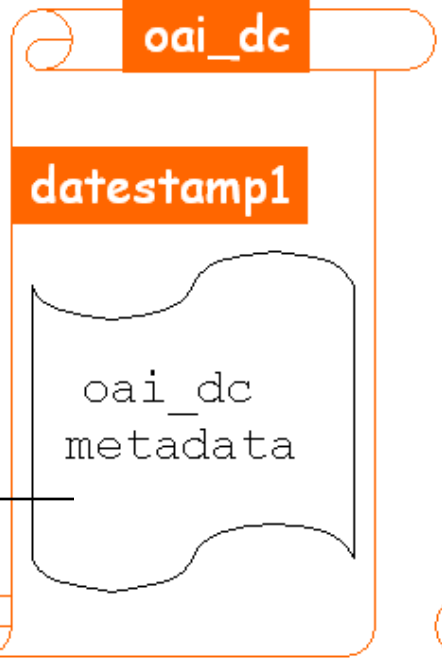
**alog:IP:128.1.22.13**

**identifier**

item



metadata records



**metadataPrefix**

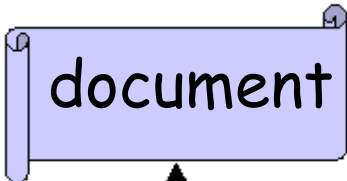
**datestamp**

docs accessed by agent

about agent



resource



about

dlog:ori:pmid:258471

identifier

item



metadata records

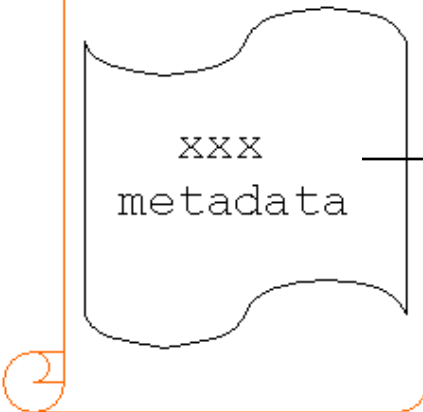


metadataPrefix

datestamp1

datestamp2

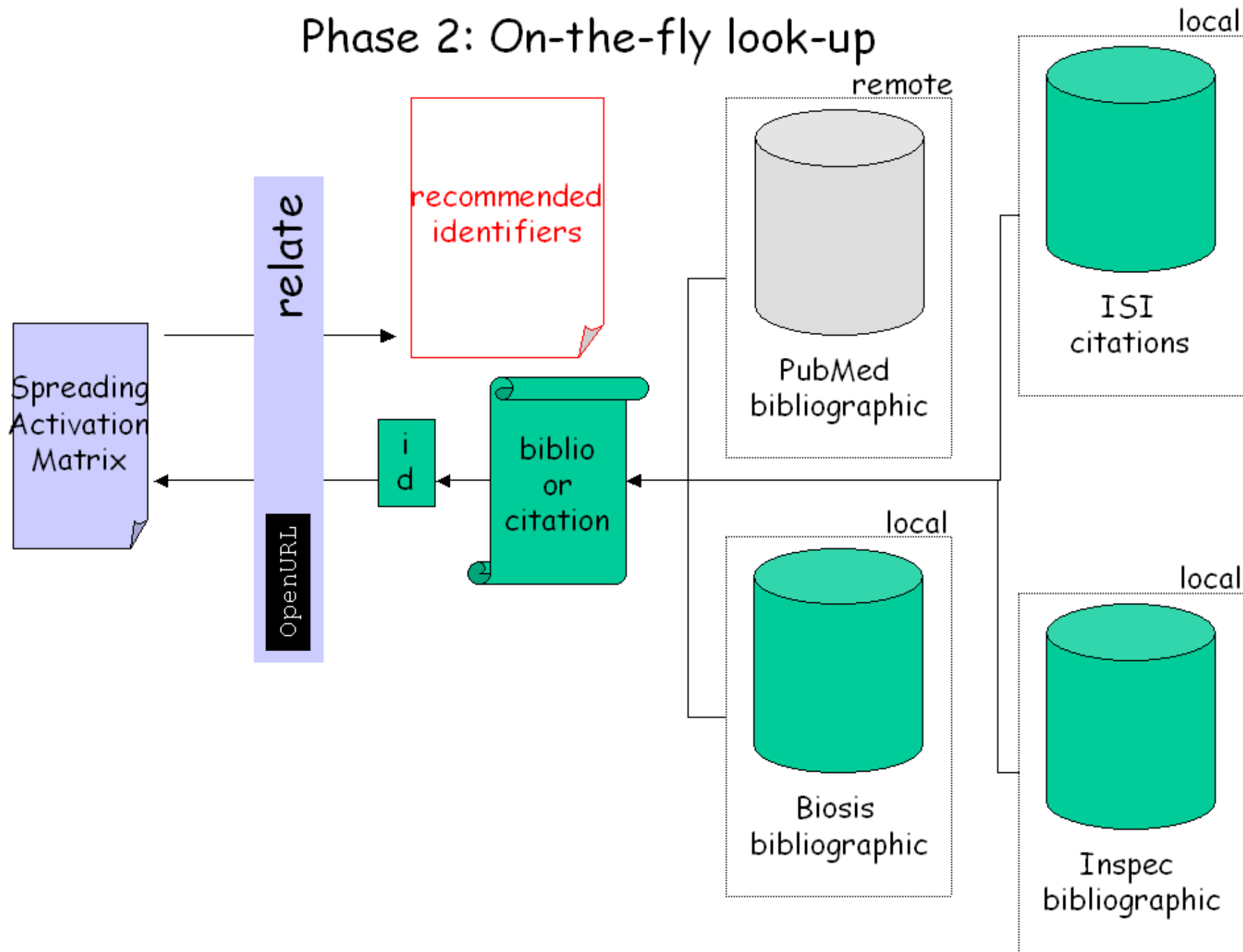
datestamp



about document

agents accessing the document

## Phase 2: On-the-fly look-up



# LANL Repository Architecture

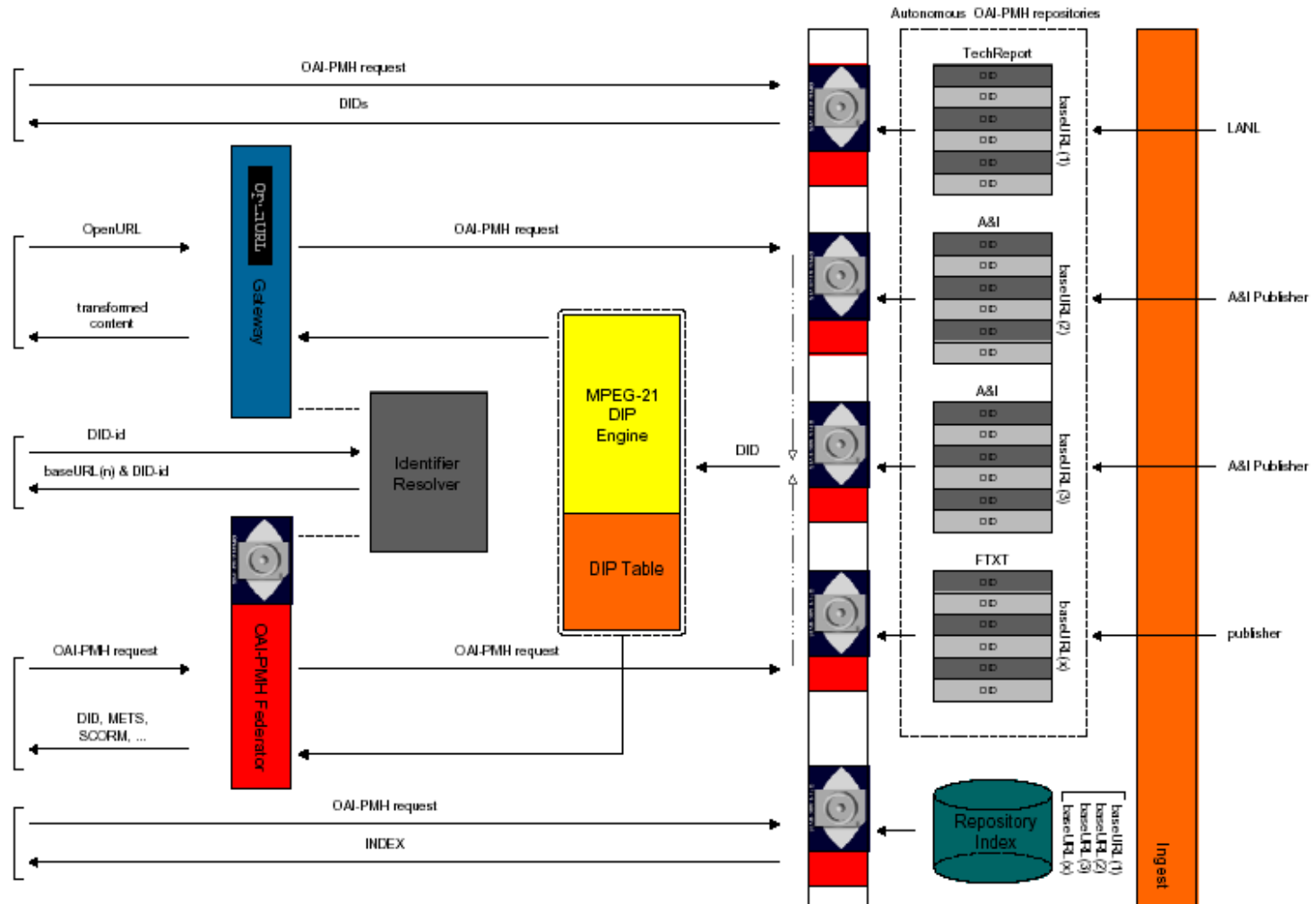
- Problem: provide multiple service access to variety of locally hosted assets
- Assets include secondary assets (ISI, BIOSIS, Inspec, etc.) and primary feeds (Elsevier, Wiley, IOP, APS, etc.)
- Common representation of assets using MPEG-21 DIDL
  - Facility for multiple disseminations
- Components of architecture federated through OAI-PMH

# LANL Repository Architecture Components

- *Asset repositories* - one per data feed with assets stored as DIDLs, harvestable by OAI-PMH
- *Repository index* - keeps track of creation and location of data repositories, harvestable by OAI-PMH
- *Identifier resolver* - single point resolution to get repository location of DIDL object.
- *OAI-PMH federator* - single point OAI access for service clients



# LANL Repository Architecture



# LANL Repository Architecture

- D-Lib nov 2003 :  
<http://dx.doi.org/10.1045/november2003-bekaert> (MPEG-21 DIDL use)
- D-Lib fed 2004 :  
<http://dx.doi.org/10.1045/february2004-bekaert> (MPEG-21 and OpenURL based dissemination architecture)
- Submission to JCDDL 2004



# Experimentation

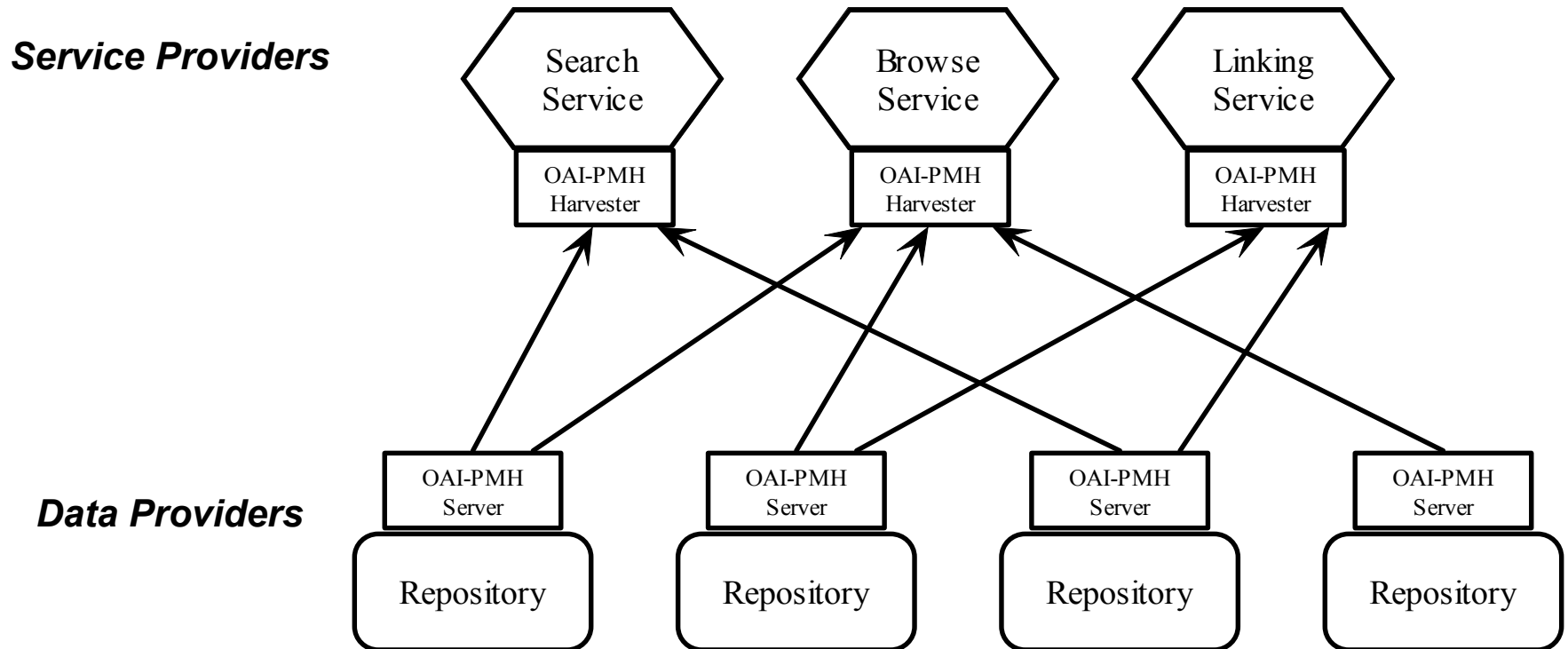
Exploration of new contexts



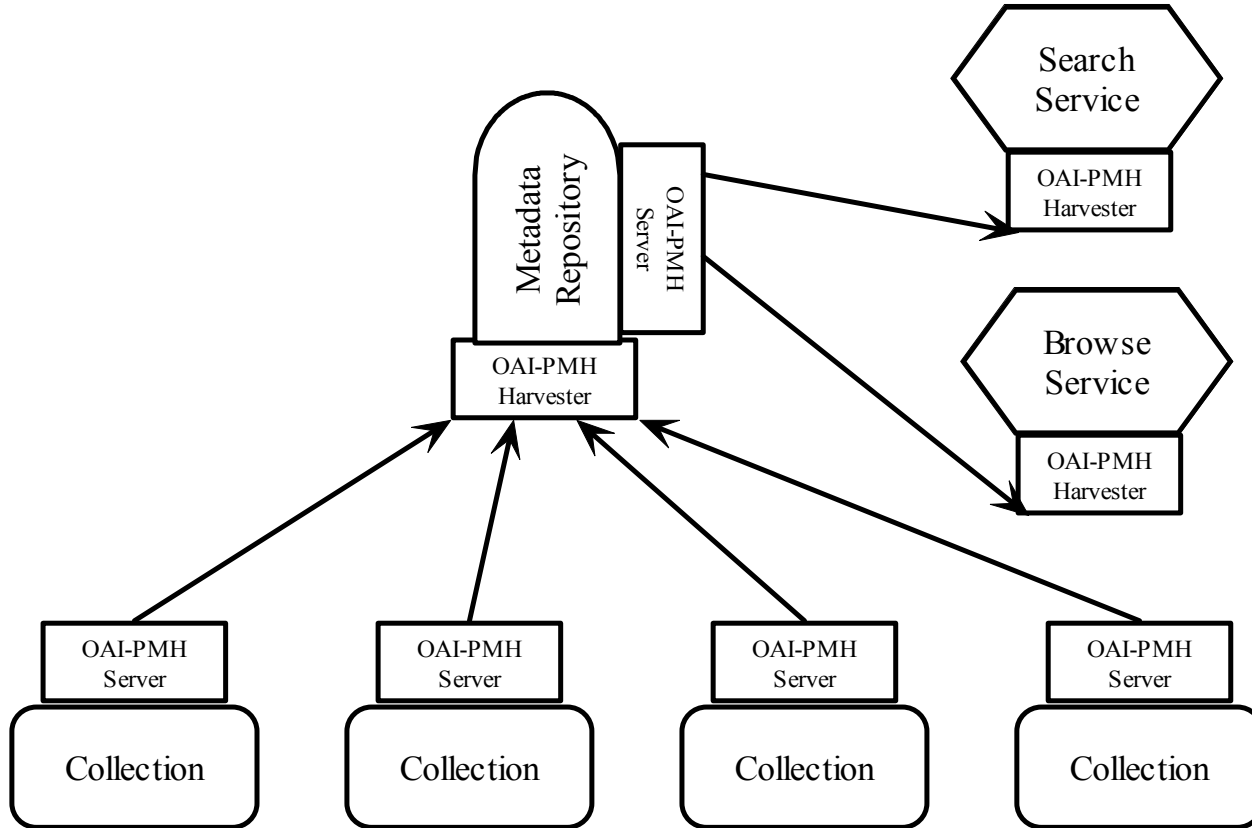
# OAI and P2P

Enabling a metadata refinement network that enables the creation of document value chains

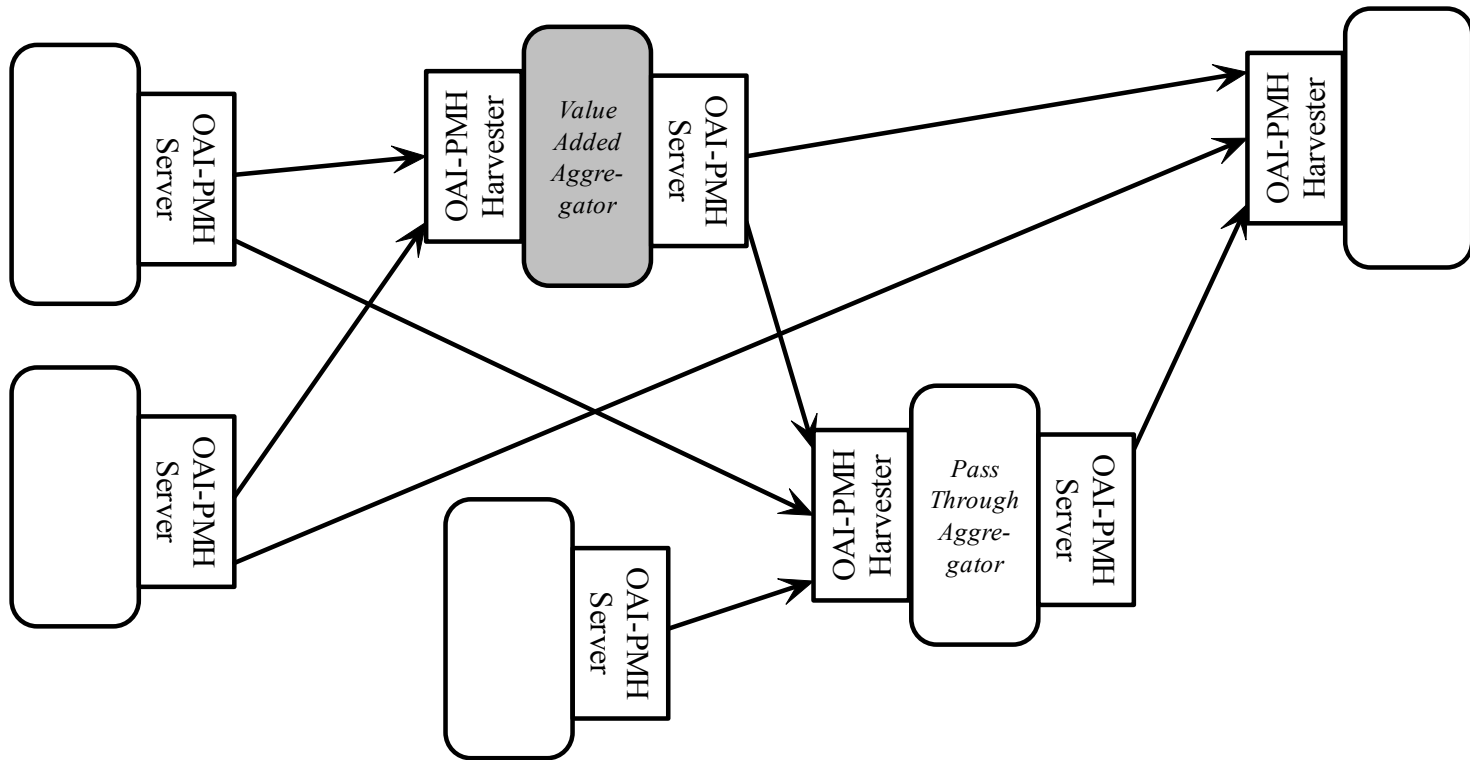
# Original OAI-PMH Model



# Hybrid Model with Aggregator



# Metadata Exchange Graph



# Implementation Questions

- Underlying framework
  - JXTA
- Metadata item/record location
  - Broadcast search
  - Distributed Hash Tables
- Provenance chains
  - Exploit provenance information in OAI-PMH
  - Logical joins based on provenance information
- Network Harvesting
  - Efficient range queries using P-trees

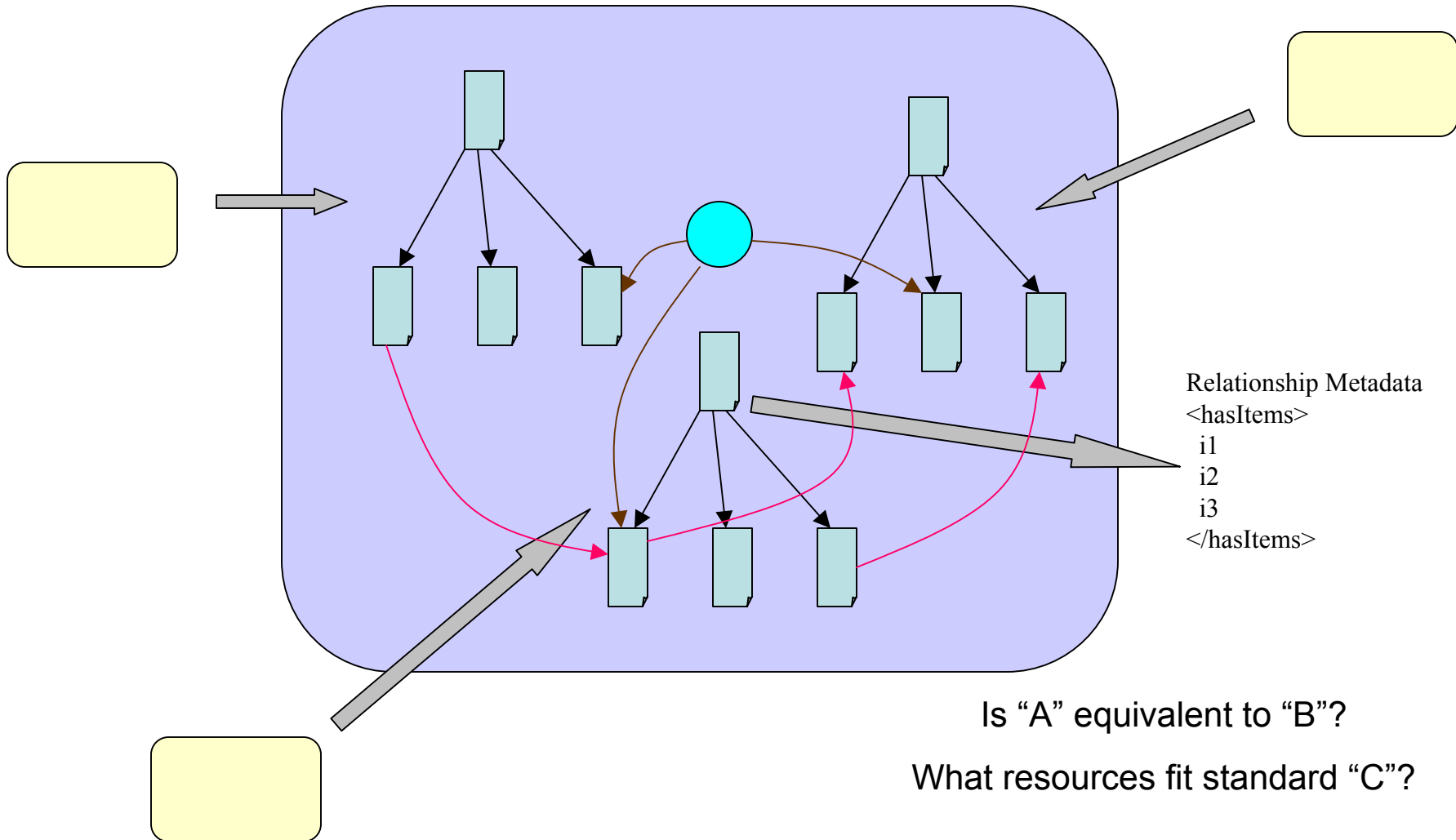




# OAI and RDF

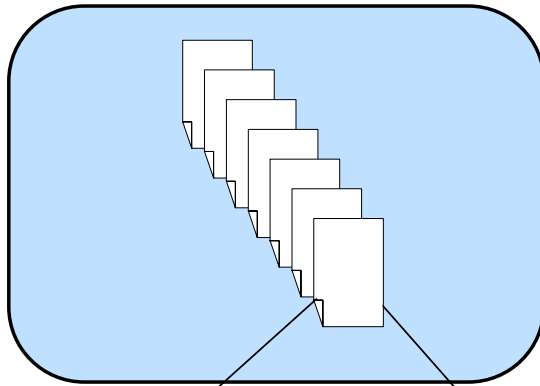
Expressing relationships among  
metadata records

# NSDL Metadata Repository (1)

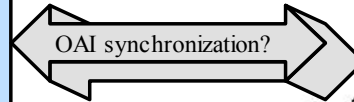
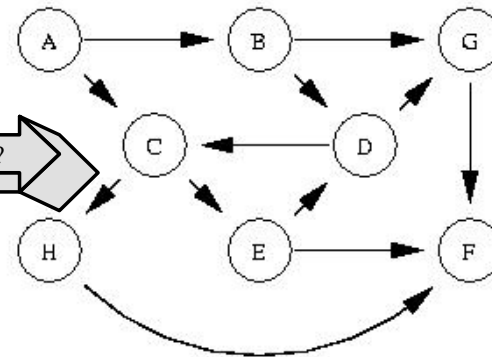


# NSDL Metadata Repository (2)

Fedora Content/Metadata Store



Jena Relationship Store



```
<rdf:Description about="ID1">  
  <nsdlrel:hasMember>ID2</nsdlrel:hasMember>  
  <nsdlrel:conformsTo>STD4</nsdlrel:conformsTo>  
</rdf:Description>
```

Issues:

- push/pull model?
- schema validation