Metadata harvesting in regional digital libraries in PIONIER Network

Cezary Mazurek, Maciej Stroiński, Marcin Werla, Jan Węglarz
Polish Optical Internet PIONIER
Digital Libraries in PIONIER (2005)
Digital Libraries in PIONIER

- Support for multiple content types
- Hierarchic collections
- Resources versioning
- Advanced support for resources’ metadata (DublinCore, MARC, RDF)
- Multiple searching mechanism
- Extended access management (LDAP, Single Sign-On)
- RSS-based notification
- Multitier, distributed and portable DL platform
dLibra Digital Library Framework

http://dlibra.psnc.pl/

System architecture

dLibra system is a multithread system composed of main module dLibra - Server and modules of client applications:

dLibra - Reader and dLibra - Editor. dLibra - Server module is responsible for overall behaviour of digital library based on dLibra system and realizes all system activities. Client application modules are user interfaces which allow using functionality of digital library.

dLibra - Editor module is an application which allows connecting authorized user to dLibra - Server and working with digital library. Interface of the application is based on GUI of operating system on which the dLibra - Editor module is running. This module has to be installed in every computer on which it is to be used.

dLibra - Reader application enables Internet users to access publications stored in digital library. This module acts as an access point to the content of the dLibra - Server. It serves publications through the web pages interface (WWW). When installed on the server side it enables all Internet users to use dLibra system functionality according to granted rights. dLibra - Reader module may have many instances in order to distribute the load in case of big digital libraries.

Software and hardware requirements for dLibra system

1. Server:
   - Software requirements:
     - Java(TM) virtual machine by Sun Microsystems Inc., according to version 1.5
     - Relational database Oracle (version 9.0 or higher) or PostgreSQL (version 8.0 or higher) or MySQL (version 4.1 or higher)
   - Hardware requirements:
     - according to point 4
   - additional space on hard drive to store all publications which are to be in the digital library

2. Reader application:
   - Software requirements:
     - Java(TM) virtual machine by Sun Microsystems Inc., according to version 1.5
     - Application server which allows running applications according to Java(TM) Servlets 2.3 specification - Apache Tomcat version 5.5 or higher is recommended
Architecture

- Readers
- Editors
- Administrator

- WWW Server
- Single Sign-On
- LDAP Server

- dLibra Distributed Services
  - Event Server
  - Search Server
  - User Server
  - Distributed Search Server
  - Metadata Server
  - Content Server

- Other OAI-PMH repositories

- Publications Archive
Service Oriented Architecture

- Each service can be deployed on different host
- Services does not need to know their locations
- Services can send and receive events
- Services implementations can be easily replaced
- Services can access each other (if authorized)
- External services can access DL system (if authorized)
  - Via dLibra interfaces with Java RMI
  - Via OAI-PMH
Service layers

User interface layer
- WWW Service
- Editor and administrator application
- Monitoring and managing interface

Services layer
- Content Service
- Metadata Service
- Search Service
- User Service
- Distributed Search Service

Supporting services
- System Service
- Event Service
Services

• Functional Services
  – Metadata Service
  – Content Service
  – Search Service
  – Distributed Search Service
  – User Service

• Supporting Services
  – System Service
  – Event Service
User interface

• WWW Service
  – Servlets
  – Portlets
• Editor/Administrator Application,
• Service Monitoring and Management Interfaces
dLibra release 2.2 reader features

• Distributed Search
• Latest Publications
• Advanced search
• RSS 2.0 feeds
• Search results as a RSS feed
• Publication description tabbed
• .....
OAI-PMH

• Open Archives Initiative
  – Protocol for Metadata Harvesting
• http://www.openarchives.org
OAI-PMH Basic Terms

- **Harvester**
  - a system (client) which wants to access metadata stored in a remote system

- **Repository**
  - a system (server) which allows to access metadata stored in it.

- **Element**
  - a basic object described in a repository, identified by an identifier unique at least at the repository level

- **Record**
  - an element’s metadata in particular format

- **Set**
  - a way of grouping elements in a repository
  - each repository accessible through the OAI-PMH protocol can define its own sets of elements. Sets can create a hierarchical structure – each set can have subsets. Each element can belong to any number of sets
OAI-PMH Other Features

• Criteria for selective harvesting
  – Last date of resource metadata record modification
  – Element membership in Sets
• Deleted records
• List of friends
  – allows to create easy discoverable networks of the connected repositories
OAI-PMH Requests

• Identify
  – gets information about the repository
• ListMetadataFormats
  – gets information about metadata formats supported by the repository. Each repository must at least support Dublin Core Metadata Element Schema.
• ListSets
  – gets information about sets defined in the repository
• GetRecord
  – gets a single record – describing a specified element in a specified metadata schema
• ListIdentifiers
  – gets a list of identifiers of elements available in the repository.
• ListRecords
  – gets a list of records available in the repository.
Metadata exchange between DLs

• Periodic metadata synchronization
  – Selective harvesting
  – Deleted records
• Collections -> Sets
• Metadata -> DC MES
• OAI-Identifiers
• Each DL has all metadata from the platform
Implementation

- dLibra became OAI-PMH repository as well as harvester
- Compatible with OAI-PMH 2.0
- Repository interface is embedded in WWW Service
  - Persistent support for deleted elements;
  - Sets
  - Flow control
  - Globally unique ids
- Harvester is provided as Distributed Search Service
  - List of repositories
  - Periodical harvesting
Content-based Services

• Distributed searching
• Virtual collections
• Distributed exhibitions
• Comments and annotations
Distributed searching

• As a result the user receives one search results list with both local and remote resources which metadata matched the user query

• Might be accessed by any other network service or information portal
Informacje

- Opis projektu
- Uczestnicy Projektu
- Informacje techniczne
- Najczęściej zadawane pytania
- Kontakt

Biblioteka

Wielkopolska Biblioteka Cyfrowa

- Działanie
- Kultura
- Muzyczne
- Archiwum
- Muzeum

Wyniki wyszukiwania ogólnego w kolekcji

Wielkopolska Biblioteka Cyfrowa

- Wyświetl wyniki jako RSS
- Liczba znalezionych publikacji: 721
- Lista publikacji spełniających warunki
- Użyj sygnatury:
- Przeszukaj zgodnie biblioteki

21. Toruń - Ratusz Staromiejski
22. Toruń - ulica Chełmińska
23. Toruń - ulica Jana Kochanowski
25. Z ludem wielkopolskim przestaw zaborcop; wspomnienie - Janusz Jacek D., Wacław
26. Księga adresowa małta Toruń; według stanu z czerwca 1936 - Reinke, Martin, dyrektor biura Zboru Mięśni w Toruniu, (oraz wiodł i wydał)
29. Szkoła Podchorążych Artylerii : ku uczciu dziesięciolecia 1922-1932
Virtual collections

• Static collections
• Dynamic collections
  – a set of conditions that object metadata must satisfy to belong to the collection
  – Based on RSS feeds
• Unified collections
  – Unified approach for collection names
  – Building virtual collections through metadata exploration
Concluding Numbers

• 9 digital libraries in PIONIER based on proprietary software
• Over 25,000 publications already digitalized and made available and searchable
• 70,000 of attribute values in dictionaries
• 320,000 references to dictionaries
Thank you!

Cezary Mazurek
mazurek@man.poznan.pl