Access IT Training

# How to set up a metadata aggregator

# Metadata aggregators

- According to the present version of Europeana Outline Functional Specification tasks for the aggregator are:
  1. To gather the information about content providers and their information systems
  2. To gather the metadata of objects that should be visible in Europeana
  3. To remove duplicates, clean-up the metadata, normalize it and enrich
  4. To confirm the accessibility of digital objects
  5. To expose the aggregated metadata for Europeana via the OAI-PMH protocol

http://dev.europeana.eu/public_documents/EDLnet%20D2.5_Outline_Functional_Specifications20090301_version%201.7_consWithoutHistory_lossless.pdf

# Available tools - aggregators

- OAICat
- RepoX
- Celestial (Perl)
- Other tools
  - OAIbiblio (PHP)
  - Ruby-oai (Ruby)
  - pyoai (Python)

# Available tools - aggregators

- List of tools comes from Julie Verleyen presentation : „Metadata Harvesting”
  - http://www.europeanalocal.eu/eng/content/download/2808/32628/version/1/file/KSW_13-01-2009_Julie_METADATA_HARVESTING.ppt

# Comparision framework

- OAI-PMH 2.0 compliance
- ESE compliance
- OAI-PMH implementation
  - Deleted records support
  - Selective harvesting capabilities
  - Incremental harvesting support
- License/Price
- Support
- Content providers information gathering aids
- Other services based on aggregated metadata
- Underlying technology
- Market share

# OAICat

- OAICat is developed by OCLC
  - http://www.oclc.org/research/software/oai/cat.htm
  - http://alcme.oclc.org/wikid/CollectionOaiCat:FrontPage
- *„OAICat was written as **open source** and includes a number of abstractions that allow it to be **customized** and **configured** for use with a variety of data sources."*
- Framework complaint with OAI-PMH v2.0
- License: Apache Software Lic. V2.0
- Java based application

# OAICat

- OAICat is included in the **Dspace** distribution
- According to the UIUC OAI-PMH registry, OAICat is used in **532** of **2242** known OAI-PMH repositories
  - http://oai.grainger.uiuc.edu/registry/ListToolkits.asp

# OAICat

- Supports  OAI sets, resumptionToken and deleted records
- It can expose any metadata scheme
- It can be also used to create aggregator
  - By addition of an OAI-PMH harvester

# RepoX

- REPOX - A Metadata Space Manager
  - http://repox.ist.utl.pt/
- REPOX  allows to:
  - Aggregate metadata from various sources
  - Expose aggregated metadata through OAI-PMH interface
- License: GPLv2
- Java standalone application with web GUI

# RepoX

- Multiple harvesting jobs, Scheduler
- Basic statistics
- Management of XML metadata repository
  - Versioning and identification of records
  - Different metadata format
  - User interface to create metadata crosswalks: Schema mapper
- OAICat + oaiharvester2 (OCLC) based
- Supports :
  - OAI sets, deleted records and resumption tokens

# Celestial

- Celestial (Perl-based)
  - http://sourceforge.net/projects/oai-perl
  - OAI aggregator/cache application that imports OAI metadata from version 1.0,1.1,2.0
  - License: GPLv2
  - Allows to re-expose that metadata through OAI-PMH 2.0 interface
  - Default configuration supports only OAI-DC format but it can  be adjusted

# Celestial

- It supports OAI sets, deleted records, resumption token
- It is used in Eprints platform
- Celestial requires:
  - oai-perl v2, MySQL, Perl 5.6.x and a CGI-capable web server

# DLF aggregation platform

- PIONIER Digital Libraries Federation
  - http://fbc.pionier.net.pl/
- Free software package which can be used to create aggregator – work in progress
- Java-based aggregation platform
- Information about content providers, harvesting statistics
- Duplicates detection, coordination of digitization

# DLF aggregation platform

- Information about content providers
- Statistics
- Duplicates detection, coordination of digitization,
- Dynamic OAI-PMH sets  support
- Allows to expose DC or ESE

# DLF aggregation platform

- Supports OAI 2.0, sets (static and dynamic), incremental harvesting, resumption token, deleted records
- DLF is a Polish national aggregator

# OAI-PMH interface validation

- How to check OAI-PMH compliance?
- OAIRepository Explorer
  - http://re.cs.uct.ac.za/
  - Website which allows to check correctness of OAI-PMH interface
- Europeana Content Checker Ingestor
  - http://contentchecker.isti.cnr.it:8080/portal/
  - Documentation is available at:
    - http://europeanalocal.avinet.no/viewtopic.php?f=5&t=22

# Basic requirements

- Each DLF content provider have to fulfill some basic requirements
  - Need to have a valid OAI-PMH 2.0 interface
    - Validation using OAI Repository Explorer
  - Register and validate repository at:
    - http://www.openarchives.org/Register/BrowseSites
- Usually we don't allow ports different than 80
  - ~~http://man.poznan.pl:8080/oai/~~
- Repository must have a domain address
  - ~~http://192.168.128.1/oai~~

# Basic requirements

- At the moment content providers don't have to sign any agreements
- Europeana Office is working on a formal agreement which will formalize cooperation between Europeana Office and aggregators
- Aggregator takes all the responsibility for content which is submitted to Europeana
  - In the boundaries described in an agreement

# Common pitfals

- Repository is down/unavailable
- Metadata contains characters which are not allowed in XML
- XML syntax errors
- Short lifetime of resumptionToken

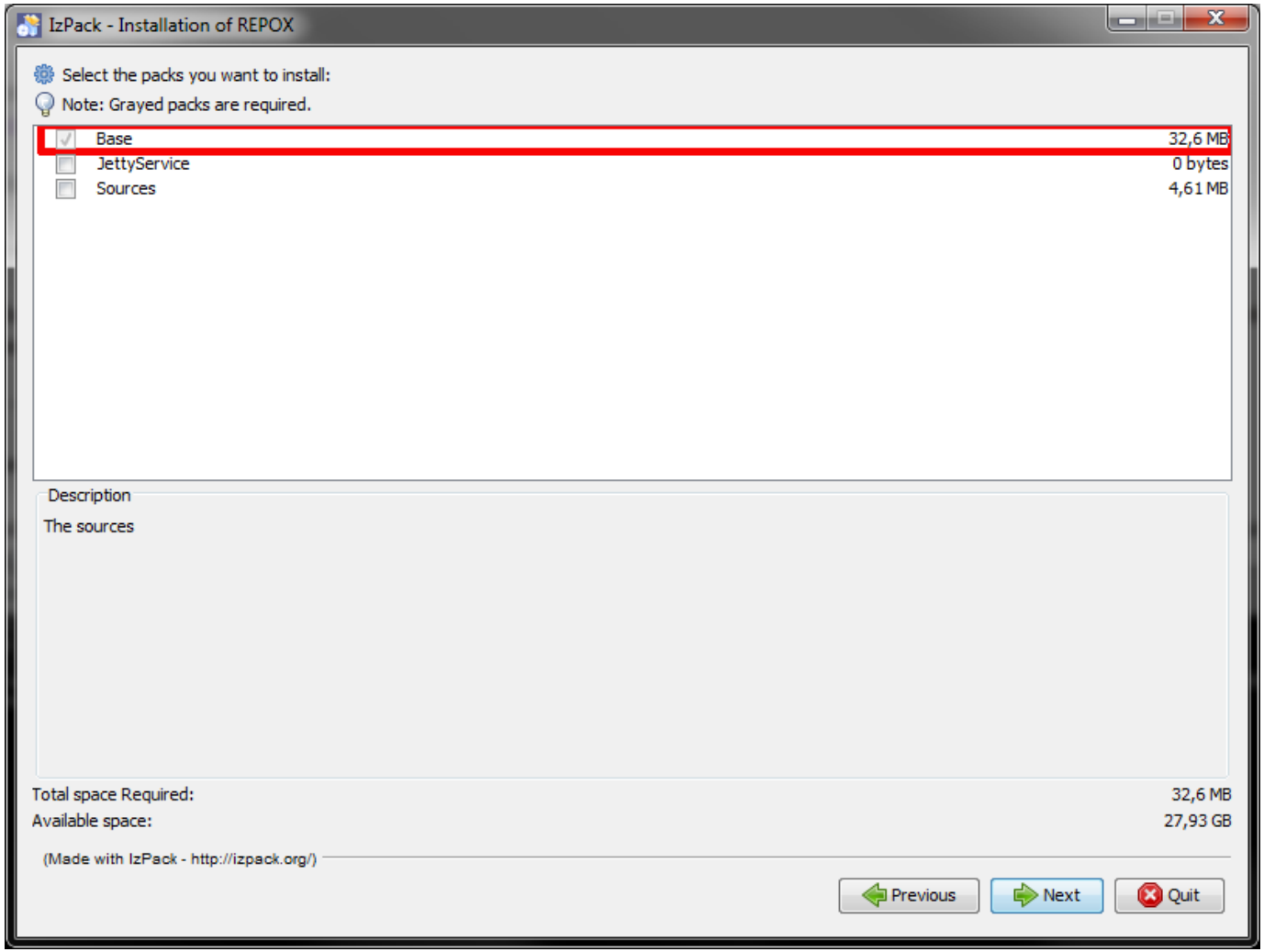
# Short demo

- How to setup a metadata aggregator using
  - RepoX
  - DLF aggregation platform

# RepoX - installation

- Ensure that you have Java 6 installed
- Download RepoX 1.4.3
  - http://repox.ist.utl.pt/REPOX_1.4.3-installer.jar
- Installation
  - Run : `java -jar REPOX_1.4.3-installer.jar`
  - Don't check install "Jetty Service"
  - Don't install RepoX in path with whitespace
    - e.g~~. c:\Program Files\repox~~
    - Use e.g. **c:\repox**

# RepoX - installation



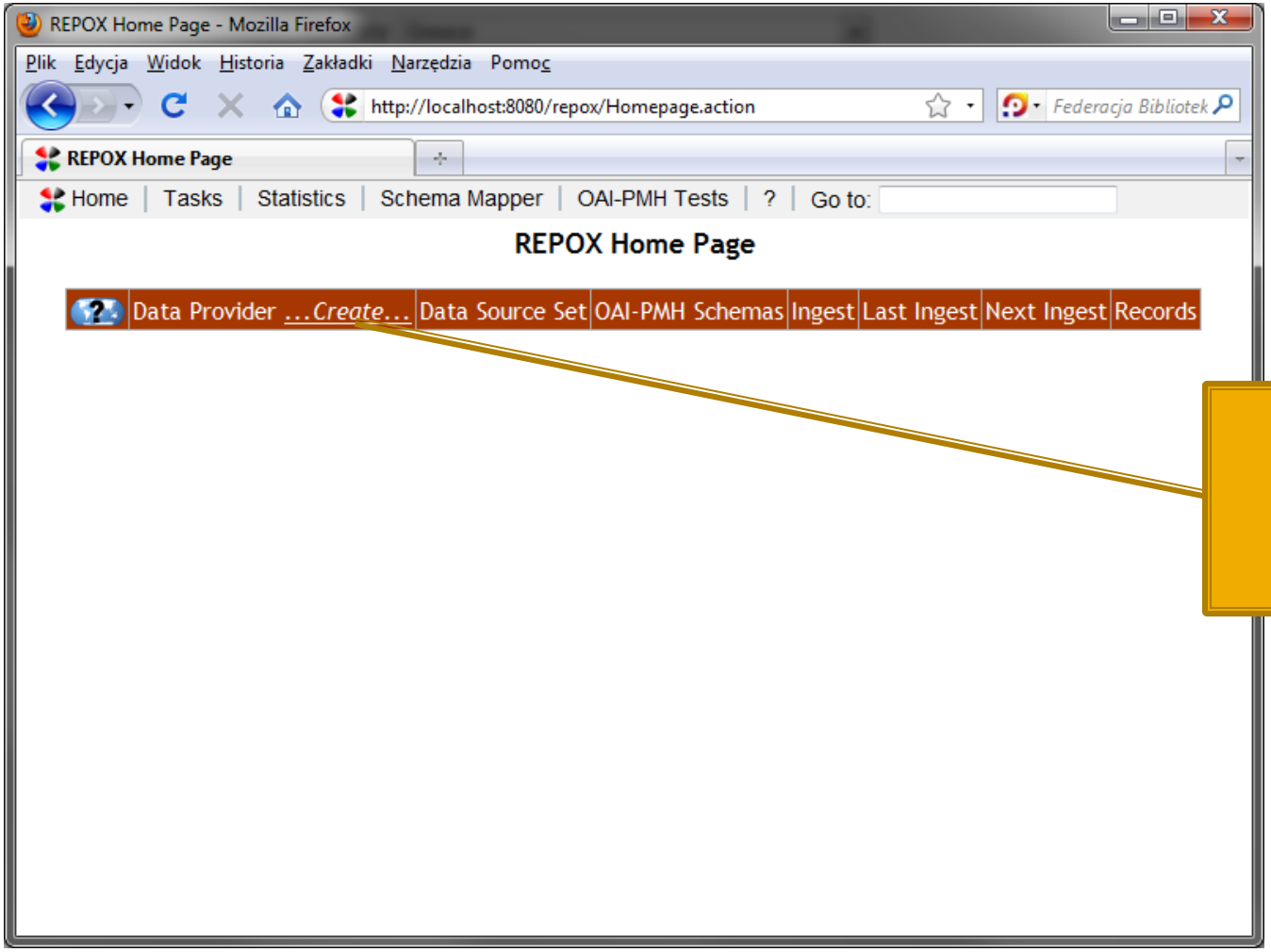
# RepoX - installation

- After installation go to RepoX folder run terminal and type in:
  - `cd jetty`
  - `java -jar start.jar`
- Now go to browser an type in: http://localhost:8080/repox
- For Linux server use dedicated installer from RepoX website

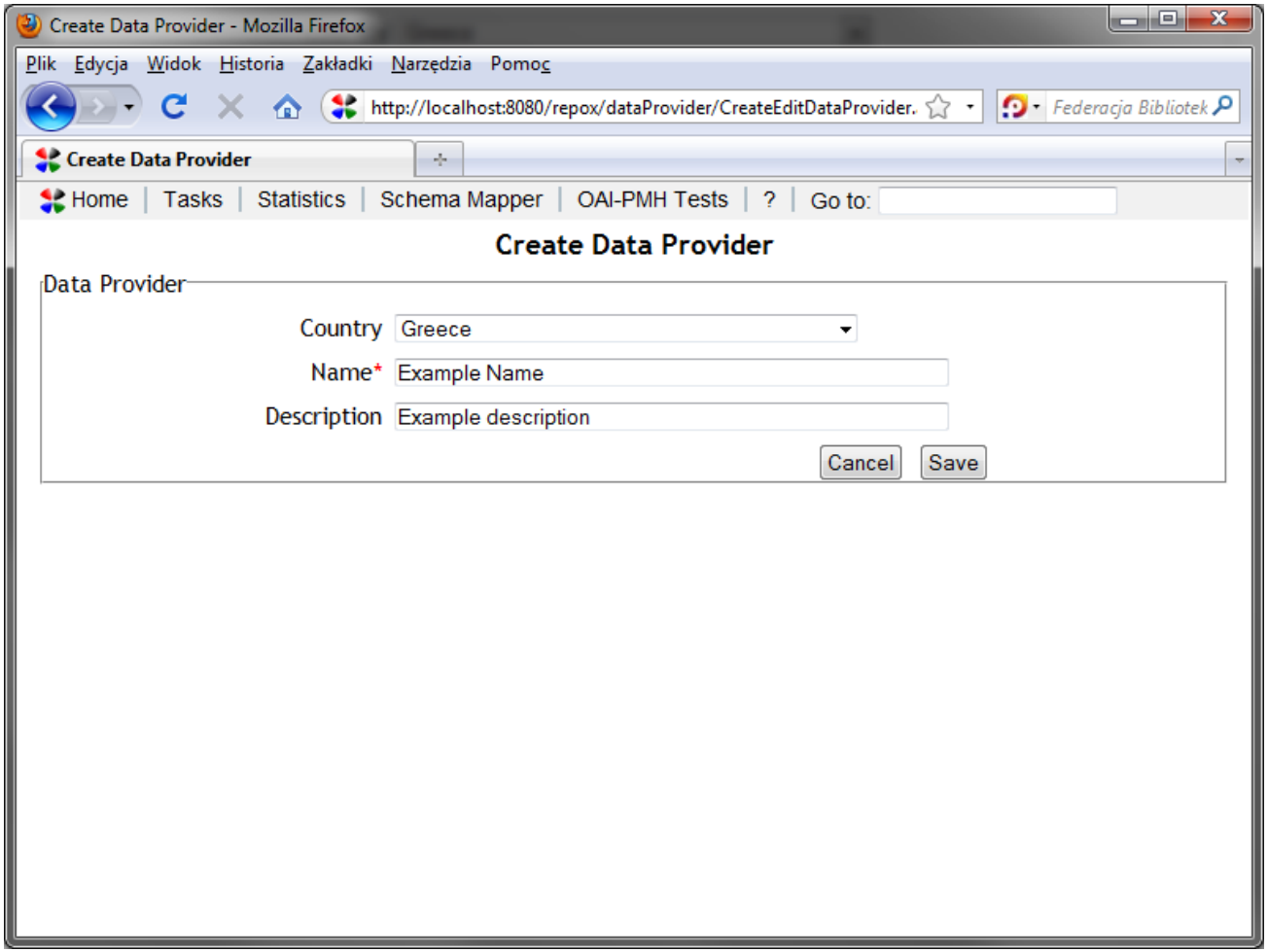
# RepoX - basic usage

- RepoX documentation is available here:
  - http://repox.ist.utl.pt/doc/usingrepox.html
- First login, than create a Data Provider and save

# RepoX - basic usage



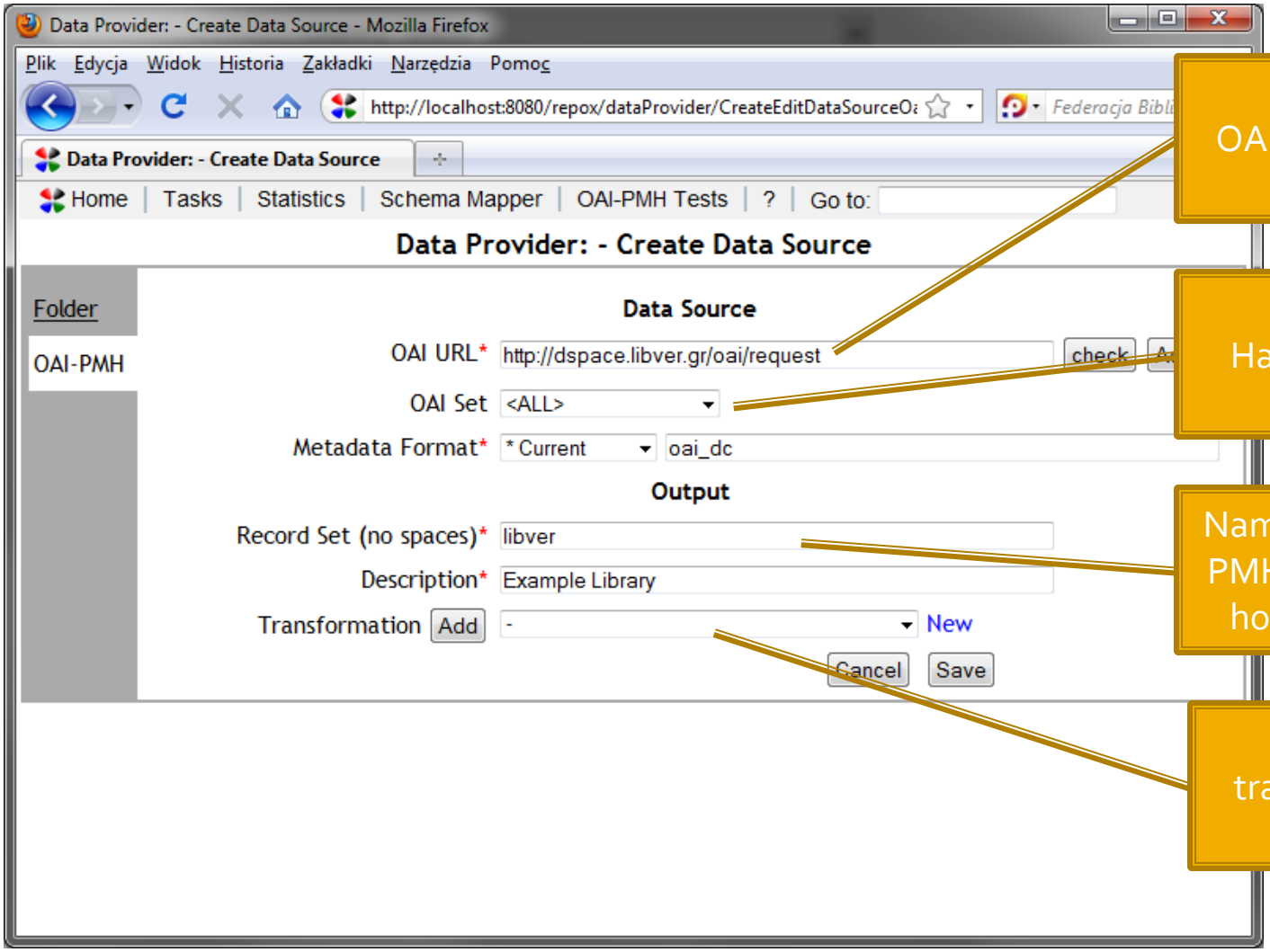
# RepoX - basic usage



# RepoX - basic usage

- Create a Data Source for given Data Provider by typing in:
  - OAI-PMH repository URL
  - Specifying harvested set name
  - Used metadata format
  - Name of set under which harvested data would be available in RepoX OAI-PMH interface
  - Applying appropriate transformations

# RepoX - basic usage



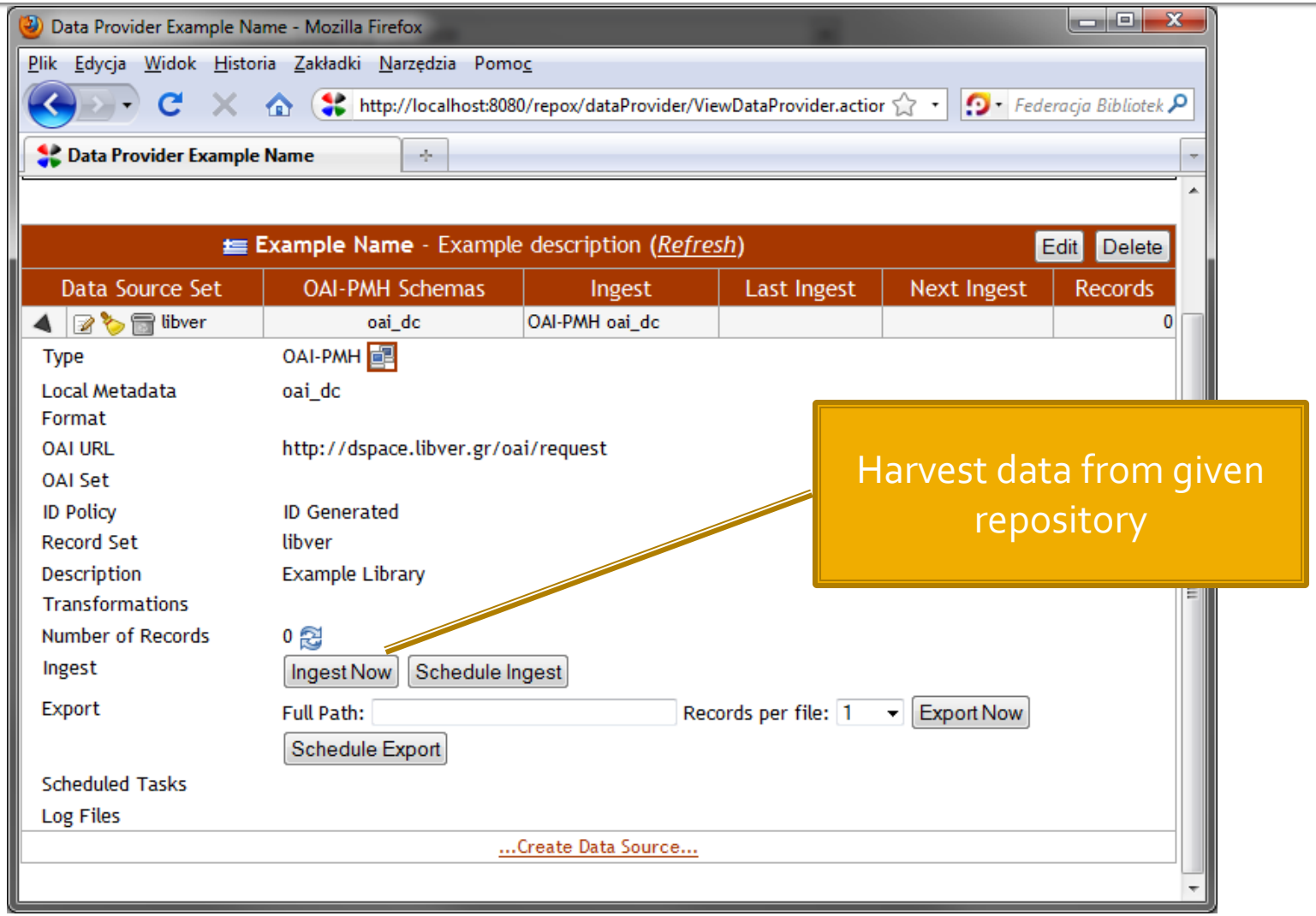

OAI-PMH repository URL

Harvested set selection

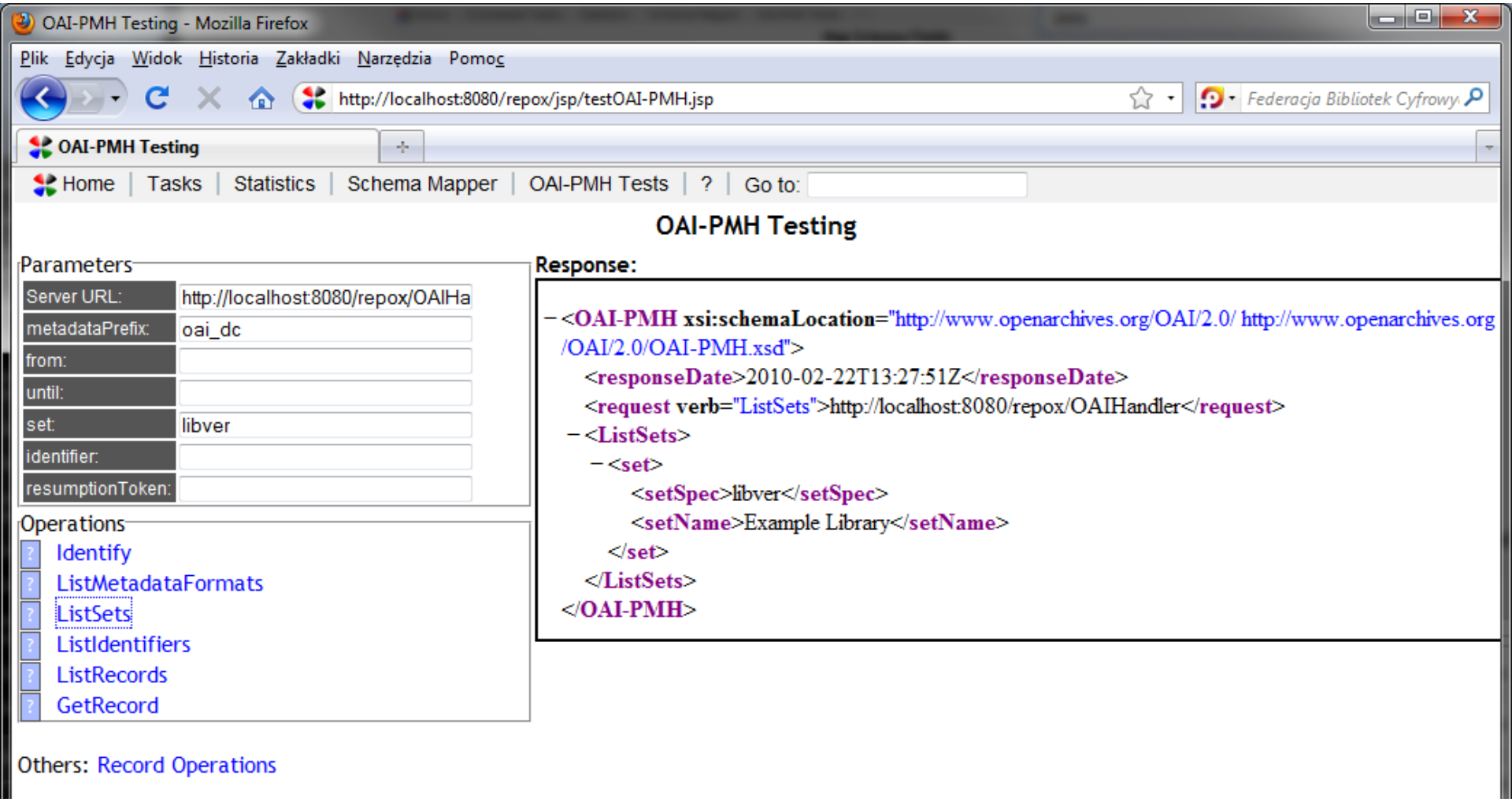
Name of set in RepoX OAI-PMH interface. This would hold all harvested data.

Choose metadata transformation or define new one

# RepoX - basic usage

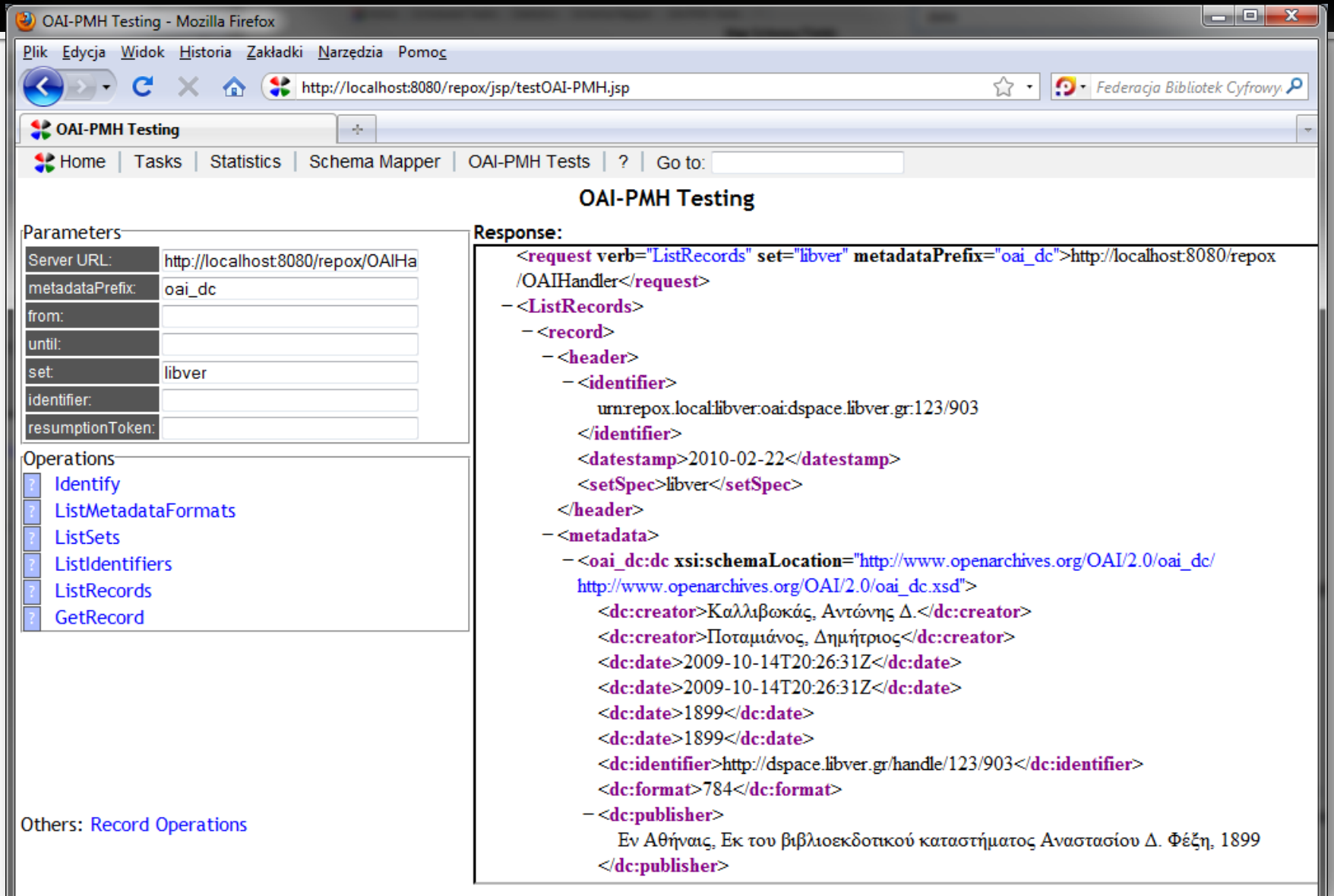


# RepoX - basic usage




Harvesting results validation

# RepoX - basic usage



# Best practices

- „DRIVER Guidelines for Content Providers"
  - http://www.driver-repository.eu/DRIVER-Guidelines.html
- TELplus D-2.1: „OAI-PMH implementation and tools guidelines"
  - http://www.theeuropeanlibrary.org/portal/organisation/cooperation/telplus/documents/TELplus_D2.1_31052008.pdf

# Best practices

- „Best Practices for OAI Data Provider Implementations and Shareable Metadata"
  - http://webservices.itcs.umich.edu/mediawiki/oaibp/index.php/Main_Page
- „Guidelines for Repository Implementers"
  - http://www.openarchives.org/OAI/2.0/guidelines-repository.htm

# Conclusions

- Good news
  - Some tools are available so there is no need to implement everything from scratch ☺
- Bad news
  - Some technical knowledge is required
    - Tomcat, XML, XSL

# Q&A

- *EuropeanaLocal technical forum*
  - *http://europeanalocal.avinet.no*