

# Service Providers: Future Perspectives

Michael L. Nelson  
Old Dominion University  
Norfolk Virginia, USA  
mln@cs.odu.edu  
<http://www.cs.odu.edu/~mln/>

2nd Workshop on the Open Archives Initiative:  
Gaining Independence With E-print  
Archives and OAI



CERN, Switzerland  
October 18, 2002



# Outline

- History of the history of OAI-PMH
- (Traditional) public service providers not present for this meeting
- Why the OAI-PMH is not important
- Defining the OAI-PMH data model
- Abusing the OAI-PMH data model
- Current and nearly-current interesting services

# OAI-PMH Meeting History

OAI Open Day,  
Washington DC  
1/2001

---

This meeting  
CERN 10/2002

---

4

Protocol definition,  
development tools

1

5

DPs, retrofitting  
existing DLs

4

1

SPs, new services

11

0

Socio-Economic-  
Political Issues

6

# Shift of Topics

- From the protocol itself, supporting & debugging tools and how to retrofit (existing) DLs...
- ...to building (new) services that use the OAI-PMH as a core technology and reporting on their impact to the institution/community

# NTRS

NASA Technical Reports Server

[About NTRS](#) [News](#) [Feedback](#) [Help](#) [Weekly Updates](#) [Simple Search](#) [Advanced Search](#) [Browse](#)

Title:

Author(s):

Date:

Report Number / Journal / Conference:

Abstract:

Combine terms with:  and  or

Combine fields with:  and  or

Order results by:

Order of results:

Results per page:

Select the databases you would like to search

NASA Archives	Non-NASA Archives
<input checked="" type="checkbox"/> Ames Research Center	<input type="checkbox"/> Aeronautical Research Committee (UK)
<input checked="" type="checkbox"/> Goddard Space Flight Center	
<input checked="" type="checkbox"/> ICASE (Langley Research Center)	
<input checked="" type="checkbox"/> Johnson Space Flight Center	
<input checked="" type="checkbox"/> Kennedy Space Center	
<input checked="" type="checkbox"/> Langley Research Center	
<input checked="" type="checkbox"/> National Advisory Committee for Aeronautics (NACA)	
<input checked="" type="checkbox"/> RIACS (Ames Research Center)	
<input checked="" type="checkbox"/> Stennis Space Center	

Debug info

- <http://ntrs.nasa.gov/>
- metadata harvesting replacement for <http://techreports.larc.nasa.gov/cgi-bin/NTRS>
  - previous NTRS was based on distributed searching
  - hierarchical harvesting
- (nigh) publicly available

# Arc

The screenshot shows a browser window with the address bar containing `http://arc.cs.odu.edu/`. The page title is "ARC - A Cross Archive Search Service" and the subtitle is "Old Dominion University Digital Library Research Group". The page features a search interface with a "Simple Search" section, a search input field, and dropdown menus for "Group Results By" (set to "Archive") and "Sort Results By" (set to "Relevance Ranking"). A "Search" button is located below the input field. To the right, there is a "Related Links" section with links to "Download Source Code", "Other OAI Service Provider", and "DP9- An OAI Gateway Service for Web Crawlers". At the bottom, there is a paragraph of text describing the service and a footer with the contact email `dlib@list.odu.edu`.

- `http://arc.cs.odu.edu/`
- harvests all known archives
- first end-user service provider
- source available through SourceForge
- hierarchical harvesting

# NCSTRL

The screenshot shows a web browser window with the address bar set to <http://www.ncstrl.org/>. The page title is "Networked Computer Science Technical Reference Library". The navigation menu includes "Simple Search", "Advanced Search", "Browse", "Register", "Submit to CoRR", "About NCSTRL", "OAI", and "Help".

**Search specific bibliographic fields**

Author   
Title   
Abstract

Combine fields with  AND  OR

**Filter options**

Archive   
Archive's Set

DateStamp   
Discovery Date

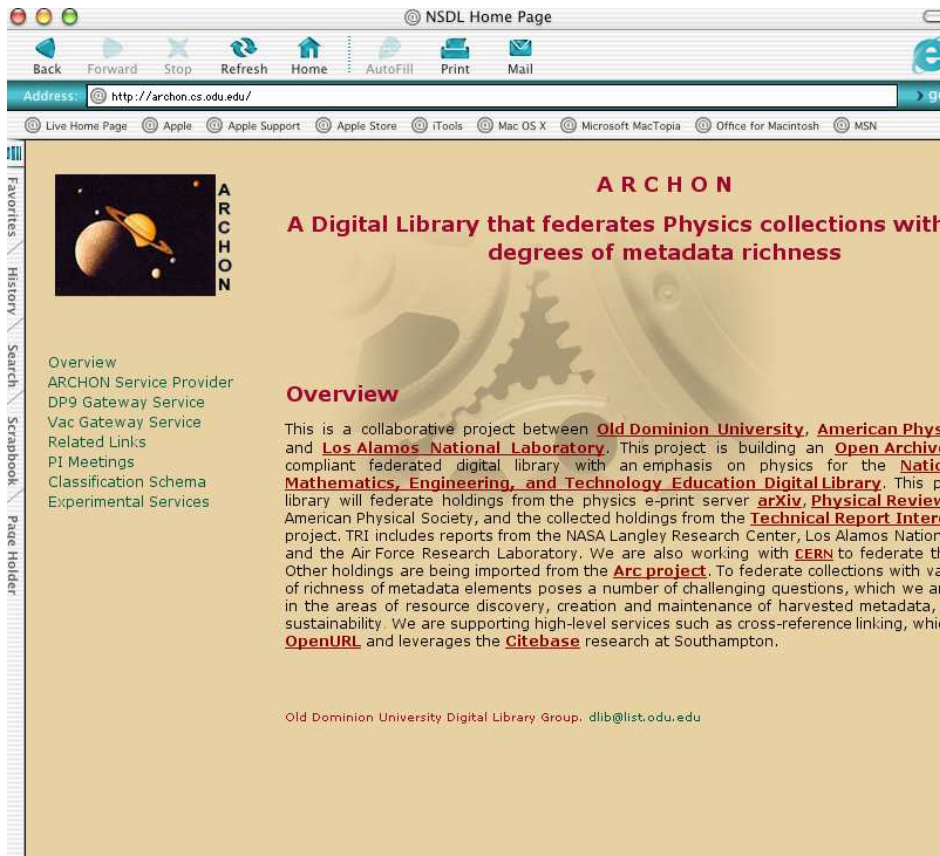
**Display options**

Group by   
Sort by

Note: If 6.0 users, please enable third party cookies  
This is a collaborative project involving [NASA Langley](#), [Old Dominion University](#), [University of Virginia](#) and [Virginia Tech](#).  
Powered by [Old Dominion University](#)

- <http://www.ncstrl.org/>
- metadata harvesting replacement for Dienst-based NCSTRL
- based on Arc
- computer science metadata

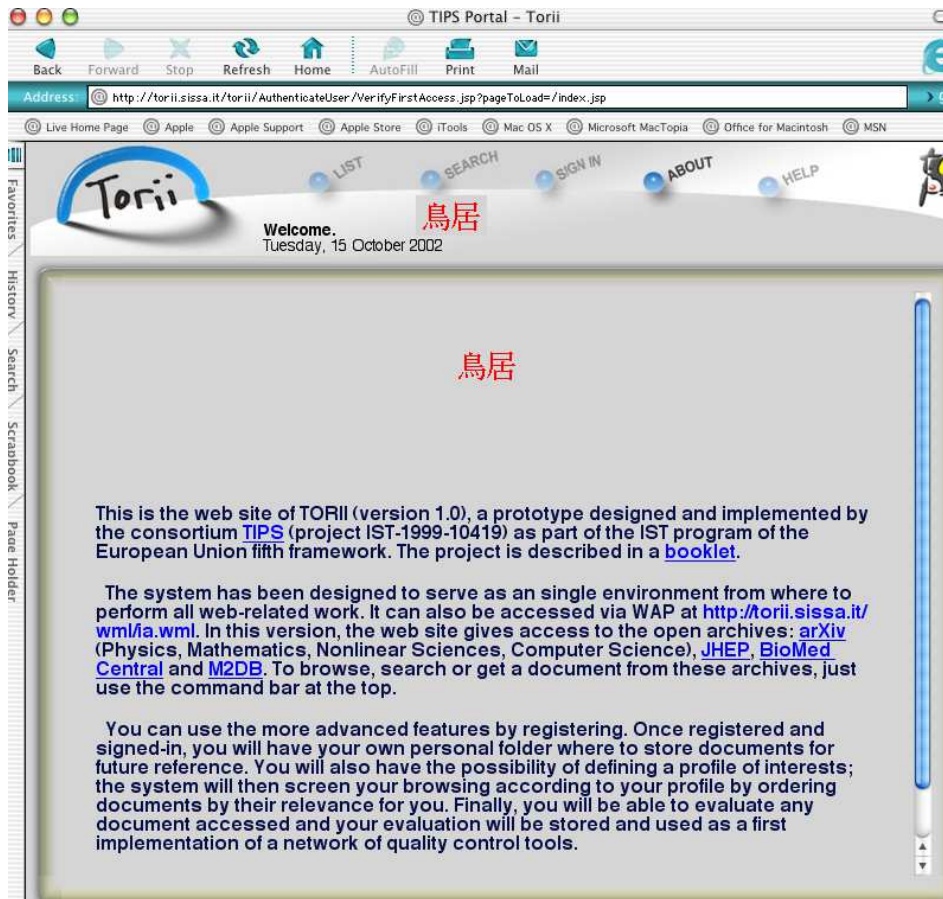
# Archon



- <http://archon.cs.odu.edu/>
- physics metadata
- based on Arc
- features:
  - citation indexing
  - equation-based searching

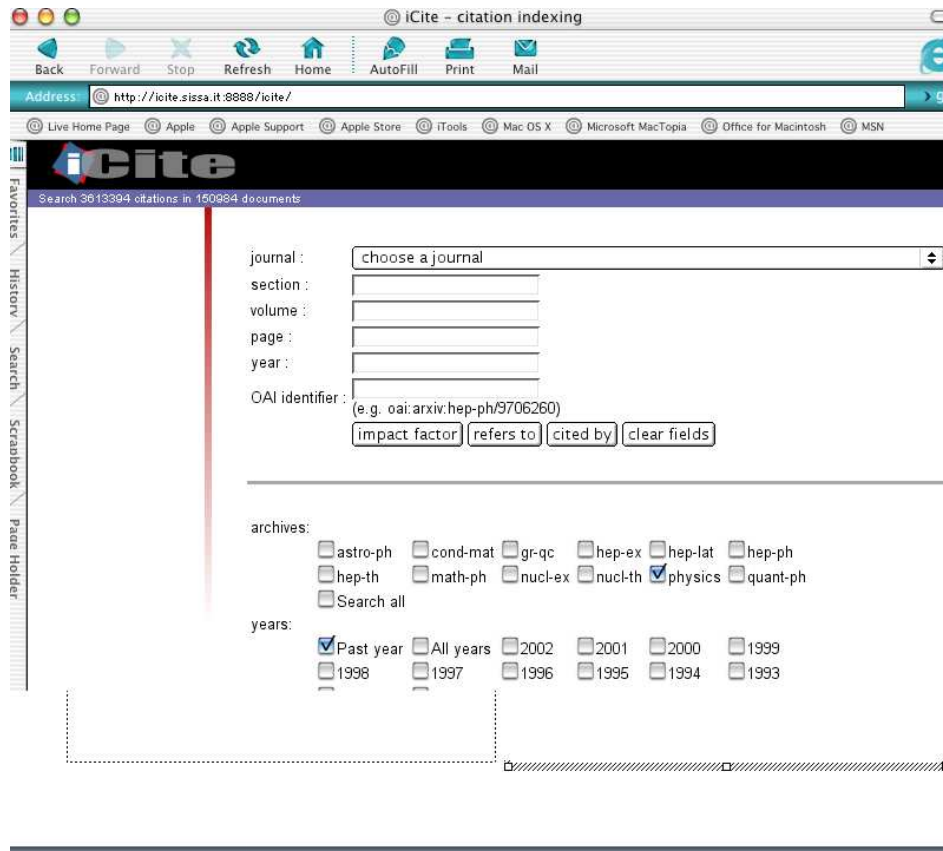


# Torii



- <http://torii.sissa.it/>
- physics metadata
- features
  - personalization
  - recommendations
  - WAP access

# iCite



- <http://icite.sissa.it/>
- physics metadata
- features
  - citation based access to arXiv metadata

# my.OAI

The screenshot shows a web browser window titled "Simple Search Form" with the address bar containing "http://www.myoai.com/search/Search.cgi/LoginForm?Login=quest&Password=quest". The page content includes a navigation menu with links for "Search Form", "Search History", "Saved Searches", "Document Folders", "User Preferences", and "Create Account". Below the menu, the "Simple Search Form" section contains several input fields and controls: a "View search results" dropdown menu with a "Do It!" button; a "Search in any field:" section with a text input and a "Learn More" link; a "Search for words in the field:" section with a dropdown menu set to "Title" and a text input, and another dropdown menu set to "Creator" with a text input, both with "Learn More" links; a "Search for:" section with a dropdown menu set to "All the terms (AND within fields)" and a "Learn More" link; a "Select the databases you wish to search:" section with a "Select all" and "Clear all" link, a list of databases including "arXiv", "BioMed Central", "Chemistry Preprint Server", "Cite-Base services", and "Humboldt University of Berlin", and a "Learn More" link; a "Number of documents to retrieve per page:" section with a dropdown menu set to "10"; and a "Sort documents by:" section with a dropdown menu set to "Date - Oldest first". At the bottom, there is a footer with links for "Home", "Help", and "Feedback", and a copyright notice: "Copyright (c) 1993-2002 FS Consulting, Inc. All rights reserved."

- <http://www.myoai.com/>
- covers all registered metadata
- features
  - result sets
  - personalization
  - many other advanced features

# Cyclades



- <http://www.ercim.org/cyclades>
- scientific metadata
- features
  - personalization
  - recommendations
  - collaboration
- status?

# citebase

- <http://citebase.eprints.org/>
- arXiv metadata
- citation based indexing, reporting

The screenshot shows a web browser window titled "citebase Search". The address bar contains the URL "http://citebase.eprints.org/cgi-bin/search". The browser's navigation bar includes buttons for Back, Forward, Stop, Refresh, Home, AutoFill, Print, and Mail. The browser's tab bar shows several open tabs, including "Live Home Page", "Apple", "Apple Support", "Apple Store", "iTools", "Mac OS X", "Microsoft MacTopia", "Office for Macintosh", and "MSN". The browser's status bar shows ".eprints.org sites at Southampton serving Open Archives" and "-- Select a site --".

The main content area of the browser displays the "citebase Search" page. The page title is "citebase Search". Below the title are links for "Help and About" and "Impact Health-Warning". A warning box states: "Citebase is currently only an experimental demonstration. Users are cautioned not to use it for academic evaluation yet. Citation coverage and analysis is incomplete and hit coverage and analysis is both incomplete and noisy." Below the warning box are three tabs: "Metadata", "Citation", and "Identifier". The "Metadata" tab is selected. The search form includes fields for "Author(s) (explanation)", "Title/Abstract Keywords", "Publication title", and "Creation Date" (with "from" and "until" sub-fields). Below the search form are two dropdown menus: "Rank matches by:" (set to "Descending") and "Citations (Paper)". There are also "Search" and "Clear" buttons. At the bottom of the page, a "User Survey" section is visible, with the text: "Links to Citebase are a new feature of arXiv. Please help us improve Citebase by performing a short evaluation exercise, which will also show you how to get more functions out of its ranked search services: <http://citebase.eprints.org/survey/>".

# OAIster

Use this form to find any digital resource (what is a digital resource?) in any digital collection (which collections are available?) in OAIster.

**Want to search for separate words in all the fields?** Use one word in the "Search all fields" box and one word in the "Keyword" box. For example, add "chicago" in the "Search all fields" box and "art" in the "Keyword" box. (You could also do this with separate phrases.) See [help](#) for more search tips.

**Search all fields**  
Use a word or phrase, e.g., "diploma\*", "fancy dress"

  
[\(help\)](#) 

**Or, search within particular fields**  
Use a word or phrase, e.g., "diploma\*", "fancy dress"

Keyword:  [\(help\)](#)

Title:  [\(help\)](#)

Author:  [\(help\)](#)

Subject:  [\(help\)](#)

Resource Type:  [\(help\)](#)

Product of DLPS  
For more info please contact  
[oaister@umich.edu](mailto:oaister@umich.edu)

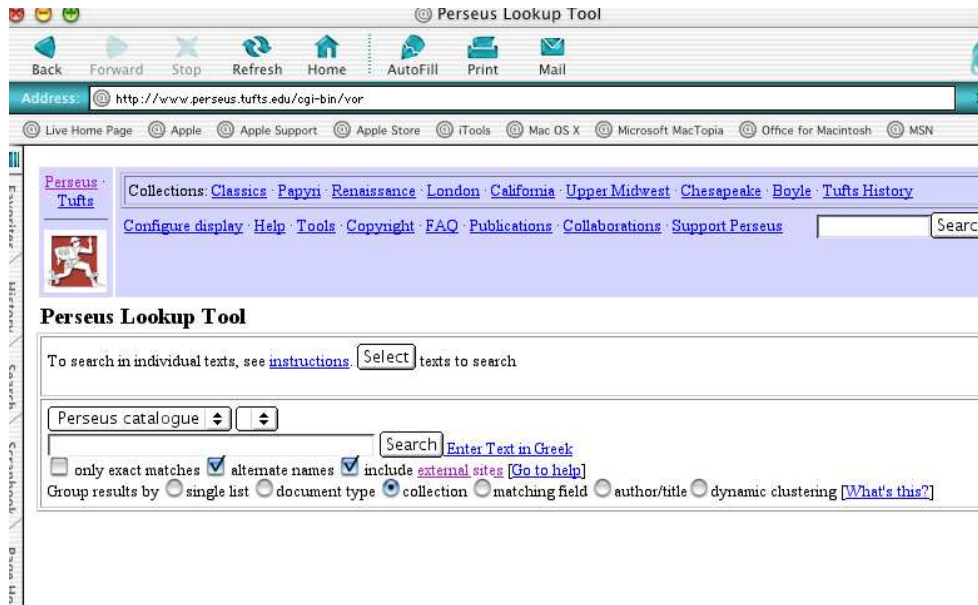
- <http://oaister.umdl.umich.edu/>
- harvests all known archives

# Public Knowledge Project

The screenshot shows a web browser window titled "PKP Open Archives Harvester: Metadata Search". The address bar contains the URL "http://www.pkp.ubc.ca/harvester/search.php?search=advanced". The page header includes the "PUBLIC KNOWLEDGE PROJECT" logo and the text "Open Archives Harvester". Below the header, there is a navigation menu with "Home > Metadata Search". The main content area is titled "Advanced OAI Search" and contains a paragraph explaining the site's adherence to the Open Archives Initiative. Below this, there are search filters: "Search PKP databases:" with a dropdown menu set to "-- All Indexed Archives --", and "Search all categories for:" with an empty text input. A section titled "Author(s):" includes fields for "Author Name:", "Affiliation:", "Title:", and "Abstract:". The "Date:" section has "From:" and "To:" fields, each with three dropdown menus for year, month, and day. Below this is the "Index terms" section, which includes a link to the "Library of Congress Classification Outline" and fields for "Discipline(s):", "Subject(s):", "Approach/Method:", and "Coverage:". At the bottom, there is a "Language:" section with a dropdown menu and the text "English=en; French=fr; Spanish=es (Additional codes)".

- <http://www.pkp.ubc.ca/harvester/>
- domain-specific filtering of harvested metadata (?)

# Perseus



- <http://www.perseus.tufts.edu/>
- they claim to harvest all DPs, but only humanities related DPs appear in the pull down menu



# Service Providers

- It is clear that SPs are proliferating, despite (because of?) the inherent bias toward DPs in the protocol
  - easy to be a DP -> many DPs -> SPs eventually emerge
  - hard to be a DP -> SPs starve
  - currently 5x DPs more than SPs
- SPs are beginning to offer increasingly sophisticated services
  - competitive market originally envisioned for SPs is emerging

# Why The OAI-PMH is NOT Important

- Users don't care
- OAI-PMH is middleware
  - if done right, the uninterested user should never have to know
- Using the OAI-PMH does not insure a good SP
- OAI-PMH is *(or is becoming)* HTTP for DLs
  - few people get excited about http now
    - http & OAI-PMH are core technologies whose presence is now assumed



# Other Uses For the OAI-PMH

- Assumptions:
  - Traditional DLs / SPs will continue on their present path of increasing sophistication
    - citation indexing, search results viz, personalization, recommendations, subject-based filtering, etc.
  - growth rates remain the same (5x DPs as SPs)
- Premise: *OAI-PMH is applicable to any scenario that needs to update / synchronize distributed state*
  - Future opportunities are possible by creatively interpreting the OAI-PMH data model

# OAI-PMH Data Model



← resource

*set-membership is item-level property*

*item = identifier*

all available metadata about *David*

← item

Dublin Core metadata

MARC metadata

SPECTRUM metadata

← records

*record = identifier + metadata format + datestamp*

# Typical Values

- repository
  - collection of publications
- resource
  - scholarly publication
- item
  - all metadata (DC + MARC)
- record
  - a single metadata format
- datestamp
  - last update / addition of a record
- metadata format
  - bibliographic metadata format
- set
  - originating institution or subject categories

# Repositories...

- Stretching the idea of a repository a bit:
  - contextually sensitive repositories
    - “personalization for harvesters”
    - communication between strangers, or communication between friends?
  - OAI-PMH for individual complex objects?
    - OAI-PMH without MySQL?!
      - Fedora, Multi-valent documents, buckets
      - tar, jar, zip, etc. files

# Resource

- What if resource were:
  - computer system status
    - uptime, who, w, df, ps, etc.
  - or generalized "system" status
    - e.g., sports league standings
  - people
    - personnel databases
    - authority files for authors

# Item

- What if item were:
  - software
    - union of versions + formats
  - all forms of metadata
    - administrative + structural
    - citations, annotations, reviews, etc.
  - data
    - e.g., newsfeeds and other XML expressible content
      - metadataPrefixes or sets could be defined to be different versions



# Record

- What if record were:
  - specific software instantiations / updates
  - access / retrieval logs for DLs (or computer systems)
  - push / pull model inversion
    - put a harvester on the client behind a firewall, the client contacts a DP and receives "instructions" on how to submit the desired document (e.g., send email to a specified address)

# Datestamp

- semantics of datestamp are strongly influenced by the choice of resource / item / record / metadataPrefix, but it could be used to:
  - signify change of set membership (e.g., workflow: item moves from "submitted" to "approved")
  - change datestamp to reflect access to the DP
    - e.g., in conjunction with metadataPrefixes of "accessed" or "mirrored"

# metadataPrefix

- what if metadataPrefix were:
  - instructions for extracting / archiving / scraping the resource
    - verb=ListRecords&metadataPrefix=extract\_TIFFs
  - code fragments to run locally
    - (harvested from a trusted source!)
  - XSLT for other metadataPrefixes
    - branding container is at the repository-level, this could be record- or item-level

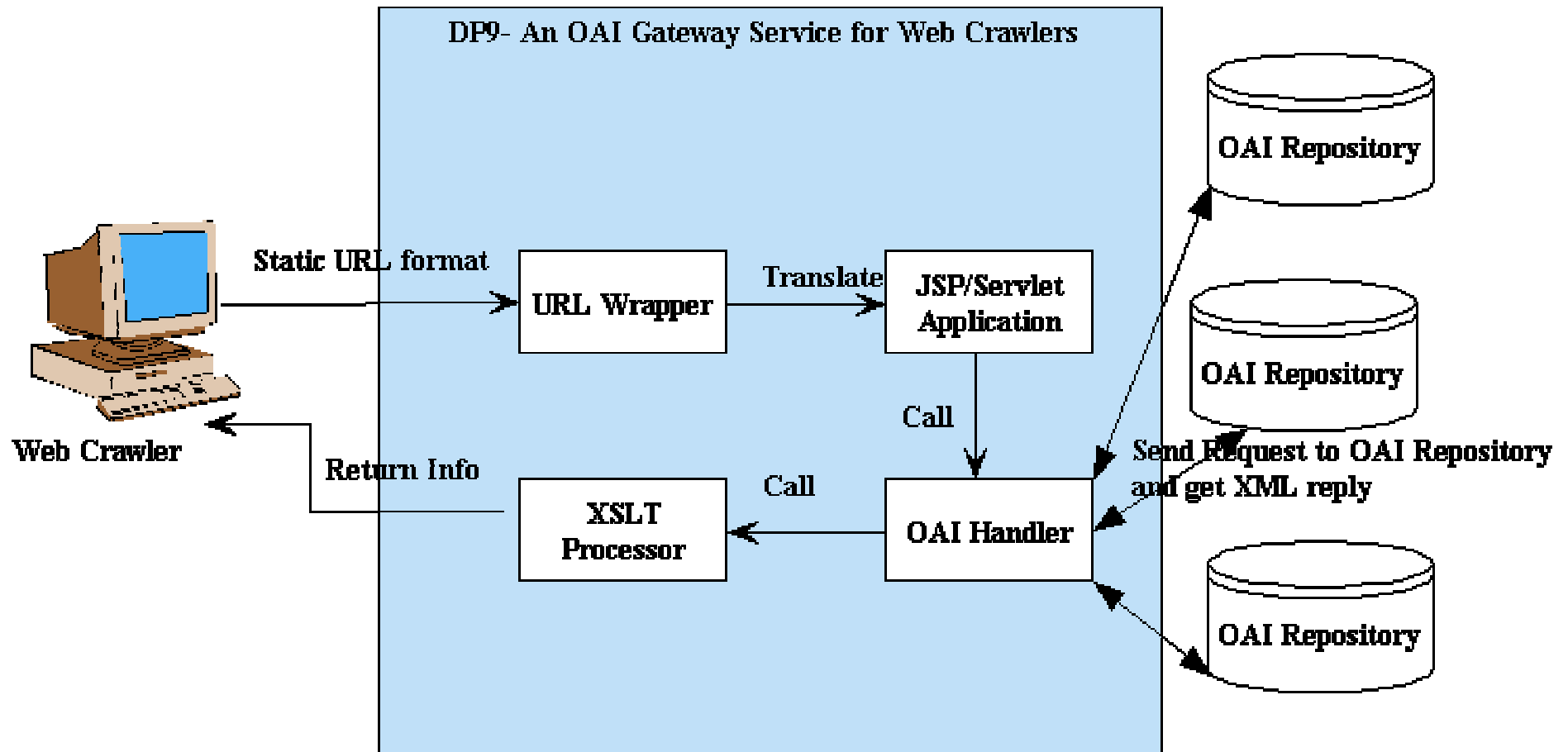
# Set

- sets are already used for tunneling OAI-PMH extensions (see Suleman & Fox, D-Lib 7(12))
- other uses:
  - in aggregators, automatically create 1 set per baseURL
  - have "hidden" sets (or metadataPrefix) that have administrative or community-specific values (or triggers)
    - set=accessed>1000&from=2001-01-01
    - set=harvestMeWithTheseARGS&until=2002-05-05&metadataPrefix=oai\_marc

# Interesting Services

- DP9
  - gateway to expose repository contents in HTML suitable for web crawlers
- Celestial
  - OAI "cache", also 1.1 -> 2.0 converter
- Static (mini-) repositories
  - XML files, based on OLAC work
- OpenURL metadata format registries
  - record = metadata format

# DP9 Architecture



see Liu et al., JCDL 2002; <http://dlib.cs.odu.edu/dp9>

# DP9 Formatting

- Format of URLs
  - `http://arc.cs.odu.edu:8080/dp9/getrecord.jsp?identifier=oai:NACA:1917:naca-report-10 &prefix=oai_dc`
  - `http://arc.cs.odu.edu:8080/dp9/getrecord/oai_dc/oai:NACA:1917:naca-report-10`
- HTML Meta tags
  - Some crawlers (such as Inktomi) use the HTML meta tags to index a Web pages; DP9 also maps Dublin Core metadata to corresponding HTML meta tags.
  - For pages that are designed exclusively for robots navigation, a noindex robots meta tag is used
  - X-FORWARDED-FOR header to distinguish between different users coming in via a proxy

# Celestial

- Developed by Brody @ Southampton
  - <http://celestial.eprints.org/>
  - designed to complement DP9
  - see Liu, Brody, et al., D-Lib Magazine 8(11)
- Where DP9 is a non-caching proxy, Celestial caches the metadata records
  - can off-load work from individual archives, higher availability
  - can harvest 1.1, 2.0; exports in 2.0



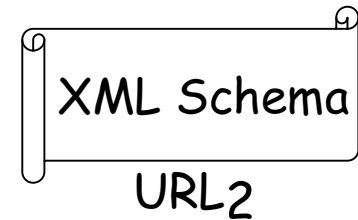
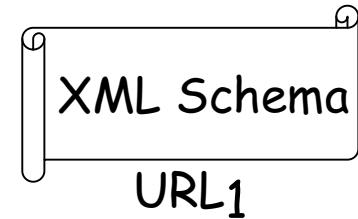
# "Static" Repositories

- Premise: a repository does not wish to have an executing program on its site, so it has a "static" XML file with some of the OAI-PMH responses in place
  - Design still being discussed
    - accessed through a proxy
    - could be a low functionality node, or the XML file could be produced by a process and moved outside a firewall
- Based on OLAC work by Bird & Simons
  - <http://www.language-archives.org/>

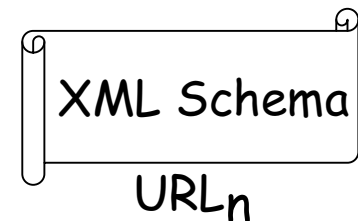
# OpenURL Metadata Registry

- Registry of metadata formats for OpenURL
  - <http://www.sfxit.com/openurl/>
  - <http://lib-www.lanl.gov/~herbertv/papers/icpp02-draft.pdf>

## registrars



•  
•  
•

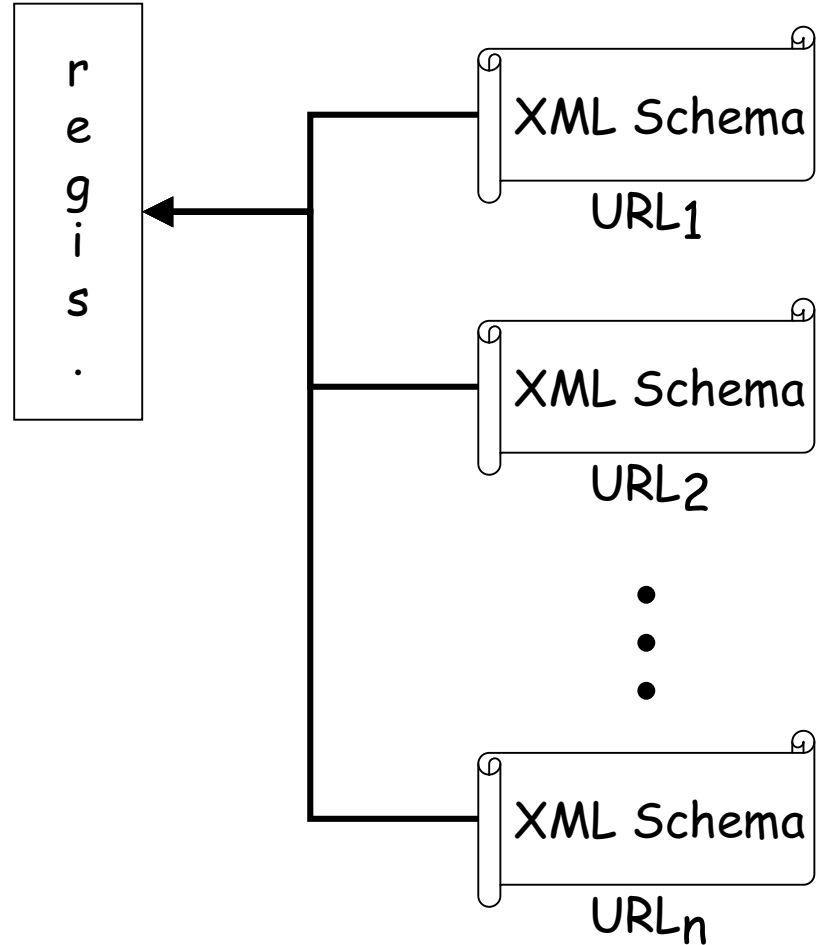


### Goal:

- inform linking servers re Schema
- ease of admin for all parties involved
- limit human overhead

central  
repository

registrars



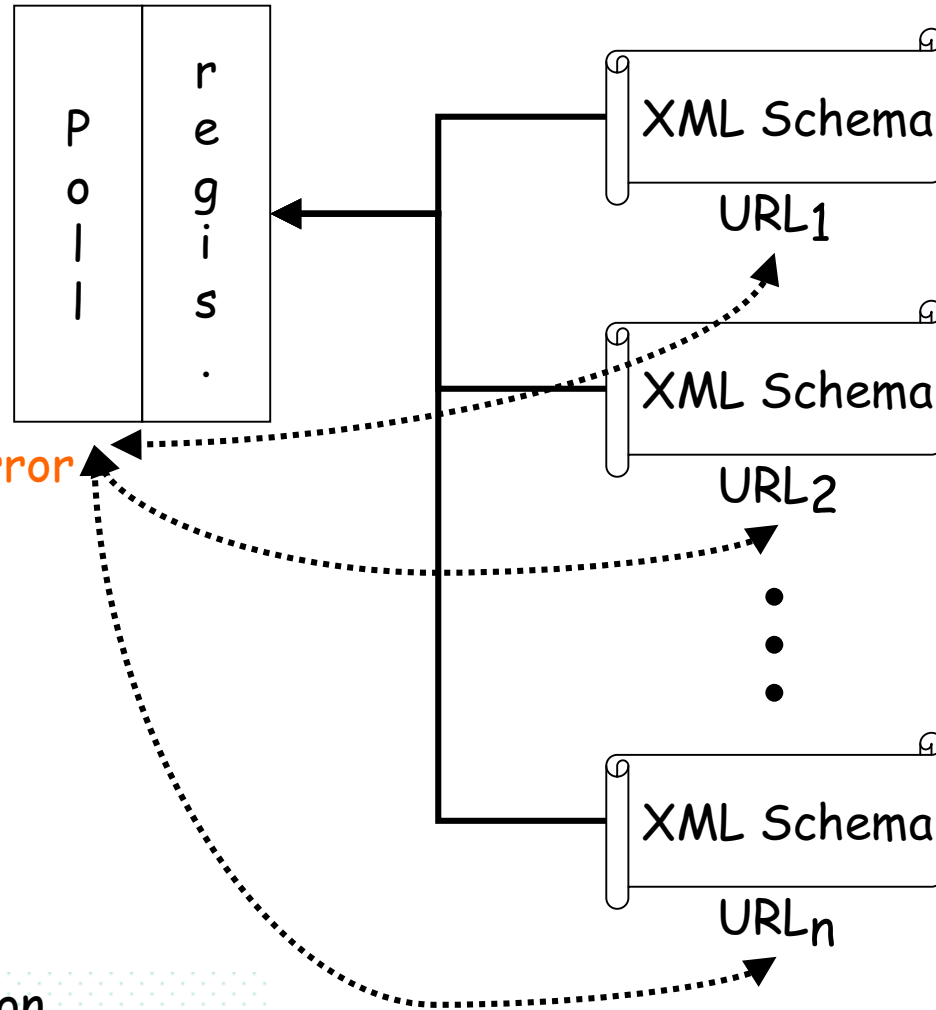
Registry:

- schemaLocation
- registration date
- mirror of Schema

← registration

central  
repository

registrars



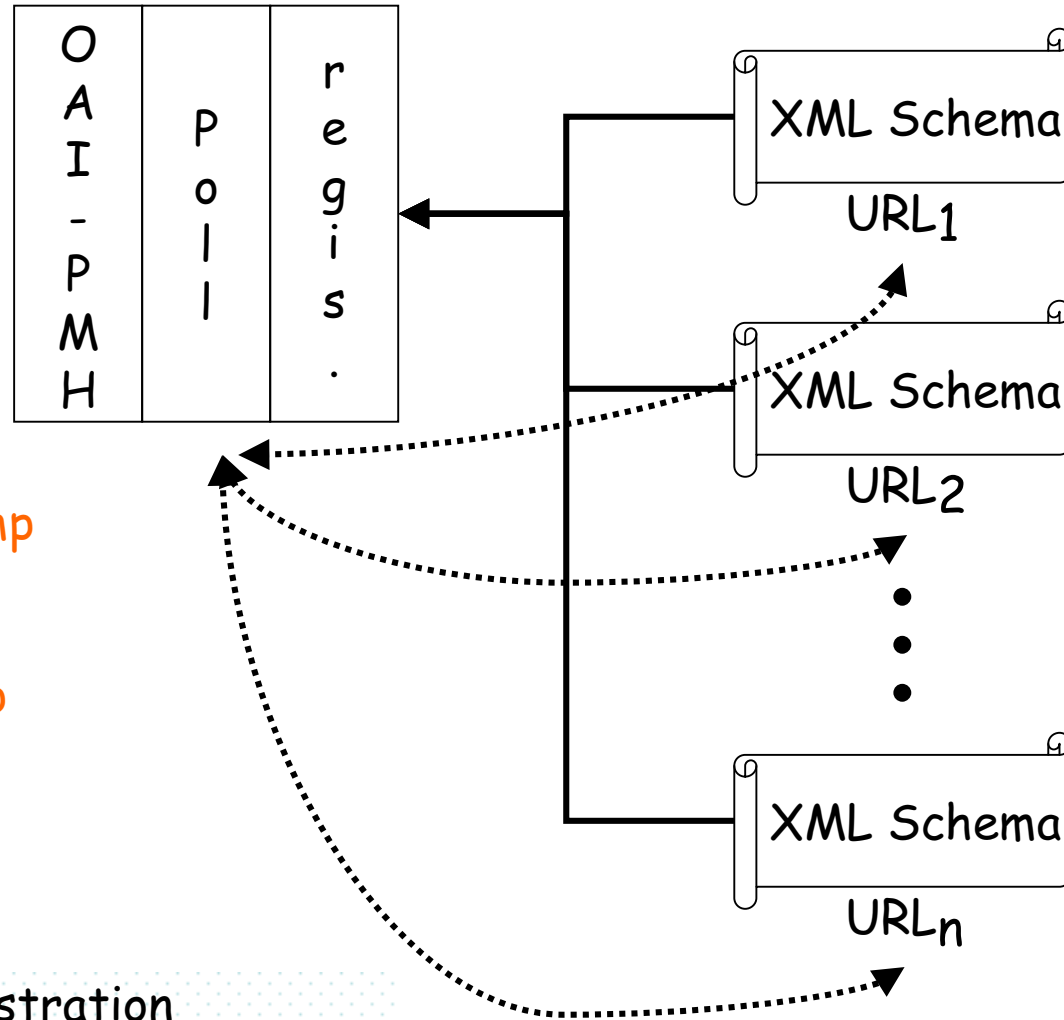
Poll:

- fetch schema at schemaLocation
- log failure/success
- compare fetched Schema with mirror
  - changed => replace mirror
  - removed => deregistered

← registration  
←·····→ polling

# central repository

# registrars

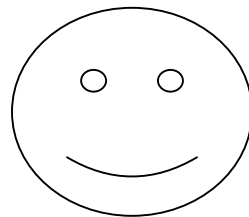
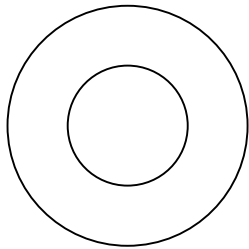
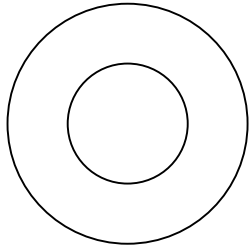


## OAI repo:

- record-ids = schemaLocation
- oai\_dc record :
  - registration info
  - (de)registration datestamp
- xsi record :
  - mirror schema
  - schema update datestamp
- poll record :
  - process info
  - recent poll datestamp

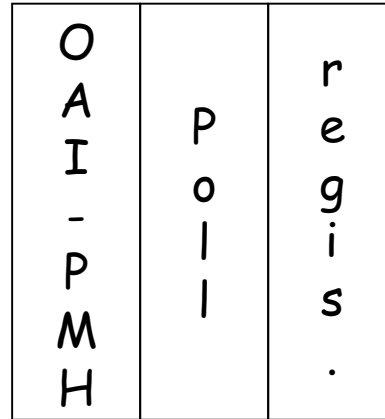
← registration  
←·····→ polling

linking  
servers

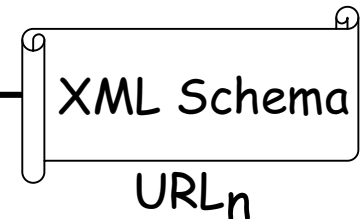
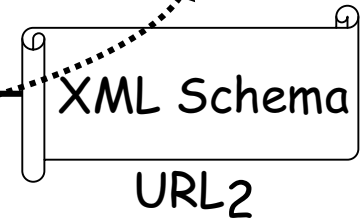
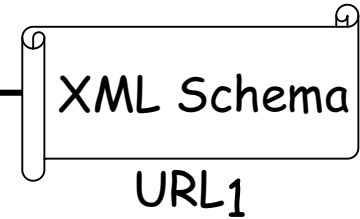


user  
service

central  
repository



registrars



# Conclusions

- DPs continue to proliferate
  - and spawn SPs!
- SPs are / are becoming a competitive market
  - e.g., at least 10 different interfaces to arXiv metadata
  - growing sophistication of services
  - differentiation of SPs will be on features that have little to nothing to do with OAI-PMH



# Conclusions

- Protocol / transport gateways
  - Dienst  $\leftrightarrow$  OAI
    - DOG, <http://www.cs.odu.edu/~tharriso/DOG/>
  - Z39.50
    - ZMARCO (UIUC)
  - SOAP
    - prototypes @ VT (Suleman) & ODU (Zubair)
  - WebDAV/DASL
    - resurrect DASL?

# OAI-PMH Will Have Arrived When:

- general web robots issue OAI-PMH verbs
  - ...DP9 will no longer be needed
  - requires shift in "control": harvester or repository?
- mod\_oai is developed and is included in the default Apache configuration
- OAI-PMH fades into the background
  - similar to TCP/IP, http, XML, etc.
  - next year's workshop is on OpenURL