Access IT Training

How objects and metadata from digital repositories can be reused?
Introduction

Metadata reuse scenarios
  - OAI-PMH

Object content reuse scenarios
  - OAI-ORE

Persistent identifiers

Conclusions
“It is highly probable that digital libraries are created for users” 😊

We assume that someone will:
- read books
- analyze manuscripts
- watch old documentaries
- ...

Metadata and object reuse
Goals of this presentation

- Go beyond “walls” (pages) of our digital library
  - Discuss how to attract users to our digital library
  - Show why it is a good thing to encourage others (users/services) to reuse digital library objects
- Show why OAI-PMH is such a big thing
- Investigate how digital libraries can be used in education
- Talk a bit about persistent identifiers
Who will reuse DL content?

Users

Services

General purpose search engines

Aggregators

Other...
What can be reused?

- **Metadata records**
  - Various protocols and formats

- **Object content**
  - Text, images, video etc.
Introduction

Metadata reuse scenarios
- OAI-PMH

Object content reuse scenarios
- OAI-ORE

Persistent identifiers

Conclusions
Metadata Aggregators

- Metadata can be reused through aggregation
- Definition:
  - “An **Aggregator** is an organization that collects metadata from its group of content providers and transmits them to Europeana, helps content providers with guidance on conformance with Europeana norms and converts metadata if necessary. The aggregator also supports the content providers with administration, operations and training.”

  Europeana Content Strategy
Metadata Aggregators

- Aggregator gathers metadata from a number of repositories/digital libraries
- On top of those data new advanced network services can be built
  - Metadata unification and cleaning
  - Forwarding metadata to other services
  - Unified access point to all resources
The model of aggregation of content is of crucial importance and will enable Europeana to reach its objectives. Aggregators, on a national, regional or vertical level, play a key role not only in aggregating content, but also in the organizational structure, standardization of content, services to end-users and future sustainability of Europeana and related projects and aggregators”

Europeana Content Strategy
Two types of aggregation

- Horizontal aggregation
  - General purpose aggregation
  - Gathers metadata of various resources from different institutions

- Vertical aggregation
  - Thematic aggregator
  - Gathers information about specific types of objects
    - e.g. European Film Gateway
Metadata aggregation and reuse

- Metadata aggregation can be realized using different means one of them is OAI-PMH
Metadata exchange using OAI-PMH

- OAI-PMH – Open Archives Initiative Protocol for Metadata Harvesting
  - Defines two actors
    - data providers (repositories)
    - service providers (harvesters)
Service providers provide services based on the metadata harvested from data providers.
Service providers provide services based on the metadata harvested from data providers.

OAI for Beginners - the Open Archives Forum online tutorial
(http://www.oaforum.org/tutorial/english/intro.htm)
Metadata exchange using OAI-PMH

- OAI-PMH – Open Archives Initiative Protocol for Metadata Harvesting
  - XML based
  - Built on top of HTTP protocol
  - Widely used
    - Over 2000 repositories worldwide containing several millions of records
  - More info: http://www.openarchives.org/
OAI-PMH Repository data structure

- Repository contains items
- Item may be described in multiple metadata schemas
- Item must be described with Dublin Core schema (15 general attributes – title, author etc.)
- Metadata is exchanged via records
  - entire description of an item in the particular schema
  - headers (basic description - id, modification date etc.)
Metadata exchange using OAI-PMH

- OAI-PMH Repository data structure
  - Repository may define sets (groups of items) – possible usage:
    - Subject sets, MIME type sets, ...
  - Set may contain subsets
  - Each item may belong to zero or more sets
  - Sets are used for selective harvesting – they allow to harvest only a part of a repository
Communication between OAI-PMH data provider and service provider is based on predefined verbs, which includes:

- **Identify**
  - Basic information about repository and its OAI-PMH interface

- **ListIdentifiers**
  - Lists identifiers for a given set and metadata format
- **ListMetadataFormats**
  - Lists supported metadata formats

- **ListRecords**
  - Lists records for a given set and metadata format
Metadata exchange using OAI-PMH

- **GetRecord**
  - Gets content of given record

- **ListSets**
  - Lists available sets
More examples can be found in: “The Open Archives Initiative Protocol for Metadata Harvesting”

- [http://www.openarchives.org/OAI/openarchivesprotocol.html](http://www.openarchives.org/OAI/openarchivesprotocol.html)
Pionier Digital Libraries Federation is a Polish national aggregator

- [http://fbc.pionier.net.pl/](http://fbc.pionier.net.pl/)

It was **created to facilitate** the use of resources from Polish digital libraries

- To **increase the visibility** of these resources in the Internet

- To **create new, advanced network services** both for end-users and digital libraries creators on the base of these resources
Basic assumptions

- **No need** nor requirement **to move** resources to the DLF
- **No fees** for the use of the DLF and for being a part of it
- **Open standards are the basis** for cooperation
  - Particular digital libraries can use different technological platforms
Available publications: 348,549
Planned publications: 5,387

Resources of "Digital Library of Polish and Poland-Related News Pamphlets"

From today we provide details about the publications of "Digital Library of Polish and Poland-Related News Pamphlets".
Overall number of digital objects
✓ 340 thousands

Number of active digital libraries:
✓ 21 regional
✓ 28 institutional

Number of cooperating institutions:
✓ over 200 libraries, museums and archives
Digital Libraries Federation as a metadata aggregator for Europeana

Metadata aggregator

Digital Libraries Federation

- Institutional
- Regional
- National (exclude??)
- Other

Institutions

- Libraries
- Archives
- Museums
- ....
Digital Libraries Federation

- Basic functions
  - Search in the available publications
    - Simple
    - Advanced
  - Digitization plans
    - Searchable
    - Report
    - API for the prevention of duplicated digitization
  - Location of digital objects on the basis of their OAI Identifiers
  - Database of Polish digital libraries
  - Statistics and reports
- Information in the DLF is updated on the daily (nightly) basis
Promotion of Polish Digital Resources

- OpenSearch plug-in for web browsers available since the beginning of the DLF
  - Must be installed manually by the user
  - The user must know that there are digital libraries and the Federation
- Several months of activities towards the inclusion of DLF plug-in into official Polish Firefox releases
  - Negotiations with Polish Mozilla team, and finally with Mozilla Foundation
  - Technical changes related to high performance requirements
    - Autosuggest service response time less than 300 ms
Digital Libraries Federation search plugin
Promotion of Polish digital resources

- June 17th, 2008 – Firefox 3.0 released with DLF plug-in included in the Polish version
- July 3rd, 2008 – Firefox 2.0.15 released with DLF plug-in included in the Polish version
- Resources of Polish digital libraries available for Polish internet users together with Google, Wikipedia etc.
  - Currently Firefox 2 and 3 have together 41% of web browsers market in Poland
- Measurable result?
Promotion of Polish digital resources
- Total number of visits in the DLF each month

DLF included in Firefox 2 and 3
The Digital Libraries Federation architecture

End-users, external systems

Digital Libraries Federation application layer – based on atomic services

Networked user profile service

Harvesting – indexing – searching service for digitisation plans

Harvesting – indexing – searching service for digital objects

Supporting services

User profile sharing

OAI-PMH-based periodic metadata harvesting

Digital libraries distributed in the PIONER network

Digital library

Digital library

Digital library

Digital library
Pionier DLF has its own OAI-PMH 2.0 compliant interface
- All gathered data are available through this interface
- It offers the support for deleted records and incremental harvesting
- It also allows to create dynamic OAI sets on a search query basis

Pionier DLF exposes unified metadata from Polish Digital Libraries in ESE format

Since **11th December 2009** that interface is harvested by Europeana
Other examples of aggregators

- OAIster
  - [http://oaister.worldcat.org/](http://oaister.worldcat.org/)
  - 23 million of records from 1100 institutions
  - Resources also visible through WorlCat.org

- ScientificCommons.org
  - [http://scientificcommons.org](http://scientificcommons.org)
  - “The major aim of the project is to develop the world’s largest communication medium for scientific knowledge products which is freely accessible to the public.”
  - 13 million of scientific publications
What about vertical services?

Europeana wants to aggregate all publicly available digital content relevant to the term “European cultural and scientific heritage”

What about vertical services based on a large scale aggregation?

- There is a need to enable precise selective harvesting of aggregated metadata
Example scenario: Thematic portal built on top of distributed OAI-PMH repositories

How to obtain the metadata?

- **Solution 1:** Harvest all records from repositories, decide what records are useful
  - A lot of useless data is harvested and processed

- **Solution 2:** Harvest only specific sets of items matching the theme of the portal
  - Each harvested repository must define a set / sets matching the theme of the portal – practically impossible

- **Solution 3:** DIY variant of scenario 2 – define a set containing items matching the theme of the portal and harvest it
  - Not supported in the OAI-PMH protocol
Dynamic sets – OAI-PMH protocol compatibility

Harvester side
- If a harvester does not support dynamic sets, it will still be able to harvest the repository supporting such sets.

Repository side
- If a repository does not support dynamic sets, it still may be harvested by a harvester supporting such sets.
- The repository extended with dynamic sets should be compatible with OAI-PMH validators.
Proposed OAI-PMH extension: dynamic sets

- Dynamic sets – implementation
  - Harvester side
    - Prepare the support for OAI-PMH set harvesting
    - Analyze the nature of metadata in particular repository and prepare proper dynamic set definition to use during harvesting
  - Repository side
    - Modify the harvesting requests processing to support the definition of dynamic sets
    - This may be based on the search mechanism already implemented in the majority of repositories – in such case the support for CQL queries must be assured
Dynamic sets – specification

- Sets defined by repository users
- Contain items that matched dynamic set definition sent by the user
- The definition is in fact a CQL query encoded into a set name
  - CQL – Contextual Query Language
    - Part of SRU protocol specification – used in integrated library systems as a replacement for the z39.50 protocol to obtain bibliographic descriptions
    - Allows to define simple and complex queries
    - Compatible with any metadata schema
    - Example: dc.creator = “Albert Einstein”
Parse requests parameters.

Is dynamic set prefix included?

Is dynamic set definition included?

Return "noRecordsMatch" error.

Prepare Lucene query based on static set, dynamic set definition and modification dates.

Run the search process.

Load identifiers of items from specified static set.

Load identifiers of items that matched the prepared query.

Phase I
Static set processing

1. Load identifiers of items from specified static set.
2. Run the search process.
3. Load identifiers of items that matched the prepared query.
4. Partition the list of identifiers into smaller lists for retrieval with resumption tokens.
5. Load the metadata of items that will be returned to the harvester in first response.
6. Transform the metadata to OAI-PMH response and return it to the harvester.

Phase II
## Tests results

<table>
<thead>
<tr>
<th>Query</th>
<th>Harvested number of repositories</th>
<th>Harvested number of records</th>
<th>Harvested % of overall number of repositories</th>
<th>Harvested % of overall number of records</th>
</tr>
</thead>
<tbody>
<tr>
<td>none (all records)</td>
<td>16</td>
<td>93681</td>
<td>100,00%</td>
<td>100,00%</td>
</tr>
<tr>
<td>dc.language eng</td>
<td>13</td>
<td>626</td>
<td>81,25%</td>
<td>0,67%</td>
</tr>
<tr>
<td>dc.language ger</td>
<td>12</td>
<td>10357</td>
<td>75,00%</td>
<td>11,06%</td>
</tr>
<tr>
<td>dc.type <em>podręcznik</em> (handbook)</td>
<td>4</td>
<td>104</td>
<td>25,00%</td>
<td>0,11%</td>
</tr>
<tr>
<td>dc.type <em>rozprawa</em> (thesis)</td>
<td>5</td>
<td>199</td>
<td>31,25%</td>
<td>0,21%</td>
</tr>
<tr>
<td>dc.type <em>czasopismo</em> (magazine)</td>
<td>16</td>
<td>28163</td>
<td>100,00%</td>
<td>30,06%</td>
</tr>
<tr>
<td>dc.type <em>gazeta</em> (newspaper)</td>
<td>4</td>
<td>33793</td>
<td>25,00%</td>
<td>36,07%</td>
</tr>
<tr>
<td>dc.subject <em>pedagogika</em> (pedagogy)</td>
<td>8</td>
<td>130</td>
<td>50,00%</td>
<td>0,14%</td>
</tr>
<tr>
<td>dc.subject <em>chemia</em> (chemistry)</td>
<td>8</td>
<td>715</td>
<td>50,00%</td>
<td>0,76%</td>
</tr>
<tr>
<td>dc.subject</td>
<td>8</td>
<td>2759</td>
<td>50,00%</td>
<td>2,95%</td>
</tr>
</tbody>
</table>
Current usage - Manuscriptorium

- eContentPlus ENRICH Project
  - Started in December 2007
  - The aim is to built a virtual European repository of manuscripts
  - The metadata about the manuscripts is harvested from multiple European repositories
  - Harvests metadata of manuscripts from several Polish digital libraries
Criteria used to get objects for Manuscriptorium

- dc.date>=1000
- dc.date<=1850
- not dc.type=czasop*
- not dc.type=prog*
- dc.format=(app* or pd* or dj* or im*)

Results: **883** objects (22.06.2009)
Jeder Montag nach Februarium, in Großberg
R. Mariae vom Gottes graelig gegr.
1490

J. Pater

Mit dem Namen von Joannes Gabriel, Escho
Scopna Gebel, und Elisaf, kleinster von Freiburg
zu folgen geschrieben in einem zu Berlin, ich habe
gegen nicht in meinem Namen, sondern wider die
gemeine

Jeder Montag nach Februarium, in Großberg
R. Mariae vom Gottes graelig gegr.
1490

J. Pater

Mit dem Namen von Joannes Gabriel, Escho
Scopna Gebel, und Elisaf, kleinster von Freiburg
zu folgen geschrieben in einem zu Berlin, ich habe
gegen nicht in meinem Namen, sondern wider die
gemeine

Jeder Montag nach Februarium, in Großberg
R. Mariae vom Gottes graelig gegr.
1490

J. Pater

Mit dem Namen von Joannes Gabriel, Escho
Scopna Gebel, und Elisaf, kleinster von Freiburg
zu folgen geschrieben in einem zu Berlin, ich habe
gegen nicht in meinem Namen, sondern wider die
gemeine

Jeder Montag nach Februarium, in Großberg
R. Mariae vom Gottes graelig gegr.
1490

J. Pater

Mit dem Namen von Joannes Gabriel, Escho
Scopna Gebel, und Elisaf, kleinster von Freiburg
zu folgen geschrieben in einem zu Berlin, ich habe
gegen nicht in meinem Namen, sondern wider die
gemeine

Jeder Montag nach Februarium, in Großberg
R. Mariae vom Gottes graelig gegr.
1490

J. Pater

Mit dem Namen von Joannes Gabriel, Escho
Scopna Gebel, und Elisaf, kleinster von Freiburg
zu folgen geschrieben in einem zu Berlin, ich habe
gegen nicht in meinem Namen, sondern wider die
gemeine
Man Monday nach Feldmarschall, zu Göttingen
König Wilhelm zum Gott gnädig 1490.
Current usage - Manuscriptorium

- Automatic conversion from DjVu to JPG
- Publication structure exposed through OAI-PMH encoded using METS
  - Links to pages and files
  - Info about original DjVu files and results of conversion
Current usage - DART

- DART-Europe E-theses Portal
  - [http://www.dart-europe.eu](http://www.dart-europe.eu)
  - „DART-Europe is a partnership of research libraries and library consortia who are working together to improve global access to European research theses. “
- Resources from Polish digital libraries are exposed through OAI-PMH interface of Pionier DLF
### Browse Results

Displaying records 1 - 15 of 300 on page 1 of 21.

<table>
<thead>
<tr>
<th>Mark</th>
<th>Title</th>
<th>Author</th>
<th>Year</th>
<th>Institution</th>
<th>Collection</th>
</tr>
</thead>
<tbody>
<tr>
<td>☐</td>
<td>A Method for Analysis of Node Position in the Network of Internet Users</td>
<td>Musial, Katarzyna</td>
<td>2009</td>
<td>Politechnika Wrocławska</td>
<td>PIONIER</td>
</tr>
<tr>
<td>☐</td>
<td>Akceptacja sprawozdawcza działania arytmetycznych w algorytmach oświetlenia globalnego</td>
<td>Tomczak, Tadeusz</td>
<td>2007</td>
<td>Politechnika Wrocławska</td>
<td>PIONIER</td>
</tr>
<tr>
<td>☐</td>
<td>Algoryt podejmowania decyzji z wykorzystaniem identyfikacji dwustopniowej</td>
<td>Brzostowski, Krzysztof</td>
<td>2009</td>
<td>Politechnika Wrocławska</td>
<td>PIONIER</td>
</tr>
<tr>
<td>☐</td>
<td>Algoryt kompensacji warunków transmisji i cech osobniczych mówcy w systemach automatycznego rozpoznawania mów</td>
<td>Mrówka, Paweł</td>
<td>2007</td>
<td>Politechnika Wrocławska</td>
<td>PIONIER</td>
</tr>
<tr>
<td>☐</td>
<td>Algoryt rozpoznawania zespołu QRS w sygnałach elektrokardiograficznych pochodzących od pacjentów z wszczepionym układem stymulującym</td>
<td>Duraj, Agnieszka</td>
<td>2007</td>
<td>Uniwersytet Zielonogórski</td>
<td>PIONIER</td>
</tr>
<tr>
<td>☐</td>
<td>Algoryt uczenia sieci neuronowych odpornych na błędy w danych</td>
<td>Rusiecki, Andrzej</td>
<td>2007</td>
<td>Politechnika Wrocławska</td>
<td>PIONIER</td>
</tr>
<tr>
<td>☐</td>
<td>Algoryt wyszukiwania drobnojakościowej równoległości w pętli programowej z zależnościami afinicznymi</td>
<td>Siedlecki, Krzysztof</td>
<td>2008</td>
<td>Zachodniopomorski Uniwersytet Technologiczny</td>
<td>PIONIER</td>
</tr>
<tr>
<td>☐</td>
<td>Algoryt zwiększające ekstrakcję równoległości w pętłach programowych</td>
<td>Pałkowski, Marek</td>
<td>2008</td>
<td>Zachodniopomorski Uniwersytet Technologiczny</td>
<td>PIONIER</td>
</tr>
<tr>
<td>☐</td>
<td>Analiza detekcji siłowych, rozproszonych, dopiero przesuniętych w częstotliwości niestacjonarnych wiązań laserowych</td>
<td>Waż, Adam</td>
<td>2008</td>
<td>Politechnika Wrocławska</td>
<td>PIONIER</td>
</tr>
<tr>
<td>☐</td>
<td>Analiza efektów działania promieniowania laserowego na strefie niestacjonarnych wiązań</td>
<td>Szajner, Hanna</td>
<td>2009</td>
<td>Uniwersytet Zielonogórski</td>
<td>PIONIER</td>
</tr>
</tbody>
</table>

[Add to marked list] [Browse again] [« Back] [1 2 3 4 5 6 Next »] [Last page »]
Other means to reuse metadata records

- Social bookmarking

- Bibliography organizers
  - Zotero
    - Handles different metadata encodings:
      - MODS, RDF, MARC, BibTeX
      - http://www.zotero.org/
  - Bibsonomy (http://www.bibsonomy.org/)
  - ...

Introduction

Metadata reuse scenarios
  - OAI-PMH

Object content reuse scenarios
  - OAI-ORE

Persistent identifiers

Conclusions
Object reuse

- Blogs
  - Good way to get external links to digital library website (good for SEO)
  - Readers know more about resources
- Accumulating user generated content
  - Europeana
  - Flickr: The Commons
  - Panoramio
Tunnel, Wombeyan Road

Format: Glass plate negative.

Rights Info: No known restrictions on publication.

Repository: Tyrrell Photographic Collection, Powerhouse Museum
www.powerhousemuseum.com/collection/database/collection=The_Tyrrell_Photographic

Part Of: Powerhouse Museum Collection

This photo also belongs to:
- Tyrrell Collection (Set)

Tags:
- Powerhouse Museum
- B&W
- Tunnel
- Trees
- Car
- wombeyan
- mittagong

5 people call this photo a favorite
General information about the Powerhouse Museum Collection is available at www.powerhousemuseum.com/collection/database

Persistent URL: http://www.powerhousemuseum.com/collection/database/?irm=28304

Acquisition credit line: Gift of Australian Consolidated Press under the Taxation Incentives for the Arts Scheme, 1985

Comments

iansand pro says:

Wombeyan Caves are limestone caves to the west of the highway between Goulburn and Mittagong. There are two ways in - from the Goulburn to Oberon road, which is on a relatively straightforward sealed road, or from Mittagong which is a narrow, twisting unsealed road. This tunnel is on the road from Mittagong.

The tunnel in March 2009

Powerhouse Museum Collection pro says:

Thanks! We have now mapped the location thanks to your geotagging on the contemporary photo!

Posted 8 months ago. (permalink)

Posted 5 months ago. (permalink)
OAI: Object Reuse and Exchange

- OAI-ORE defines standards for the description and exchange of aggregations of Web resources
  - [http://www.openarchives.org/ore/](http://www.openarchives.org/ore/)
- The goal of these standards is to expose the rich content to applications that support:
  - authoring, deposit, exchange, visualization, reuse, and preservation.
OAI: Object Reuse and Exchange

Source: ORE User Guide – Primer
(http://www.openarchives.org/ore/1.0/primer.html)
Some solution exists in this space already (METS)
- OAI ORE is built on top of successful OAI-PMH
- It’s a very flexible spec, promising low adoption curve
- ORE can be used to aggregate whole objects (+metadata)
- More about ORE in the last part of tutorial
Agenda

- Introduction
- Metadata reuse scenarios
  - OAI-PMH
- Object content reuse scenarios
  - OAI-ORE
- Persistent identifiers
Nowadays it is very easy to create new website, publish content on the web. Over time, more and more of these hyperlinks are "broken". Organization’s website was re-organized. Persistent identifiers introduce a service which records how precise location of document changes over time.
When end-user wishes to access a document, the identifier in his request is „resolved”

The correct document is retrieved

- User don’t need to know the exact location of the document

Support for persistent identification some administrative effort
Persistent identifiers

- Various propositions:
  - Handles,
  - Digital Object Identifier (DOI)
  - Archival Resource Keys (ARK)
  - Persistent Uniform Resource Locators (PURL)
  - Uniform Resource Names (URN)
  - OAI Identifiers
The syntax of URNs was fully specified in 1997 in another RFC, “URN Syntax”

**Global uniqueness**: different resources cannot have the same URN

**Persistence**: in the URN context, the name's lifespan is permanent, regardless of the lifespan of the named resource.

**Scalability**: room to accommodate the number of names required in the next centuries.
Uniform Resource Names

- Syntax: `urn:<NID>;<NSS>`
- Every URN begins with the ‘urn:’ character string, followed by the Namespace Identifier (NID, e.g. ISBN)
- Namespace Specific String (NSS) syntax depends on the namespace identified by the NID
- Example: `urn:isbn:3-938616-59-8`
OAI Identifier

- **Syntax:** `oai:<NI>:<LI>`
- **Persistent identifier for repositories with OAI-PMH interface**
- **Identifier points to resource which holds metadata record**
- **NI** – namespace identifier – repository domain
- **LI** – local identifier – points to resource in repository
OAI Identifier

- It does not require central resolution service
- No fees for using this solution
- Example:
  - oai:www.wbc.poznan.pl:8711
Conclusions

- The more doors we have in our library there is a bigger chance that someone will come in.
- Reuse is a key to enrichment
  - Chance for User Generated Content
  - Chance to attract new users
- Whole Web 2.0 is about reuse and exchange
  - Open protocols and public APIs
Conclusions

- Search Engine Optimization
  - Source of external links to our content
  - Everything works fine as long as links are working - persistent identifiers might be useful