POZNAŃ SUPERCOMPUTING AND NETWORKING CENTER

POZNAŃ SUPERCOMPUTING AND NETWORKING CENTER

dLibra Digital Library Framework Overview

Marcin Werla
mwerla@man.poznan.pl
Poznań Supercomputing and Networking Center, Poznań, Poland

Disclaimer

The aim of this presentation is to give a quick overview of the dLibra system with the focus on distributed services giving the functional basis for dLibra end-user applications. Therefore this presentation omits vast functionality implemented in the end user applications.

- dLibra website: http://dlibra.psnc.pl/

Agenda

- Introduction and background information
- Functionality overview
- Architecture
- Development process
- Users community
- Licensing
- Future works

Introduction and background information

- dLibra development was started in 1999 in the PSNC's Network Services Department
- It was a continuation of earlier works in the digital libraries area
- Initially the project was focused on electronic publishing
- When the Polish Optical Internet PIONIER (the Polish NREN) implementation was started, the dLibra software became a possible basis for new type of services in this network – digital libraries for scientific and cultural institutions

Introduction and background information

- Thanks to the cooperation with Poznan Foundation of Scientific Libraries in 2002 first dLibra-based digital library was made available publicly
- It was the Digital Library of Wielkopolska, which for the last few years is the largest Polish digital library, with more than 125 000 objects today
- In the next years dLibra was adopted by several other institutions crossing the level of 60 deployments in 2010

Introduction and background information

- The number of institutions that use dLibra can be counted in hundreds because the dLibra architecture facilitated regional cooperation of several institutions in one digital library
  - This became the dominating organizational model of digital libraries in Poland
- From today's perspective we may say that in the context of science dLibra-based digital libraries provide source material for Humanities and facilitate popularisation of research results from many domains
- Besides of deployments in Poland we have also few abroad:
  - Test phase: Lviv (Ukraine), Goeteborg (Sweden)
  - Almost in production: Belgrade (Serbia) and Jerusalem (Israel)

Introduction and background information

- In 2009 we started cooperation with National Museum in Warsaw on the software package for digital museums named “dMuseion”
- It is developed on the basis of the set of services created for dLibra
- The differences in functionality are visible (and implemented) only on the level of end-user applications and configuration of services

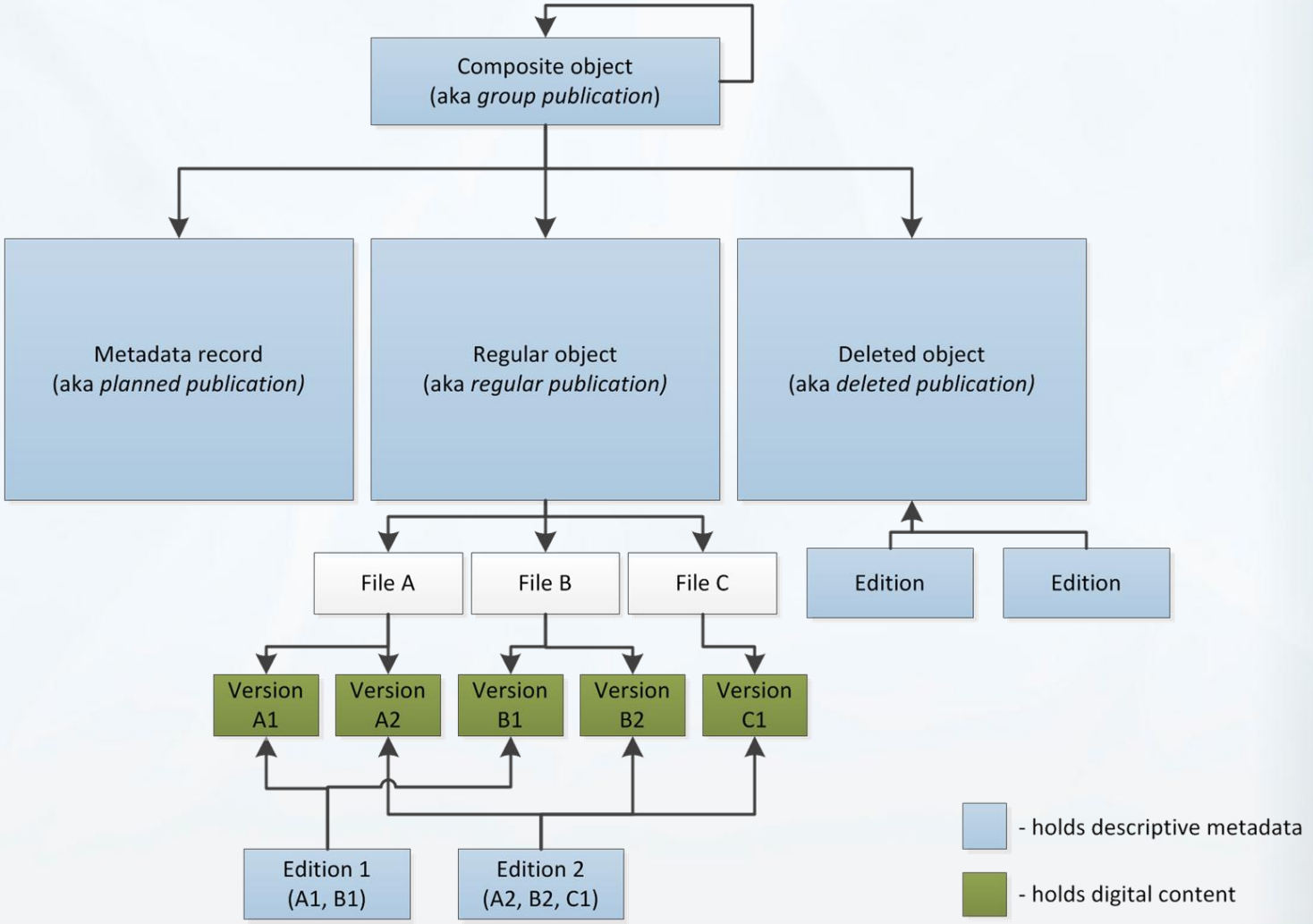
Functionality overview

- Storing digital objects of any type with associated metadata
- Metadata is indexed for full text searching
- Textual content of digital objects stored in one of several supported formats is also indexed for full text indexing
  - Support for new formats can be implemented as a plugin
- Digital objects are organized into collections (*:*) and directories (*:1)
  - Both collections and directories can be organized into hierarchical structures
  - Collections are designed for readers, allow to organise the same content in different “views”
  - Directories are for the editors, allow to organise objects in digital library in a way similar to directories in the filesystem

Functionality overview

- Descriptive metadata schema can be configured
  - There is one metadata schema definition per digital library
  - The schema consists of elements which can be grouped into hierarchical structure
  - Each object can be described with 0 or more values of each metadata elements
  - In general values are treated as text (no typed values)
  - There are automatically created dictionaries for each metadata element
  - dLibra is able to store distinctive dictionaries for different languages
- Administrative metadata is defined by dLibra features
- Structural metadata is defined by dLibra digital object's data model

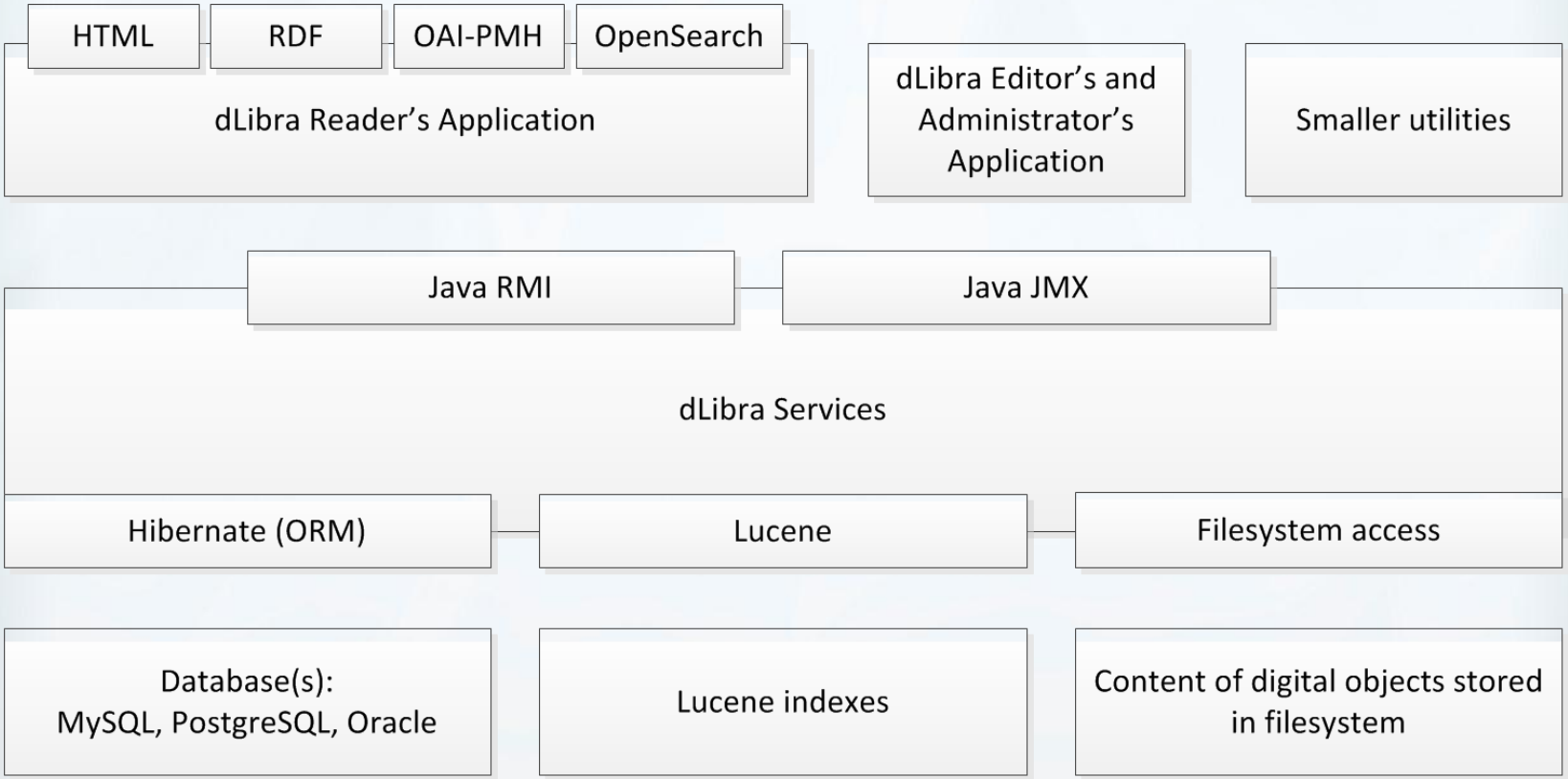
Digital object's data model



Functionality overview

- Content stored in dLibra undergoes periodic consistency checks (based on MD5 checksums)
- dLibra offers wide authentication and authorization features based on:
  - User data stored internally
  - User data stored in external user management systems (like LDAP)
  - IP/domain address of the user
- dLibra offers basic copy-protection features for selected digital formats (HTML, PDF, DjVu, JPEG etc.)
- There are many other features of dLibra connected with the cultural heritage context and implemented in end-user applications

High level architecture



Key technologies on services level

- Java 6
  - Java RMI
  - Java JMX
- Lucene
- Hibernate
  - In production environment we support three types of external databases: MySQL, PostgreSQL and Oracle
- Java Plug-in Framework
- HTTP (+XML) access to selected functionality is available via the Reader's Application

dLibra services architecture

- “dLibra server” is in fact a set of distributed services offering parts of the core dLibra functionality
- Each service can be deployed on different machine (VM) but this is just an option
- Services can be divided into two groups
  - Internal services
    - Independent from the dLibra functional services
    - Being a basis for the services system
  - Functional services
    - Utilise internal services
    - Offer the core digital library functionality

dLibra services architecture

- Internal services
  - System Service
    - A registry of all services in particular digital library
    - Responsible for services discovery
    - Responsible for services authentication
      - Each service is identified by IP address (+ TCP port), service type and password
  - Event Service
    - Responsible for asynchronous services communication
    - A service can register in Event Service to receive particular type(s) of events
    - A service can submit events to the Event Service
    - Events are organized into hierarchy facilitating the selective notifications
    - Types of events depend on the functionality of system utilising the Event Service (beside of few core events like “service connected”, “service disconnected”)

dLibra services architecture

- Event Service example
  - Indexing Service registers for any type of events related to modification of descriptive metadata
  - Metadata Service submits event of type “Edition metadata modified” containing the identifier of the modified edition
  - Event Service forwards this event to any interested service, including the Indexing Service
  - Indexing Service (using System Service) connects to Metadata Service, obtains new metadata of the modified edition and updates search indexes

dLibra services architecture

- Functional services
  - Metadata Service
  - Content Service
  - Indexing Service
  - Search Service
  - User Service
  - Profile Service
- If service provides a lot of functionality, it is divided into so called managers, grouping closely related groups of functions

dLibra services architecture

- Metadata Service
  - DirectoryManager – digital library directories
  - PublicationManager - publications
  - FileManager – files of publications
  - LibCollectionManager – digital library collections
  - AttributeManager – metadata schema management
  - AttributeValueManager – management of dictionaries of the metadata schema
  - ElementMetadataManager – management of metadata of digital library objects
  - LanguageManager – languages for metadata and for user interfaces
  - ReportManager – reports on the metadata

dLibra services architecture

- Content Service
  - Content storage/access
  - MD5 Checksums
  - Transformations
    - PDF and DjVu to JPEG in various configurations of output size and compression level
  - Compressions
    - Entire edition into single ZIP file

dLibra services architecture

- Indexing Service
  - Registers for events associated with modification (incl. creation and deletion) of content and metadata of various digital library objects
  - Maintains several indexes
    - Textual content
    - Descriptive metadata of editions in digital library schema
    - Descriptive metadata of publications in digital library schema
    - Descriptive metadata of editions in DMCS schema
    - Descriptive metadata of publications in DCMS schema
  - Exposes the index for Searching Service

dLibra services architecture

- Search Service
  - Periodically gets newest copy of indexes from indexing service
    - If services are deployed in one VM, they share the indexes without copying
  - Allows to search in indexes with all Lucene features
  - Supports queries combined on several indexes (e.g. content + publications metadata + editions metadata)

dLibra services architecture

- User Service
  - GroupManager – management of the groups of users
  - RightManager – authorization information management
  - UserManager – authentication information management

dLibra services architecture

- Profile Service
  - Stores personal profiles of users

Development process

- Three level product versioning
  - dLibra x.y.z (e.g. 4.0.24)
    - x – new functionality without downgrade possibility (improvements requiring the modification of data stored in the system)
    - y – new functionality with downgrade possibility (minor improvements)
    - z – only bug fixes
- Maven + Cruise control used for builds and integration
  - Two instances of Cruise Control
    - One for official distributions
    - One for continuous integration
- JIRA for issue management
- SVN for code versioning
  - Trunk for integration of the current stable version
  - Branches for bugfix versions
  - Branches for larger new functionalites
  - Tags for released versions
    - Both public versions and internal development milestones
- SCRUM for development process management
- Confluence for public documentation

Users community

- In 2004 we organized first “Digital libraries” workshop, which became an impulse for further dLibra deployments
- The workshop became an annual event with 40-60 participants each year
- In 2008 we have organized the first “Polish Digital Libraries” conference and it also became annual event with around 150 participants
- This workshop and conference became a meeting place for all people interested in the development of digital libraries, especially in the cultural heritage domain

Users community

- Those events formed the dLibra community
- The community consists mostly of librarians
- The community supports us by
  - Providing interface translations
  - Providing feature requests/bug reports
  - Providing basic support for other community members
  - Providing small tools around dLibra (e.g. query log analyser)
- The dLibra community initiated the creation of Polish “Library 2.0” community

Licensing

- dLibra as a whole is not a free system
  - License for one dLibra instance in particular major version, without any additional limitations, costs around 300 EUR / 260 GBP
  - It is several hundred times less than comparable commercial products
  - But it is not free… because of the tax system and the fact that it is co-funded from the PSNC internal budget

Works planned for 2011

- Development in the cultural heritage context
  - Support for full digital library workflow from digitisation to on-line publishing and archival storage
- Development in the context of our other activities
  - Upgrade of (communication) technologies to allow wider adoption of dLibra services
  - Stronger separation of services
  - Support for Semantic Web technologies in the descriptive metadata area

POZNAŃ SUPERCOMPUTING AND NETWORKING CENTER

Thank you for your attention!

Visit us at http://dl.psnc.pl/

Poznań Supercomputing and Networking Center
affiliated to the Institute of Bioorganic Chemistry of the Polish Academy of Sciences,
ul. Noskowskiego 12/14, 61-704 Poznań, POLAND,
Office: phone center: (+48 61) 858-20-00,
fax: (+48 61) 852-59-54,
e-mail: office@man.poznan.pl, http://www.man.poznan.pl