

## An introduction to OAI-PMH

---

**Timothy W. Cole (t-cole3@illinois.edu)**

*Library Digital Content Access Lead*

*Head, Mathematics Library*

*Prof. of Library Administration*

*Prof. of Library & Info. Science*

***(with acknowledgements to MJ Han, Sarah Shreeves, Muriel Foulonneau, Simeon Warner)***

University of Illinois at Urbana-Champaign

28 September 2010



## Goals

- The **information landscape** can be seen as a contour map in which there are mountains, hillocks, valleys, plains and plateaus.... A specialized collection of particular importance is like a sharp peak. Upon a plateau there might be undulations representing strengths and weaknesses.... The landscape is, however, multidimensional. Where one scholar may see a peak another may see a trough. The task is to devise mapping conventions which enable scholars to read the map of the landscape fruitfully, at the appropriate level of generality or specificity.

*Michael Heaney (2000), "[An Analytical Model of Collections and their Catalogues.](#)"*





# OAI Protocol for Metadata Harvesting

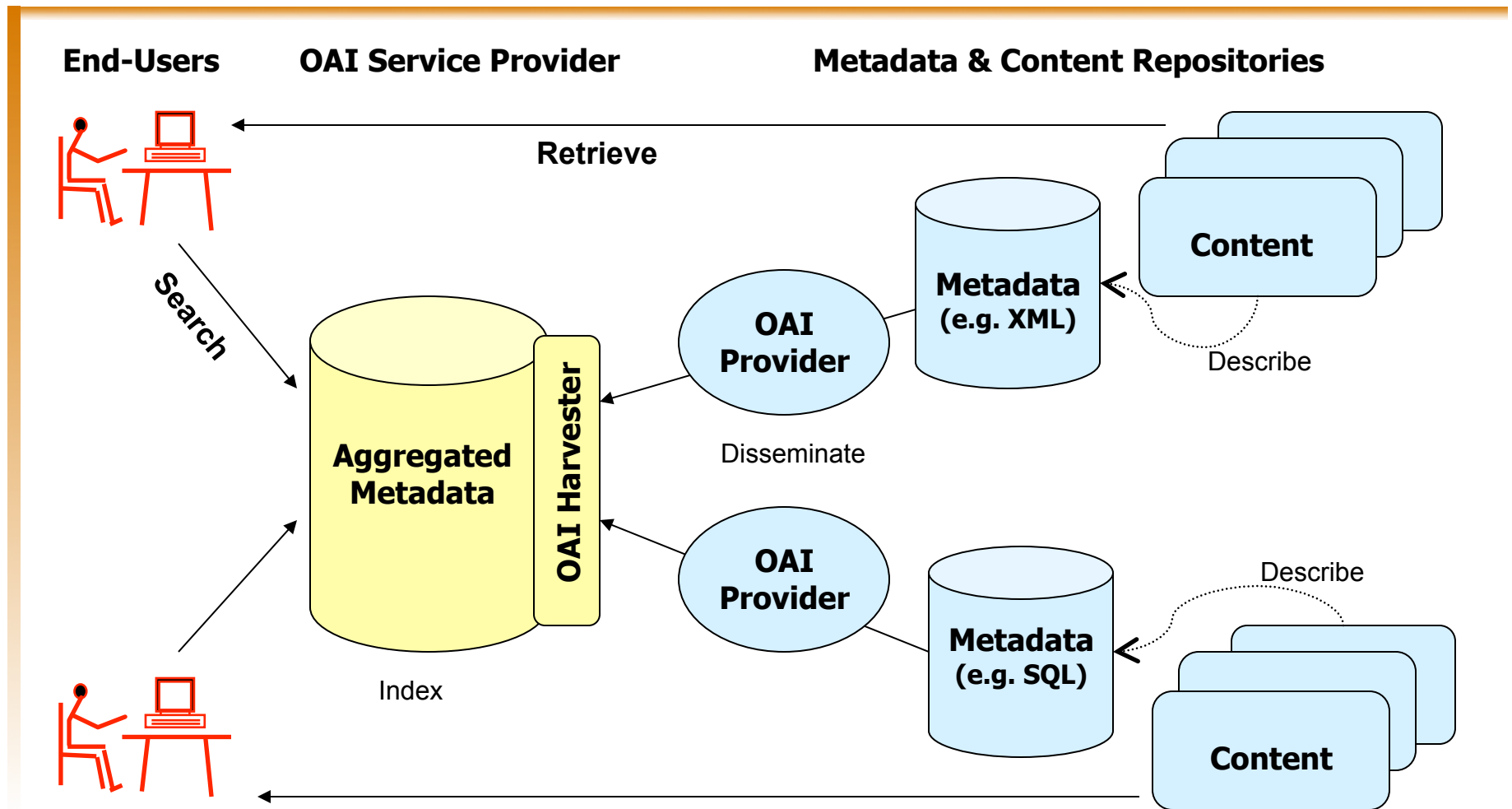
- 'Harvesting' approach to interoperability at metadata level
- Divides world into Metadata Providers & Service Providers
- Builds on HTTP, XML, & community metadata standards

<http://www.openarchives.org/>

The screenshot shows the Open Archives Initiative website. The browser title is "Open Archives Initiative - Microsoft Internet Explorer". The address bar shows "http://www.openarchives.org/". The website has a navigation menu with links for Home, Documents, Tools, Community, News, and Organization. The main content area is titled "News from the OAI Community" and lists several news items, including "Static Respository Specification", "mod\_oai Project Aims at Optimizing Web Crawling", "OLAC Archive on board European Space Agency mission", and "U-M expands access to hidden electronic resources with OAIster". On the right side, there are sections for "Read Core Documents" (with links to Harvesting Protocol, Migration Instructions, Implementation Guidelines, and FAQ), "Join the OAI Community" (with buttons for "Join OAI General list", "Join OAI Implementers list", "Register as a data provider", and "Submit a post to our web site"), and "Contact us: openarchives@openarchives.org". At the bottom, there is a footer with support information and logos for Digital Library Federation, CNI, and another organization.



# Metadata Harvesting Model



## Metadata Harvesting Model (cont.)

- OAI Service Provider (harvester) is middleman between content provider and end-user for selected metadata-based transactions – e.g.,
  - Resource discovery
  - Value-added link mediation
  - ...
- Transactions involving full content still conducted directly between end-users and content provider – e.g.,
  - Delivery of complete document in desired format
- OAI-PMH is not synonymous with Open Access



## OAI- PMH is a tool

- The protocol provides the technical framework and a set of detailed rules telling implementers how to disseminate and share metadata records, irrespective of the intellectual content of those records.
- All about moving metadata (not data) around
- Assumes widely distributed content, but centralized indexing & services
- Build once, use for many applications – a building block for metadata-mediated digital library services

The purpose of OAI-PMH is to share metadata and foster interoperability





## OAI-PMH itself is not....

- A metadata scheme
- A search tool
- A database





## History of OAI

- Originated in the e-print archive community
  - Creation of interoperability tools for between archives of e-prints
- Santa Fe Meetings - 1999 and 2000
  - Convened by Paul Ginsparg, Rick Luce, & Herbert Van de Sompel
- OAI – PMH version history:
  - First Alpha Release, Sept. 2000
  - 1.0 (Beta) Release January 2001
  - 1.1 (Beta 2) Release July 2001
  - 2.0 (Production) Release June 2002





## Organization of the OAI (historical)

- OAI Executive:
  - Carl Lagoze, Cornell University
  - Herbert Van de Sompel, Los Alamos National Laboratory
  
- OAI Steering Committee – funders & stakeholders, e.g.:
  - Mellon Foundation, DLF, CNI, Library of Congress, ...
  
- OAI Technical Committee – early adopters & experimenters



## Harvested Search vs. Distributed Search

- Distributed/Broadcast searching: search and discovery by simultaneously querying remote services and data
- Harvesting is when data/metadata is transferred from the remote source to the destination where the services are located (e.g. Union catalogs)

*Competing – but not incompatible – approaches to interoperability*



## Some of the underlying concepts

- OAI Data providers – support OAI PMH as a means to expose and disseminate metadata
- OAI Service providers – ‘harvest’ metadata from data providers via the OAI-PMH
- OAI-PMH relies on HTTP, XML, Community metadata stds.
  - Requires use of W3C XML Schema Language
- OAI-PMH requires use of simple Dublin Core
  - BUT supports and encourages use of other metadata schemas
- RSS and ATOM are newer alternatives to OAI-PMH (broader adoption but more limited metadata implementation options)



## Reliance on HTTP & XML

- OAI-PMH is a REpresentational State Transfer (REST) protocol (unlike RPC, SOAP)
  - OAI requests and responses are sent via the HTTP protocol
  - OAI requests encoded as HTTP GET or POST operations
  
- OAI responses are valid XML documents
  - Consistency and data “quality” is ensured by using XML Schema Definitions (XSD) for all responses
  - XML Namespaces used to identify which parts of response are metadata and which parts support the Protocol



# Terminology

| <b>Term</b>                  | <b>Meaning in context of OAI-PMH</b>   |
|------------------------------|--|
| <b>Resource</b>              | The information object of interest; the "stuff" described by a <i>metadata item</i> .  |
| <b>Metadata item</b>         | All the metadata held in the repository about a given <i>resource</i> ; may be a virtual construct only.   |
| <b>OAI identifier</b>        | The persistent identifier by which the <i>metadata item</i> may be referenced. The <i>OAI identifier</i> must be unique at least within the repository; properly constructed it is globally unique.  |
| <b>Metadata record</b>       | A dissemination of the <i>metadata item</i> in a particular metadata format and at a particular point in time; a <i>metadata record</i> is uniquely defined by its <i>identifier</i> , <i>metadata prefix</i> and <i>datestamp</i> .   |
| <b>metadataPrefix</b>        | A label for the format of a <i>metadata record</i> ; the "oai_dc" <i>metadata prefix</i> must be supported by all repositories, all other <i>metadata prefixes</i> should be assumed repository-specific   |
| <b>OAI datestamp</b>         | The UTC date and optionally time-of-day when a <i>metadata record</i> was last modified; this value changes when the underlying <i>metadata item</i> changes and also when the mapping of the <i>metadata item</i> to <i>metadata record</i> in the metadata format of the <i>metadata record</i> changes. |
| <b>datestamp granularity</b> | There are two allowed granularities for <i>OAI datestamps</i> : YYYY-MM-DD (i.e., day granularity) and YYYY-MM-DDThh:mm:ssZ (seconds granularity)  |

## Protocol Details

- An OAI Transaction consists of:  
An OAI (HTTP) request & a corresponding OAI response (XML)
  - Transactions initiated by harvester
  - Optional flow control mechanisms to manage provider load
- OAI Item Identifiers – persistent & unique
- Item (Metadata) Date Stamps – support selective harvesting
- OAI supports multiple metadata formats
  - Distinguishes between an ITEM (complete metadata) & a RECORD (disseminated item of metadata in given format)



# How OAI-PMH Works

## ■ OAI "VERBS"

**Identify**

**ListMetadataFormats**

**ListSets**

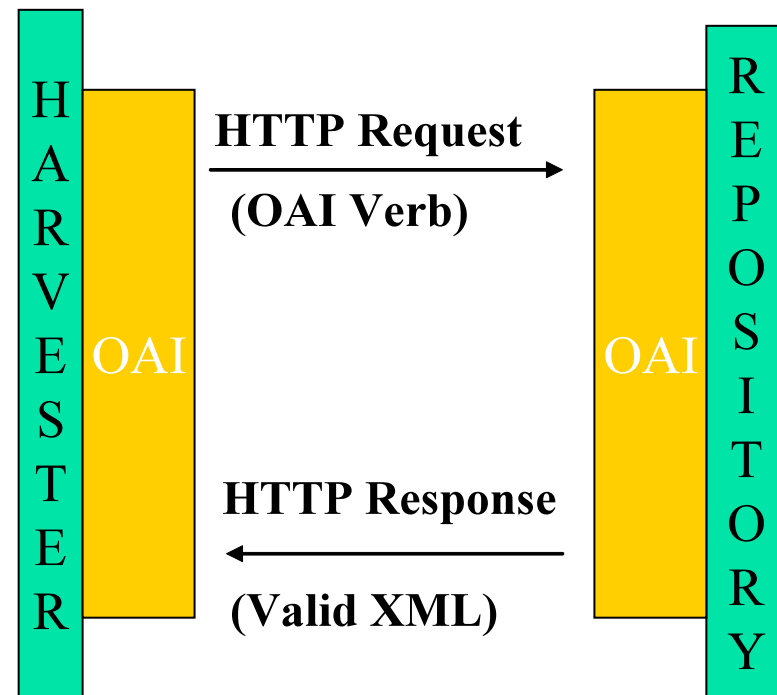
**ListIdentifiers**

**ListRecords**

**GetRecord**

Service Provider

Metadata Provider



# Identify

- Purpose
  - Return general information about the archive and its policies (e.g., datestamp granularity)
- Parameters
  - None
- Sample URL

<http://aerialphotos.grainger.uiuc.edu/oai.asp?verb=Identify>





# ListSets

- Purpose
  - Provide a listing of sets in which records may be organized (may be hierarchical, overlapping, or flat)
- Parameters
  - None
- Sample URL:

<http://aerialphotos.grainger.uiuc.edu/oai.asp?verb=ListSets>



## ListMetadataFormats

- Purpose
  - List metadata formats supported by the archive as well as their schema locations and namespaces
- Parameters
  - identifier – for a specific record (O)
- Sample URL

<http://aerialphotos.grainger.uiuc.edu/oai.asp?verb=ListMetadataFormats>



# ListIdentifiers

- Purpose
  - List headers for all items corresponding to the specified parameters
- Parameters
  - from – start date (O)
  - until – end date (O)
  - set – set to harvest from (O)
  - metadataPrefix – metadata format to list identifiers for (R)
  - resumptionToken – flow control mechanism (X)
- Sample URL (slow, don't try to open during class)

[http://aerialphotos.grainger.uiuc.edu/oai.asp?verb=ListIdentifiers&metadataPrefix=oai\\_dc](http://aerialphotos.grainger.uiuc.edu/oai.asp?verb=ListIdentifiers&metadataPrefix=oai_dc)



# GetRecord

- Purpose
  - Returns the metadata for a single item in the form of an OAI record
- Parameters
  - identifier – unique id for item (R)
  - metadataPrefix – metadata format for the record (R)
- Sample URL

[http://aerialphotos.grainger.uiuc.edu/oai.asp?  
verb=GetRecord&identifier=oai:aerialphotos.grainger.uiuc.edu:AP-1A-1-1940&metadataPrefix=oai\\_dc](http://aerialphotos.grainger.uiuc.edu/oai.asp?verb=GetRecord&identifier=oai:aerialphotos.grainger.uiuc.edu:AP-1A-1-1940&metadataPrefix=oai_dc)



# ListRecords

- Purpose
  - Retrieves metadata records for multiple items
- Parameters
  - from – start date (O)
  - until – end date (O)
  - set – set to harvest from (O)
  - resumptionToken – flow control mechanism (X)
  - metadataPrefix – metadata format (R)
- Sample URL (slow, don't try to open during class)

[http://aerialphotos.grainger.uiuc.edu/oai.asp?verb=ListRecords&metadataPrefix=oai\\_dc](http://aerialphotos.grainger.uiuc.edu/oai.asp?verb=ListRecords&metadataPrefix=oai_dc)



## Unique Identifiers

- Each OAI item must have a unique identifier
- Identifiers must follow rules for valid URIs
- Example:
  - `oai:<archiveId>:<recordId>`
  - `oai:etd.vt.edu:etd-1234567890`
- Each identifier must resolve to a “single” item and always to the same item
  - Can't reuse OAI item identifiers



## Datestamps

- Needed for every OAI record to support incremental harvesting
- Must be updated when addition or modification or deletion made in order to ensure changes are correctly propagated to harvesters
- Different from dates within the metadata – OAI datestamp is used only for harvesting
- Can be either YYYY-MM-DD or YYYY-MM-DDThh:mm:ssZ (must be GMT timezone)



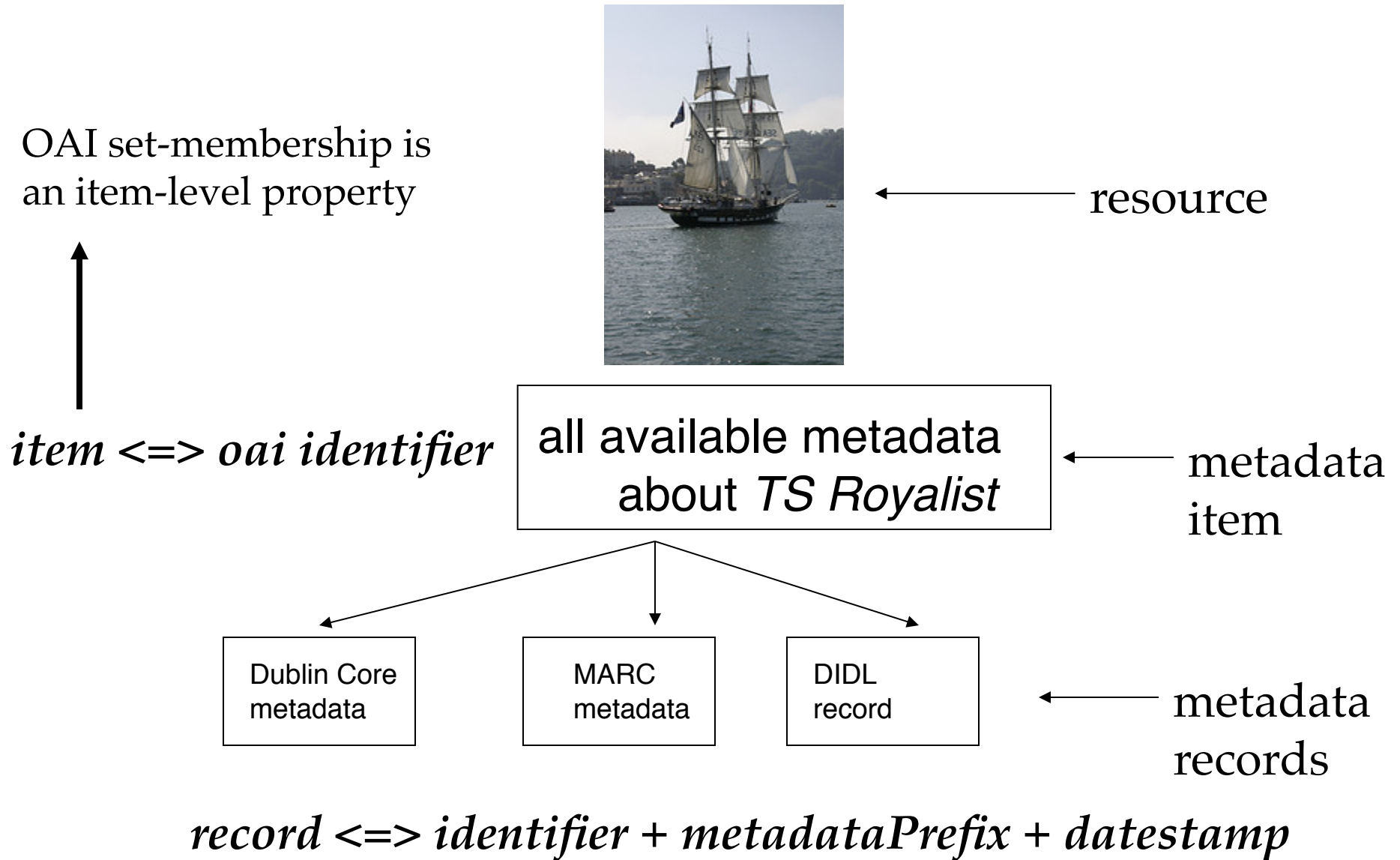
## OAI Items vs. OAI Records

- An OAI ITEM is the complete set of metadata you possess describing an object in your repository
  - Items exist only in OAI Metadata Provider database
  
- An OAI RECORD is an OAI Item disseminated in a particular metadata format – e.g., DC or MARC
  - Records are what get harvested by OAI Service Providers
  
- OAI IDENTIFIERS are Item-Level
- OAI DATESTAMPS are Record-Level





# OAI data model: resource-item-record(s)



## Optional Features

| <b>Optional OAI-PMH Feature</b>                            | <b>Reason to implement or select a turnkey solution that supports the feature</b>   |
|--|---|
| <b>Multiple metadata formats</b>                           | To expose additional, richer or more community-specific metadata than can be exposed using simple Dublin Core metadata format alone.  |
| <b>Description, setDescription and/or about containers</b> | To provide service provider with added, useful information about the data provider repository, about set organization of the repository, and/or about individual metadata records.                    |
| <b>OAI sets</b>  | So that a service provider can choose to harvest only a portion of the metadata items contained in the full repository; especially desirable for larger repositories containing multiple collections. |
| <b>resumptionTokens</b>                                    | To facilitate handling of OAI responses involving large numbers of identifiers or metadata records, i.e., more than a few thousand.   |
| <b>Persistent deleted records policy</b>                   | To allow harvester to better know when a record has been removed; this facilitates incremental harvesting and obviates need for frequent full harvests.   |
| <b>datestamp granularity of second</b>                     | To allow a harvester to perform more precise incremental harvesting based on more granular datestamp; essential when used that implementation is correct.   |

## What it takes to implement OAI

- Dynamic Web server functionality (e.g., CGI)
- Capacity to respond with XML
- Descriptive metadata in standard format(s)
  - OAI persistent identifiers & date stamps may require changes to metadata creation workflow
- Open source implementations available (starting points)
- OAI-PMH included in turnkey publishing solutions:
  - Public Knowledge Project (UBC)
  - Open Repository (BioMed Central), ...
  - Eprints.org, DSpace, CONTENTdm, ARNO, CDSware, ...



## Provider Performance Issues

- Database design biggest impact on performance
  - e.g., load to dynamically map to DC, other formats
  - Webserver performance load can be kept quite low
  
- Use resumptionTokens, other flow control mechanisms to improve performance
  - Fetch only records needed to satisfy current request
  - resumptionTokens should retain state information for best performance and for idempotency

Scale example: OCLC repository with 4+ million records



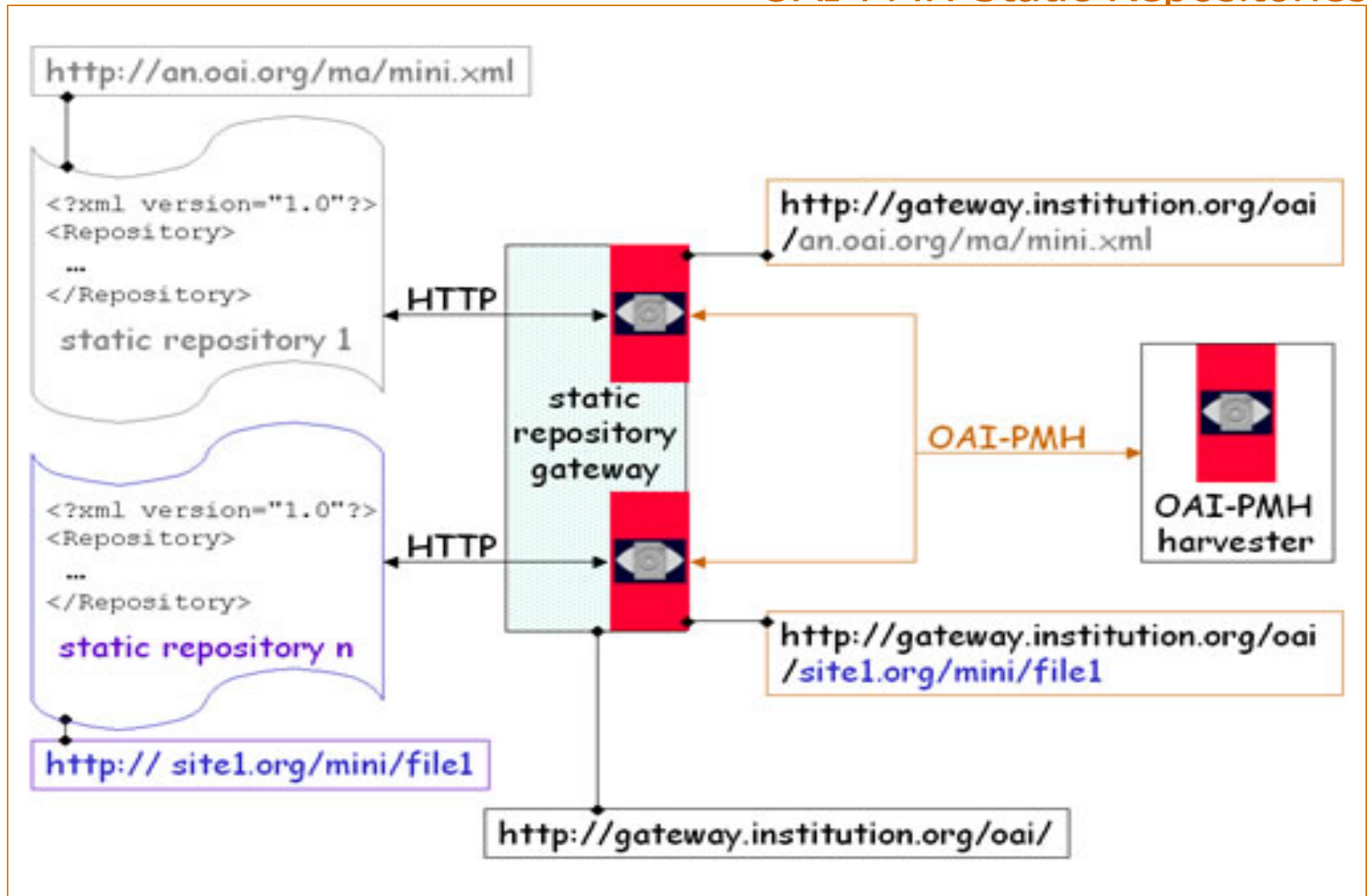
## OAI Implementation Guidelines for Repositories

- Tools Required
- Basic program strategies (incl. object-oriented approaches)
- Guidance for use of
  - optional container elements
  - Metadata generation / mapping, data cleaning
  - Use of OAI Sets
  - resumptionToken, flow control, load-balancing
  - Denial-of-service prevention
  - Error handling
- Strategies for deleted metadata records

<http://www.openarchives.org/OAI/2.0/guidelines-repository.htm>



## OAI-PMH Static Repositories



From: <http://www.openarchives.org/OAI/2.0/guidelines-static-repository.htm>

## Who's Using OAI?

OAI Data Provider Registry (<http://oai.grainger.uiuc.edu/registry>)

- As of Sept 2010: 2,300+ active OAI metadata provider repositories
  - Range in size from millions of items, to less than 100 items
  - More than half are institutional repositories or eprint archives
- Handful of publisher / publisher-aggregators, e.g.:
  - PubMed Central; BioOne; BioMed Central (partial); Project Euclid; Africa Journals Online; Institute of Physics (user id & password); American Physical Society (restricted access); ...
- Handful of individual journals, e.g.:
  - J. of STEM Education; Electronic J. of Probability; J. of Cognitive Affective Learning; Canadian J. of Communication; ...



## Who's Harvesting Metadata Using OAI-PMH

- Portals, including many encouraging Open Access, e.g.:
  - WorldCat; Public Knowledge Project; Citebase; Cyclades; ...
  - NSDL (STEM Education); NCSTRL (computer science); SAIL (physical science e-prints); ...
- Local harvesting projects
  - As way to share data internally (e.g., UIUC e-book records)
  - As a collation service to their users – e.g., Grainger Search Service; OAI harvesting supported by some Library meta-search utilities such as Illinois Harvest
- Research Projects
  - IMLS Digital Collections & Content
  - DLF Aquifer (American History Online)

Not for end-user, not particularly well-suited to Web 2.0, Linked-data, ...





## So, you want to use OAI-PMH?

- Enhance visibility of your resources to other portals
  - Enables reuse of digital content in new contexts
- Support interoperability across collections
  - Enables repurposing digital content to support new services
- Facilitate discovery of new linkages and relationships between and among content objects, metadata, and services
  - Ultimately supports a “recombinant” vision of digital libraries

### **A step toward enhanced models of digital libraries, e.g.,**

Lorcan Dempsey, et al. “Metadata Switch: thinking about some metadata management and knowledge organization issues ....”

<http://www.oclc.org/research/publications/archive/2004/dempsey-mslitaguide.pdf>



## Why Share?

- Sharing benefits your institution
  - Increases exposure of collections
  - Broadens user base
  - Potentially adds collaboration opportunities
  
- We can no longer assume that users will come through the front door, sharing metadata gets us “in the flow” (Lorcan Dempsey)



# Challenges

- Metadata providers (OAI Data Providers)
  - Technical requirements
  - Political / organizational requirements
  - Resource requirements
  
- Aggregators (OAI Service Providers)
  - Communicating scope to user – virtual buildings
  - Diversity of content and standards
  - Lack of documentation
  - Technical challenges



## To consider when implementing OAI-PMH

- Audience/Purpose: Who are you describing it for? Why?
- Standards: Are there standards you can use?
- What: Are you describing the digital manifestation of a work? A physical object that has been digitized? Both?
- Granularity: What level of description? Item level? Archival unit? Collection level? (FRBR issues here as well)
- Context: In what setting do you anticipate metadata will be used?



## Shareable Metadata... (more next week from Sarah)

- Is *quality* metadata
- Promotes search interoperability - “the ability to perform a search over diverse sets of metadata records and obtain meaningful results” (Priscilla Caplan)
- Is human understandable outside of local context
- Is *useful* outside of local context
- Preferably is machine processable



## Implementing OAI-PMH – Finding the right balance

- Metadata providers know the materials
  - Be consistent & document practices
  - Document encoding schemes & controlled vocabulary use
  - Ensure record validity
  
- Aggregators have scale, design the portal
  - Format conversion
  - Reconcile known vocabularies
  - Normalize data
  - Batch metadata enhancement



## OAI-PMH depends on shared work model

OAI-PMH / metadata sharing depends on a collaboration between data providers and service providers

- Data providers:
  - Conform to OAI-PMH (incl. [Implementation Guidelines](#))
  - Provide shareable metadata of appropriate richness
  - Provide access to resources described by metadata
- Service providers:
  - Conform to OAI-PMH (incl. [Implementation Guidelines](#))
  - Filter, normalize, & enrich metadata for purpose
  - [Present](#) aggregated metadata without bias
  - Do no harm



## Data providers need to help service provider

- Provide documentation on choices made when providing metadata for exposure via OAI – remember, metadata itself has provenance
  - Provide human-readable documentation on your Website
  - Utilize OAI optional containers to help service providers
- Document especially:
  - Use of terminologies / controlled vocabularies
  - Source of metadata (was it transformed from different format?)
  - Value encoding practices for: names, dates, identifiers, ...
  - Local practices for quality control, updating frequencies, ...
- Most service providers have some capacity to normalize harvested metadata on a data provider-by-data provider basis





## Appropriate representation of resources

- OAI-PMH record provides only one view of a resource
  - Records shared via OAI should be appropriate for purpose
- Consider:
  - Contents: what to include in OAI metadata record
  - Context: make explicit that which might be implicit in your local system; OAI metadata records must stand alone
  - Metadata format:
    - **provide multiple** if possible
    - select metadata schema(s) appropriate to content
    - retain richness of native scheme as much as possible





## DLF-NSDL OAI Best Practices

[http://webservices.itcs.umich.edu/mediawiki/oaibp/?OAI\\_Best\\_Practices](http://webservices.itcs.umich.edu/mediawiki/oaibp/?OAI_Best_Practices)

- Collaborative effort of the Digital Library Federation and the National Science Digital Library (U.S.) – data providers & service providers
  - Scope: Develop technical implementation & shareable metadata best practices for OAI data providers
  - Targeted for broad audience, but group drawn mostly from U.S. library & digital library communities
- July 2004 – First meeting of Working Group
- August 2005 – First review draft
- Summer 2007 – Completed work
- Intended to complement existing resources: OAI listservs, OAForum, community specific guidelines & best practices



## Common problems (1)

- Analyzed validation 2004 logs for validator:  
<http://www.openarchives.org/Register/ValidateSite>  
(paper [arXiv:cs.DL/0506010](http://arxiv.org/abs/cs.DL/0506010) describes in more detail)
- 1893 requests with sensible baseURL
- 18% no Identify response
- 21% of cases returned invalid XML (Xerces output)
- 7% bad `adminEmail`, 0.3% bad protocol version
- 24% other errors with Identify -- usually quickly fixed
- 1% excessive (>5 in a row) 503 Retry-after
- 3% no identifiers from ListIdentifiers
- 2.5% no datestamp in sample record - fundamental problem!



## Common problems (2)

- 927 completed validation requests
- 34% successful
- 22% errors in handling exception conditions
- 44% other (more serious) errors

Most common errors:

1. XML problems – Failed schema validation, incorrect character encoding
2. Empty response with known good `from` and `until`
3. Empty `resumptionToken` to request without `resumptionToken`
4. Malformed response if identifier is `invalid" id`
5. Granularity of `earliestDatestamp` doesn't match `granularity` value



## XML Schema / Namespace problems

- OAI-PMH response must specify the correct namespaces and schemaLocations for the OAI-PMH schema and the oai\_dc schema, e.g.

```
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"  
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"  
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/  
    http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
```

and

```
<oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"  
  xmlns:dc="http://purl.org/dc/elements/1.1/"  
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"  
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/  
    http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
```

(Hint: just copy from spec.)

- Use standard namespaces and schemas for other formats where possible



## Tricky datestamp and timezone

- One useful test is to check that a given header/record is returned when the `from` and `until` dates of a `ListIdentifiers/ListRecords` are set to its `datestamp`.
- Second most “popular” error after parsing failures.
- Usually quickly corrected.
- One as yet unsolved case with a DSpace instance in Australia, operating in a timezone with a half-hour offset from UTC/GMT. The `from` and `until` must be set half a hour off to get the correct record, clearly broken!



## identifier="invalid" id

- The most common responses to this input condition are:
  1. invalid XML returned
  2. 500 server error
  
- Particularly troubling as these case imply
  1. lack of systematic parameter checking (should have checks at least as strict as OAI spec, perhaps more so to limit to local context)
  2. lack of systematic output encoding (plain " can't go in an XML attribute even if one mistakenly wants to include it, use &quot; instead)
  
- Such failures are asking for trouble!



## More subtle errors -- hidden updates

- OAI-PMH is designed to allow incremental harvesting
- Record timestamps change when metadata item change affects record & when item --> record mapping changes
- Updates must be available by the end of the period of the timestamp assigned, i.e.
  - Day granularity => during same day
  - Seconds granularity => during same second
  - Reason: harvesters need to overlap requests by just one timestamp interval (one day or one second)





## More subtle errors -- resumptionToken & idempotency

- idempotency of List requests: return same incomplete list when resumptionToken is re-issued
  - while no changes occur in the repository: strict
  - while changes occur in the repository: all items with unchanged datestamp
- Means that harvester can recover from a bad transmission by repeating request at any point in a long response sequence
- IMPLICATION: data-provider must accept both the most recent resumptionToken issued and the previous one



## Frustrating errors – invalid resource URLs

- Harvest is correct & metadata looks correct
- But when try to use metadata-embedded URL to direct users to resource, URL turns out to be broken link





## Identifiers, URLs, & linking

- Data providers should use persistent URIs & recognized standard identifiers (e.g., ISBNs, ISSNs, DOIs, PURLs, EPICUR, OpenURLs, ...)
  - **Provide one unambiguous URI for primary user access to resource**
  - If schema allows, encode the nature of each identifier provided
  - For analog resources, still provide URL for how to access item
  - Other links: rights statement; access restrictions; collection or institution home page; curator contacts; alternate versions; ...
  - Don't confuse resource identifiers & OAI identifiers ([Guidelines](#))
  - Express multiple identifiers in repeated fields; but only include multiple where meaning/function of each identifier is clear – consensus regarding “actionable” resource URLs still evolving



"We'll have lots to eat this winter, won't we Mother?"



**Grow your own  
Can your own**

Illustration of metadata  
for purpose

[http://  
images.library.illinois.edu:  
8081/u?/tdc,107](http://images.library.illinois.edu:8081/u?/tdc,107)

```
http://images.library.uiuc.edu:8081/cgi-bin/oai.exe?verb=GetRecord&metadataPrefix=oai_dc&identi - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://images.library.uiuc.edu:8081/cgi-bin/oai.exe?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:images.library.uiuc.edu:tdc/107 Go Links

<?xml version="1.0" encoding="UTF-8" ?>
- <OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2005-10-13T16:29:02Z</responseDate>
  <request verb="GetRecord" metadataPrefix="oai_dc"
    identifier="oai:images.library.uiuc.edu:tdc/107">http://images.library.uiuc.edu:8081/cgi-bin/oai.exe</request>
  - <GetRecord>
  - <record>
  - <header>
    - <header>
      <identifier>oai:images.library.uiuc.edu:tdc/107</identifier>
      <timestamp>2003-09-18</timestamp>
      <setSpec>tdc</setSpec>
    </header>
  - <metadata>
    - <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:dc="http://purl.org/dc/elements/1.1/"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
      <dc:format>ww20015p.jpg</dc:format>
      <dc:type>image</dc:type>
      <dc:title>We'll Have Lots to Eat this Winter, Won't We Mother?; Grow Your Own, Can Your Own</dc:title>
      <dc:title />
      <dc:coverage>1943</dc:coverage>
      <dc:description>Poster, color, 16 x 22.6 in., published by the United States Government Printing Office</dc:description>
      <dc:description>Many foods, including coffee, tea, butter, meat, frozen and canned vegetables were rationed during World
        War II. Americans were encouraged to plant "victory gardens" to help provide food for their families and neighbors.
        Women often preserved the excess produce from these gardens through home canning. Canning is a food preservation
        system that involves precooking and then air-tight sealing of food in jars, which are then immersed in a hot water "bath"
        for a specified period of time. This hot water "bath" is intended to kill off contaminants that may have survived the
        processing. If the contents of the jars were very acidic, such as tomatoes, there was less danger of the canned food
        spoiling. Other, less-acidic, home-canned foods were protected from spoiling by the use of brines, sugar, or salt as
        preservatives and for flavor.</dc:description>
      <dc:description>World War II;</dc:description>
      <dc:description>14 Political systems; 16 History;</dc:description>
      <dc:creator>United States. Office of War Information</dc:creator>
      <dc:contributor>Parker, Albert [artist]</dc:contributor>
      <dc:source />
      <dc:publisher>Illinois State Library</dc:publisher>
      <dc:language>English</dc:language>
      <dc:subject>Canning and preserving--United States ; Food conservation--United States ; World War, 1939-1945--Food
        supply--United States ; War posters, American ; World War II; Winter</dc:subject>
      <dc:rights>http://images.library.uiuc.edu/projects/tdc/conditions.htm</dc:rights>
      <dc:identifier>ww20015p</dc:identifier>
      <dc:date>11-20-01</dc:date>
      <dc:relation />
      <dc:identifier>http://images.library.uiuc.edu:8081/u?/tdc,107</dc:identifier>
    </oai_dc:dc>
  </metadata>
</record>
</GetRecord>
```

[http://images.library.illinois.edu:8081/cgi-bin/oai.exe?verb=GetRecord&metadataPrefix=oai\\_dc&identifier=oai:images.library.illinois.edu:tdc/107](http://images.library.illinois.edu:8081/cgi-bin/oai.exe?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:images.library.illinois.edu:tdc/107)

```
http://snuffy.lib.umn.edu:8080/image/oai/HandleRequest.do?verb=GetRecord&metadataPrefix=oai_dc& - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://snuffy.lib.umn.edu:8080/image/oai/HandleRequest.do?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:digital.lib.umn.edu:mpw00250
Go Links

<?xml version="1.0" encoding="UTF-8" ?>
- <OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2005-10-13T16:37:18Z</responseDate>
  <request identifier="mpw00250" metadataPrefix="oai_dc"
    verb="GetRecord">http://snuffy.lib.umn.edu:8080/image/oai/HandleRequest.do</request>
- <GetRecord>
- <record>
- <header>
  <identifier>oai:digital.lib.umn.edu:mpw00250</identifier>
  <datestamp>2005-02-22</datestamp>
</header>
- <metadata>
- <oai_dc:dc xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
  <dc:identifier>mpw00250</dc:identifier>
  <dc:identifier>OCLC40637019</dc:identifier>
  <dc:identifier>http://snuffy.lib.umn.edu/image/srch/bin/Dispatcher?mode=600&id=mpw00250</dc:identifier>
  <dc:title>"We'll have lots to eat this winter, won't we Mother?" : grow your own can your own</dc:title>
  <dc:date>1943</dc:date>
  <dc:rights>Subject to U.S. and international copyright laws.</dc:rights>
  <dc:subject>World War, 1939-1945. United States. Posters</dc:subject>
  <dc:subject>Food conservation</dc:subject>
  <dc:subject>Agriculture</dc:subject>
  <dc:subject>Victory gardens</dc:subject>
  <dc:subject>War work</dc:subject>
  <dc:language>eng</dc:language>
  <dc:creator>Parker, Alfred, 1906-</dc:creator>
  <dc:contributor>United States. Office of War Information</dc:contributor>
  <dc:description>This object is held by: Minneapolis Public Library</dc:description>
  <dc:description>"OWI Poster No. 57. Additional copies may be obtained ... from the ... Office of War Information,
    Washington, D.C."/"U. S. Government Printing Office : 1943--O-520465"</dc:description>
  <dc:description>A mother and daughter with matching blonde pony-tails and aprons process food at home, while behind
    them is a shelf packed with canned fruit and vegetables.</dc:description>
  <dc:relation>IsPartOf http://www.mplib.org/wpdb/</dc:relation>
  <dc:relation>This object is part of a larger series in the University of Minnesota digital collections database, available at
    http://digital.lib.umn.edu. The specific series is War Posters (Rationing and conservation).</dc:relation>
  <dc:format>image/jpeg</dc:format>
</oai_dc:dc>
</metadata>
</record>
</GetRecord>
</OAI-PMH>
```

[http://snuffy.lib.umn.edu:8080/image/oai/HandleRequest.do?verb=GetRecord&metadataPrefix=oai\\_dc&identifier=oai:digital.lib.umn.edu:mpw00250](http://snuffy.lib.umn.edu:8080/image/oai/HandleRequest.do?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:digital.lib.umn.edu:mpw00250)

## Granularity, differentiation, & context

- Appropriate granularity defined by anticipated use
  - Individual images on a page, full pages, or whole book?
  - Individual letters or complete archive?
  - General advice: smallest granularity appropriate for resource – consider user & use common sense
  
- OAI-PMH tends to encourage de-contextualization
  - Records shared via OAI often lose implicit linkages and context of local implementation
  - Use OAI Set Descriptions & item-level relation fields to preserve as much useful context as possible







## Metadata formats

MYTH - OAI only allows exposure of simple Dublin Core (DC) records  
MYTH - OAI allows exposure of items in only a single metadata format

- There is a distinction in OAI between **items** & **records**
  - **Item** is all available metadata for a resource
  - **Record** is dissemination of an item in specific metadata format
- **Expose richest metadata you can**; crosswalk to less rich formats as needed (oai\_dc) and as makes most sense for purpose
  - OAI requires XML Schema (.xsd) for all formats exposed
  - Must list available formats in ListMetadataFormats response; Should also list formats for given set in <setDescription>
- Consider QDC, DARE, MODS, MARCXML, MABXML, IMS, METS, MPEG, ...





## Service Providers – approaches for metadata analysis

- Statistical analysis of the metadata population
  - Size of collection
  - Structure of records
    - Metadata elements: distinct/repeated
    - Uniqueness of values
  - Length of values
    - Empty metadata elements (18% of records in CICHarvest)
- Analysis of metadata values (sampling)
  - Implicit or explicit encoding schemes
  - Necessary transformations: Clean up, splitting ....



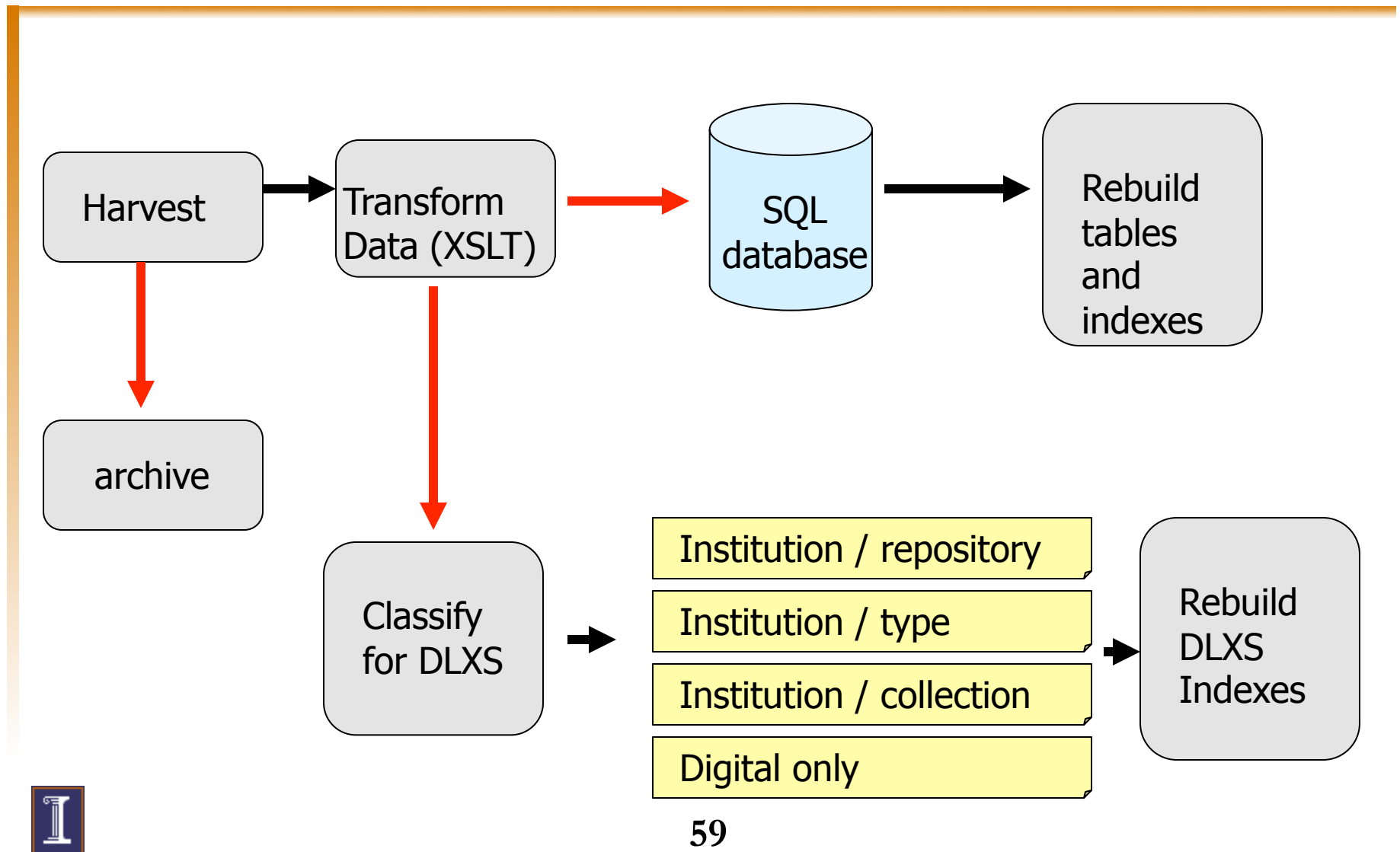


## Types of reprocessing

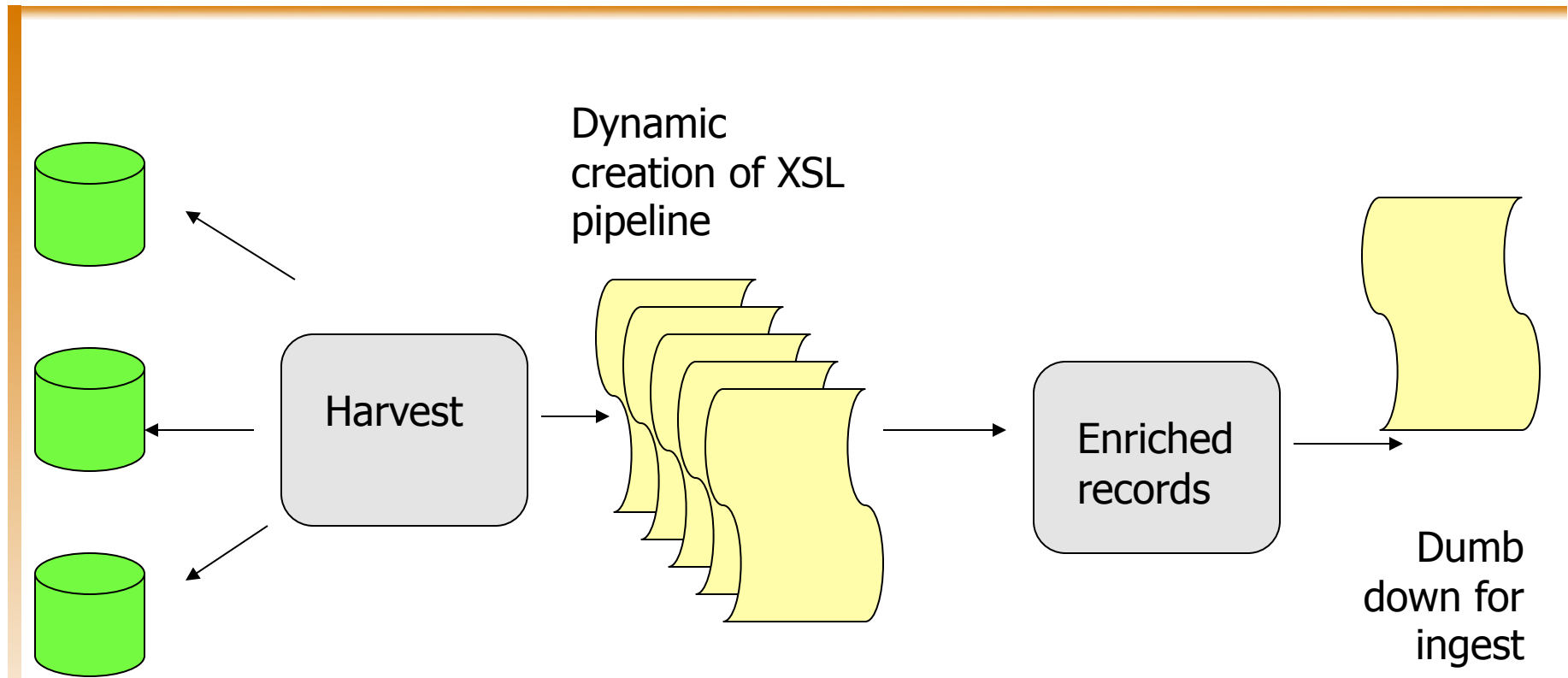
- Selection
  - What is the SP's Collection development policy
- Normalization
  - `<description>16 History; 17 Geography;</description>`
  - Semantic / syntactic
- Augmentation
  - Add / link to collection level information
  - Additional fields containing derived or inferred values
- Customization for ingest in DL applications



## CIC Project Workflow : 3 streams



# Data transformation



## What's the metadata value being used for?

- Audience is
  - machine
    - Dc:type "Text.Correspondence.Letter"
    - Dc:language "wln"
  - human
    - Dc:type Correspondence
    - Dc:language "wallon"
  
- May want different values for browse vs. search
  - Browse – limited number of controlled vocabulary values
  - Search – expressive, potentially verbose free-text values





# Functional requirements for metadata

**Results by** Collection

- Collection
- Institution
- Type
- Online Material Only

**Indiana Un**  
Sam DeVincent Collection of American Sheet Music  
**2 records**

**Michigan State University:**  
Vincent Voice Library  
**12 records**

**Ohio State University:**  
Ohio State University Thesis and dissertations  
**3 records**

**University of Chicago:**  
Digital South Asia Library  
**2 records**

**University of Illinois at Urbana-Champaign:**  
Teaching with Digital Content  
**119 records**

**University of Michigan:**  
American Council of Learned societies - History e-Book project  
**1 record**

Miscellaneous Resources, University of Michigan  
**1 record**

University of Michigan Museum of Art  
**12 records**

Art, Architecture and Engineering Library, Lantern Slide Collection  
**3 records**

Art, Architecture and Engineering Library  
**5 records**


Sort by: title

Jump to Records: 1 | 11 | 21 | 31 | 41 | 51 | 61 | 71 | 81 | 91 | 101 | 1

---

**Record 1 of 214**


|                |   |
|----------------|---|
| Title          | ...Because Somebody Talked!   |
| Author/Creator | United States. Office of War Information; Wesley [artist]   |
| Type           | Image   |
| URL            | <a href="http://images.library.uiuc.edu:8081/tdc/image/99.jpg">http://images.library.uiuc.edu:8081/tdc/image/99.jpg</a> |
| Collection     | Teaching with Digital Content   |




---

**Record 2 of 214**

|                |   |
|----------------|---|
| Title          | 1778, 1943: Americans will always fight for liberty   |
| Author/Creator | United States. Office of War Information; Perlin, Bernard, 1918- [artist]   |
| Type           | Image   |
| URL            | <a href="http://images.library.uiuc.edu:8081/tdc/image/2184020252002_ww20166p.jpg">http://images.library.uiuc.edu:8081/tdc/image/2184020252002_ww20166p.jpg</a> |



**COLLOCATE**

**SELECT INTERPRET**

**OBTAIN**





## Loss of context / reference

- Most descriptive metadata formats focus on discovery
  - Don't include thumbnails, other content useful for interpretation, selection, & use of resource
  - Limited ways to express relationships / context
- This is changing –
  - Thumbnails: NLA Qualified DC, MODS ver. 3.2
- More advanced formats offer more ways to express relationships



## Match cases for multiple term queries (MT)

|               | <b>Item desc.</b> | <b>Collection desc.</b> |
|---------------|-------------------|-------------------------|
| <b>Case A</b> | Part of Query     | Rest of Query           |
| <b>Case B</b> | No match          | All of Query            |
| <b>Case C</b> | All of Query      | No match                |
| <b>Case D</b> | Part of Query     | All of Query            |
| <b>Case E</b> | All of Query      | All of Query            |
| <b>Case F</b> | All of Query      | Part of Query           |
| <b>Case G</b> | Part of Query     | No match                |
| <b>Case H</b> | No match          | Part of Query           |
| <b>Case J</b> | Part of Query     | Part of Query           |





## Test with 1638 multiple term queries from OAIster

Partial  
match

Rest of  
match

Full  
match

No  
match

|               | # of queries with at least 1 item-level match of the case | % of queries with at least 1 item-level match of the case |
|---------------|---|---|
| <b>Case A</b> | 287   | 17.00%  |
| <b>Case B</b> | 21  | 1.20%   |
| <b>Case C</b> | 761   | 45.10%  |
| <b>Case D</b> | 25  | 1.50%   |
| <b>Case E</b> | 20  | 1.20%   |
| <b>Case F</b> | 222   | 13.10%  |
| <b>Case G</b> | 1,639   | 97.00%  |
| <b>Case H</b> | 940   | 55.70%  |
| <b>Case J</b> | 945   | 56.00%  |



## A concrete example

- Search “Harriet Beecher Stowe”
  - appears in 40 item records
  
- Q1 = “Poetry of Harriet Beecher Stowe”
  - **0** item records contain all words
  - **6** item records contain *Harriet Beecher Stowe* AND appear in collections having to do with *poetry*
  
- Q2 = “Harriet Beecher Stowe in Antebellum History”
  - **0** item records contain all words
  - **14** items contain *Harriet Beecher Stowe* AND appear in collections having to do with *antebellum history*



# Thumbnails enrich display of search results

Refine search

Sort by


title

Next

Jump to Records: [1](#) | [11](#) | [21](#) | [31](#) | [41](#) | [51](#) |


---

**Record 1 of 82**

|                |  |   |
|----------------|--|---|
| Title          | Dievotchka pokupaet varenye grushi. Une petite fille a chetant des poires. Ein kleines mÄxchen kauft Birnen. |  |
| Author/Creator | (unknown)  |   |
| Type           | Image  |   |
| Collection     | <b>Russian Publics</b>   |   |

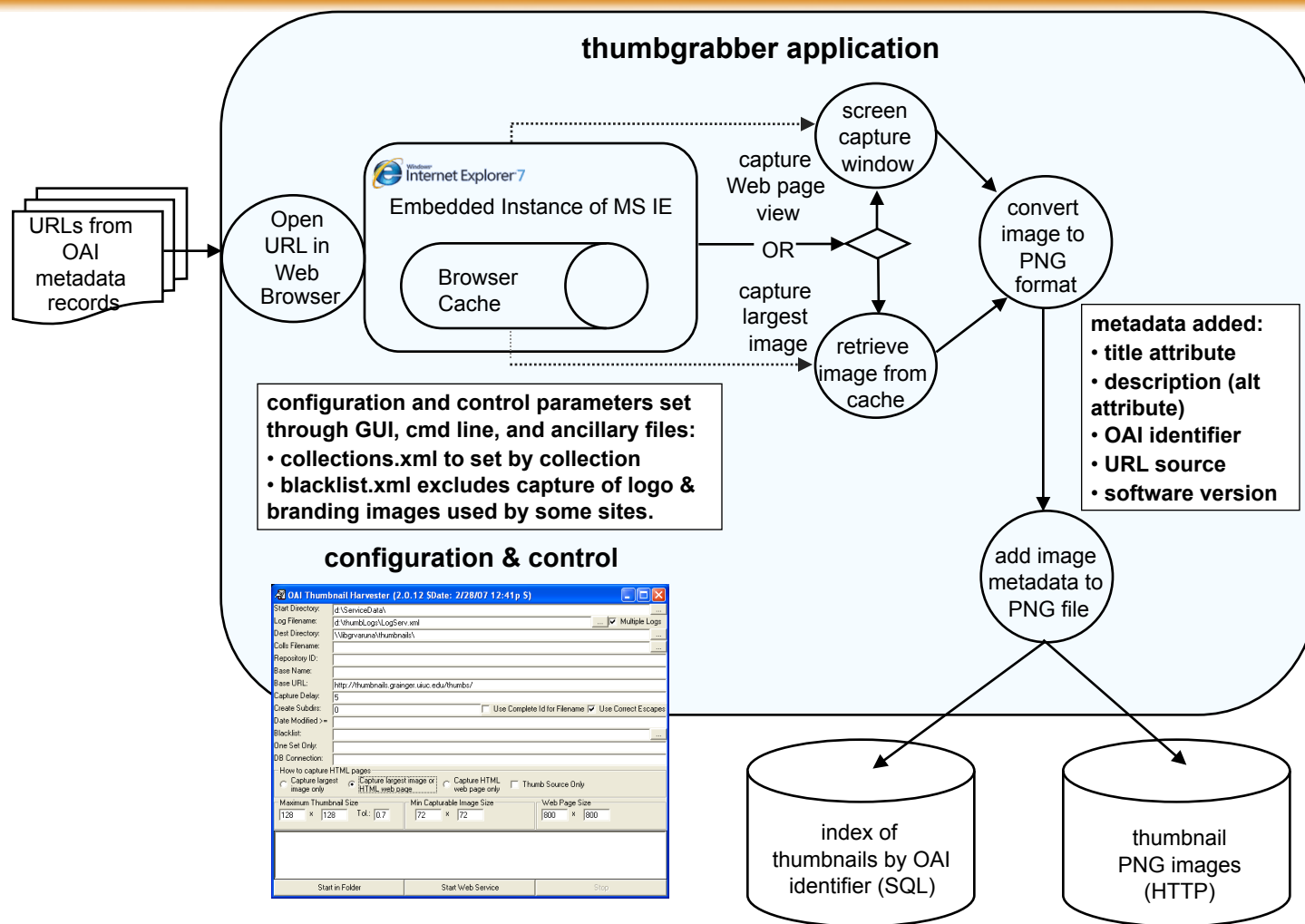
---

**Record 2 of 82**

|                |   |   |
|----------------|---|---|
| Title          | Dvornik i Trubochist. Le portier et le ramoneur. Der Hausknecht und der Schornsteinfeger. |  |
| Author/Creator | (unknown)   |   |
| Type           | Image   |   |
| Collection     | <b>Russian Publics</b>  |   |

# Thumbgrabber application architecture

(from T. Cole, T. Habing, & M. Foulonneau. 2007. "An OAI-PMH Based Thumbnail Capture and Delivery Service." Poster presented at OAI5 Workshop, Geneva, CH, April 2007.)



## Collection information can provide useful context

- Potential influence on
  - Result ranking when limited number of results matching the full query
  - Co-location – suggested related resources through collection access
- Property inheritability
  - Always -> no added value of combined operator
  - Never -> no interest in collection / item retrieval
  - Sometimes -> A non matching record is in a collection where other records match, therefore ....
- User expected granularity level
  - Future user testing
- Other pieces contain context
  - resources
- Data providers are exposing more CLD
  - What is the best way to share them?

**Resources**



**Metadata**

**Collections**

**services**