OAI Overview

OAI Workshop

Nuno Freire
(nuno.freire@bn.pt)

What is the OAI?

- What is the Open Archive Initiative (OAI)?
  - Organization dedicated to solving problems of digital library interoperability by defining simple protocols and standards
  - Grew out of the e-prints (arXiv) community at Los Alamos
- What is the OAI Protocol for Metadata Harvesting (OAI-PMH)?
  - Protocol to transfer metadata from a source archive to a destination archive
- How is the OAI-PMH Being Used?
  - The OAI-PMH has been adopted in several projects as a primary means of gathering and sharing metadata among contributors

OAI divides the world between data providers and service providers

Data Providers vs. Service Providers

- Data Providers refer to entities who possess metadata and are willing to share this with others (e.g. collection builders)
- Service Providers are entities who harvest data from Data Providers in order to provide higher-level services to users (e.g. searching, browsing, recommender systems, etc.).
- Some systems may act as both…

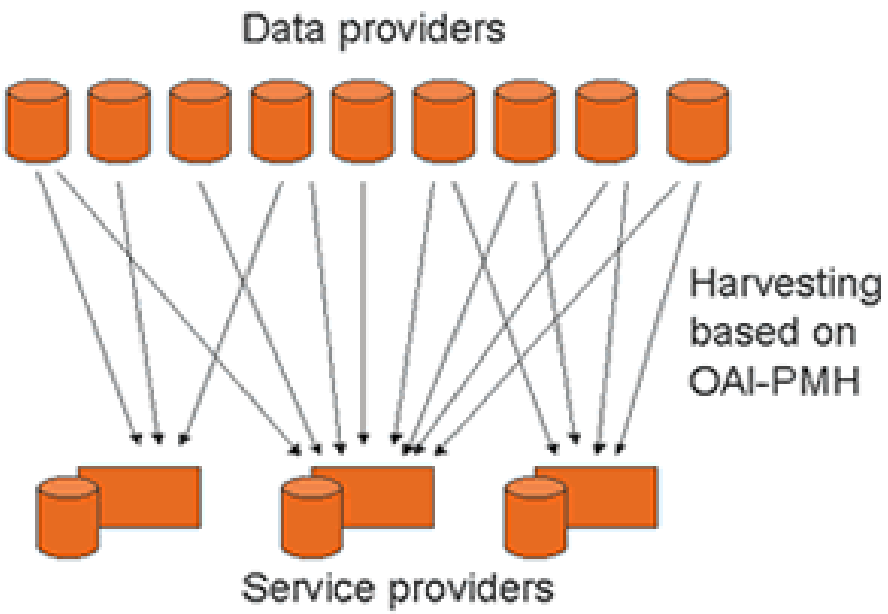
Data Providers vs. Service Providers

- Data Providers deploy an OAI-PMH compliant **repository** (also called archive)
- Service Providers deploy an OAI-PMH **harvester**

...

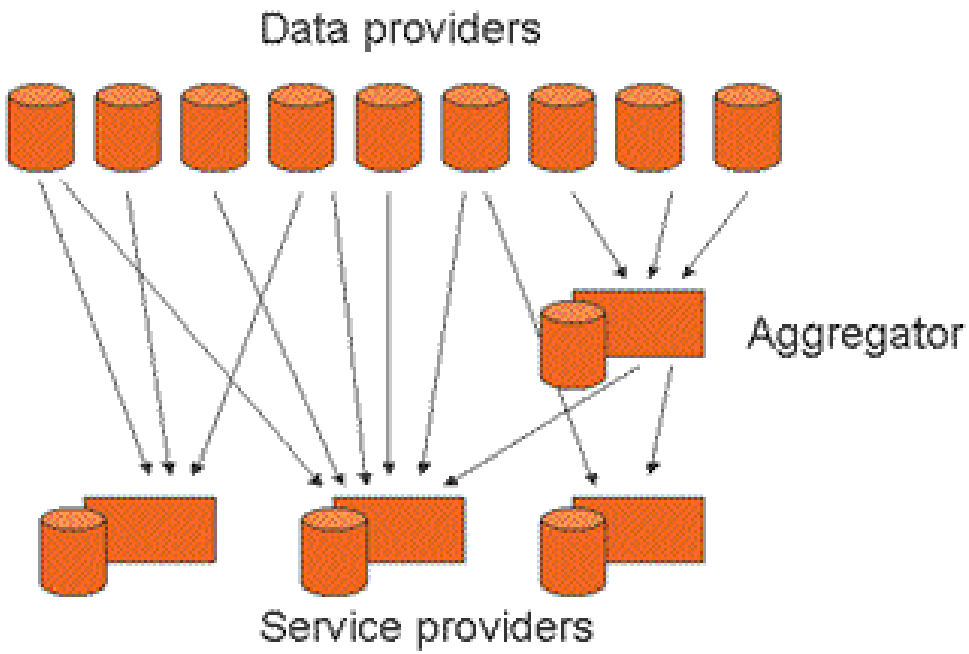
Flexible deployment

Multiple Service Providers



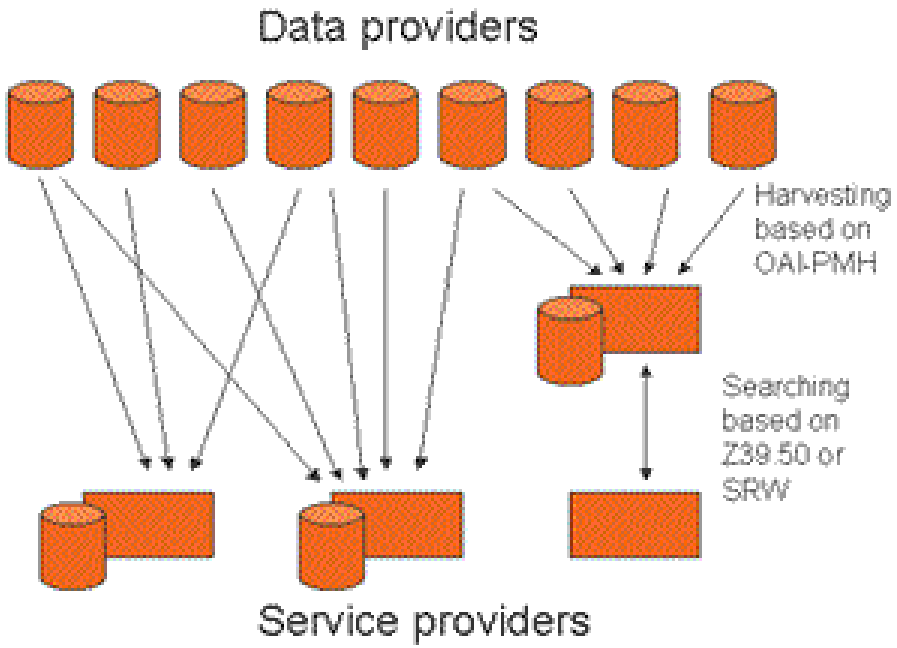
Flexible deployment

Aggregators

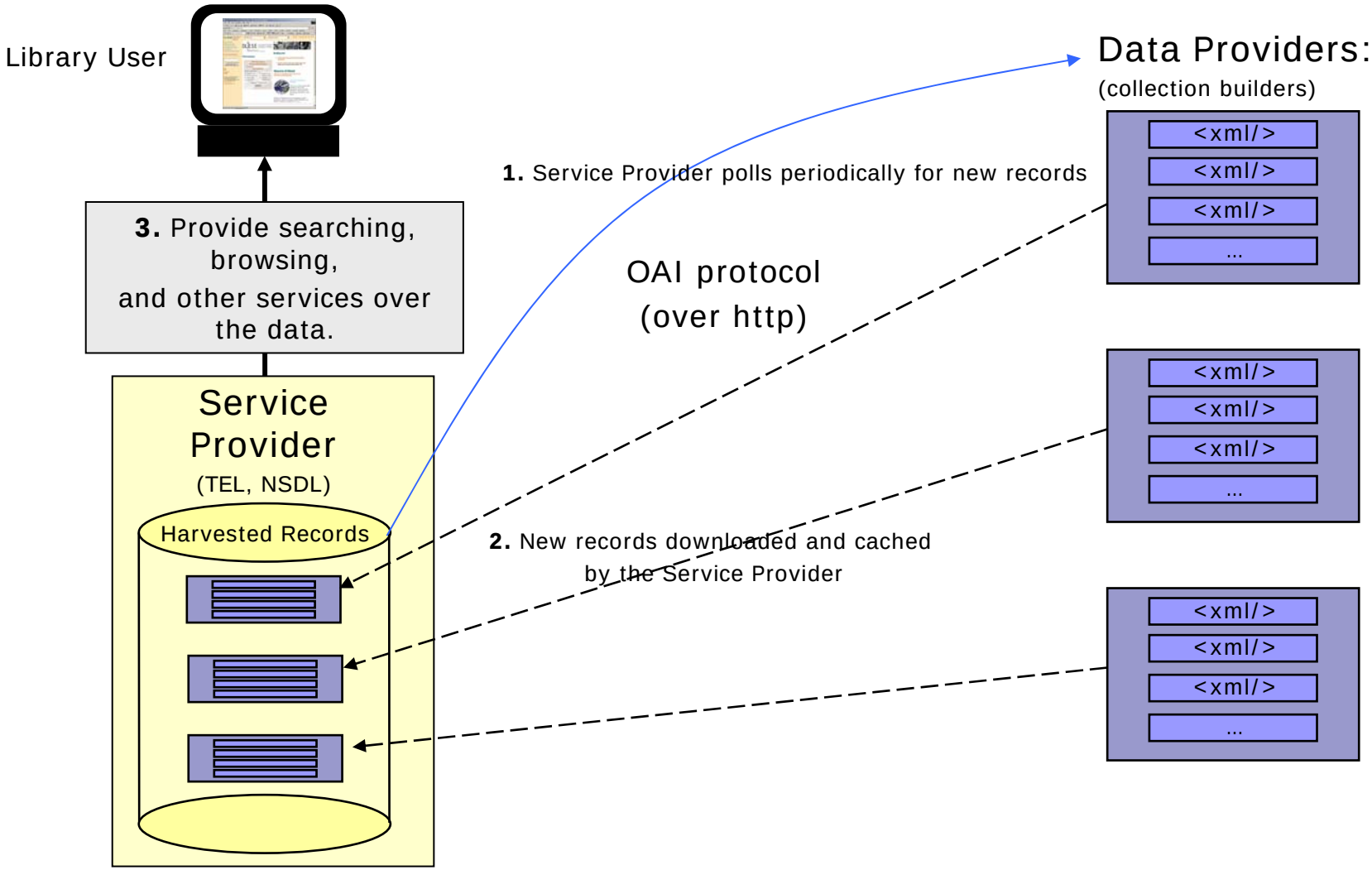


Flexible deployment

Harvesting combined with searching



Metadata Harvesting Framework



Multiple representations of an object

MARC Record

In XML

Dublin Core Record

In XML

Qualified Dublin Core Record

In XML

MODS record

In XML

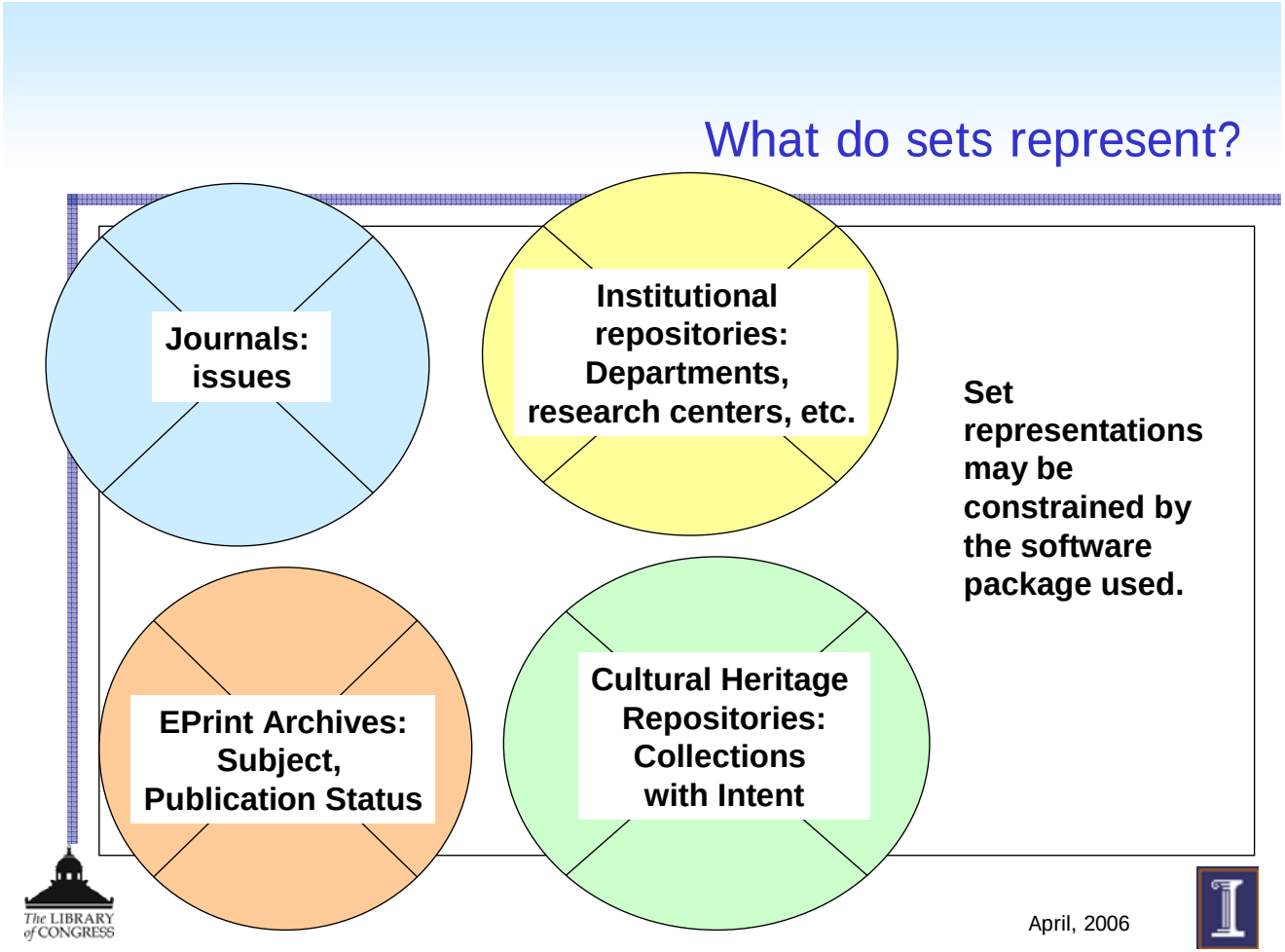
TEL AP Record

In XML

OAI repositories can be organized in sets

- OAI-PMH mechanism to allow for harvesting of sub-collections
- Semantics for sets are defined outside of the protocol
- Sets are defined by conventions established between data and service providers, or just by the data provider

OAI repositories can be organized in sets



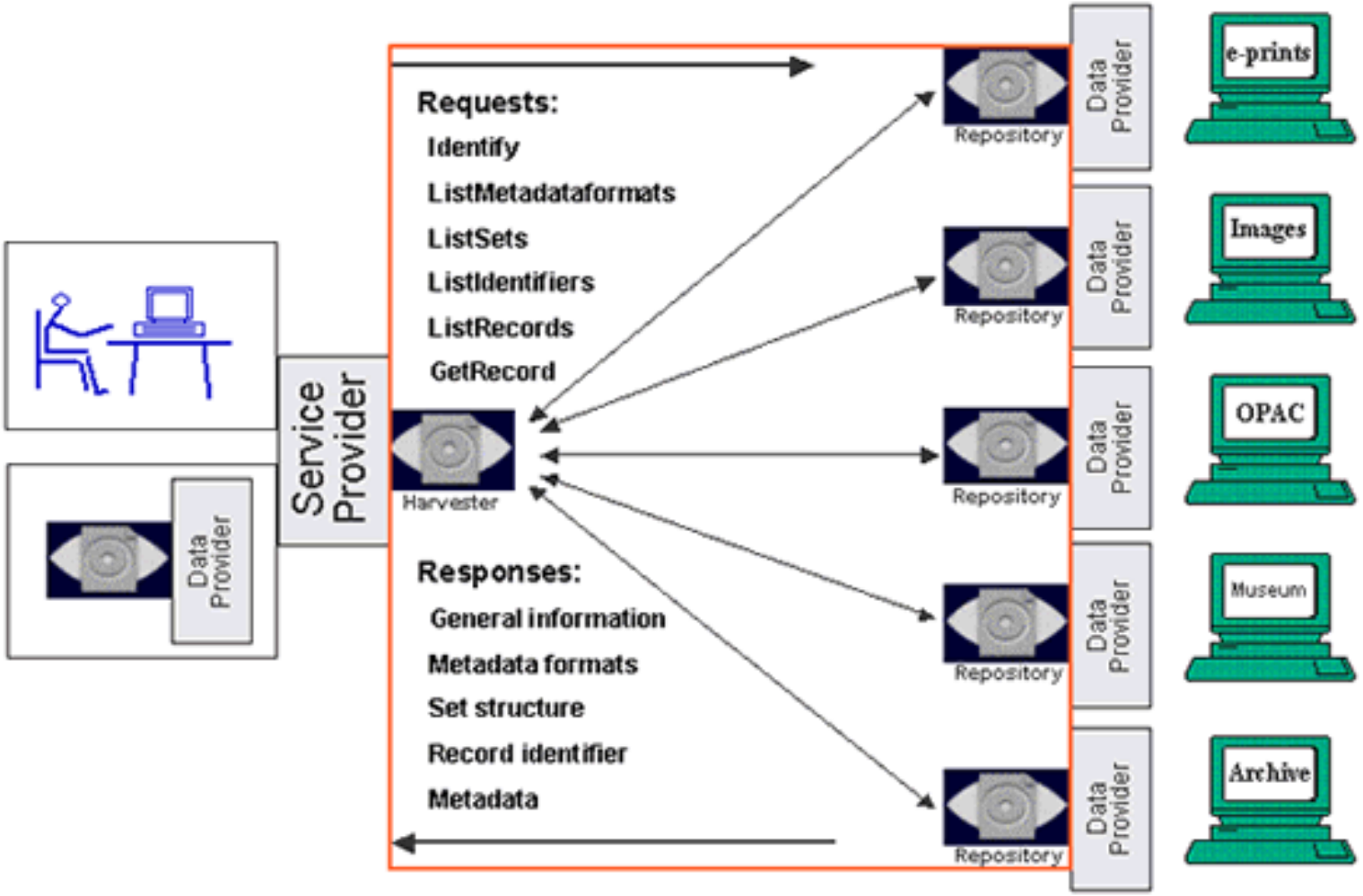
Requirements to be a Data Provider

- Source of metadata
  - Human or automated resource catalogers
- Metadata mappings
  - Crosswalks from native formats to DC or other formats
- Server technology
  - Handled by the OAI software
- Datestamps
  - Indicates when the item was last changed (handled by the OAI software)
- Deletions
  - Indicates if the item has been deleted and should be removed (handled by the OAI software)
- Unique identifiers
  - Used to uniquely identify each item across repositories

The OAI-PMH

- OAI-PMH Requests
  - Identify
  - ListMetadataFormats
  - ListSets
  - GetRecord
  - ListIdentifiers
  - ListRecords
- Resumption Tokens
  - Used for flow control when large responses are required

OAI-PMH: overview and structure model



Implementing OAI-PMH

OAI Workshop

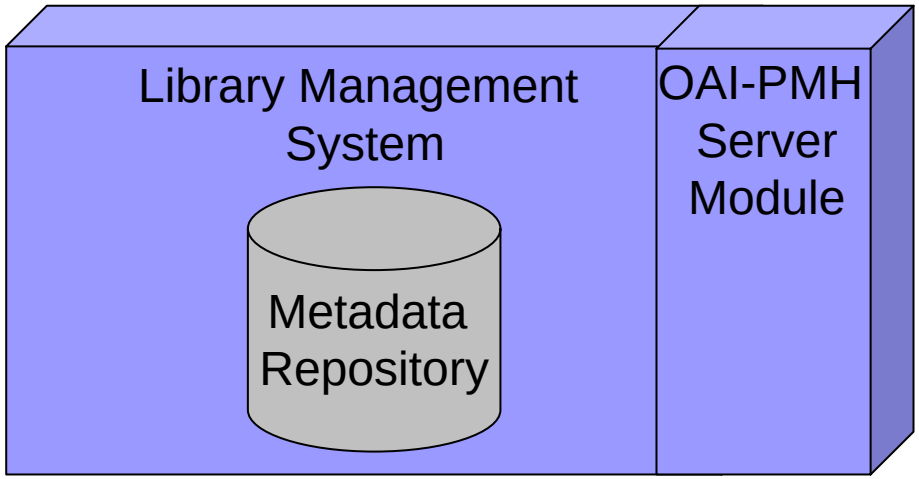
Nuno Freire
(nuno.freire@bn.pt)

The three most common scenarios of OAI-PMH implementation

- Library Management System (LMS)
  OAI- module provided by the vendor
- In-house OAI-PMH server development
- Standalone OAI server

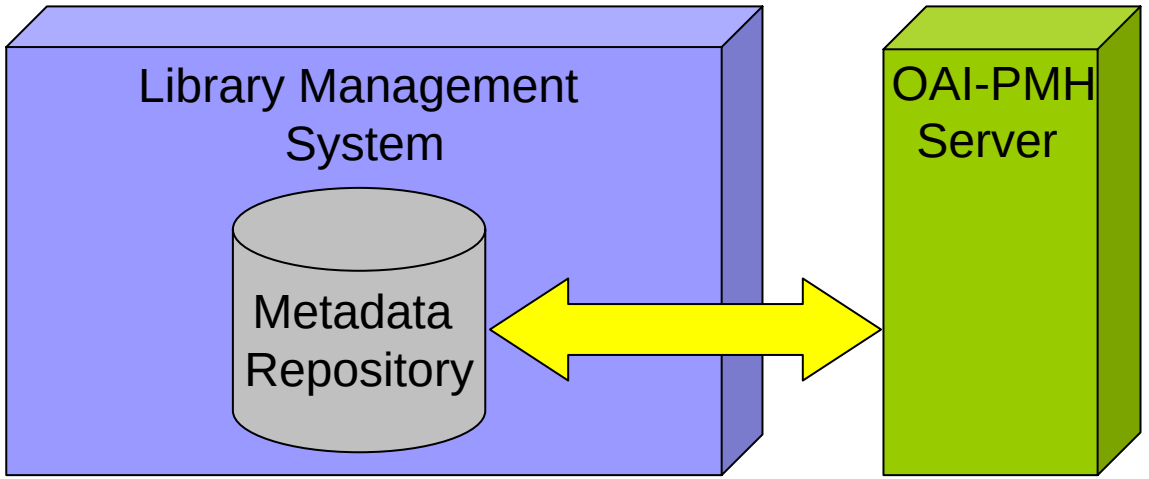
The three most common scenarios of OAI-PMH implementation

- Library Management System (LMS)
  OAI- module provided by the vendor



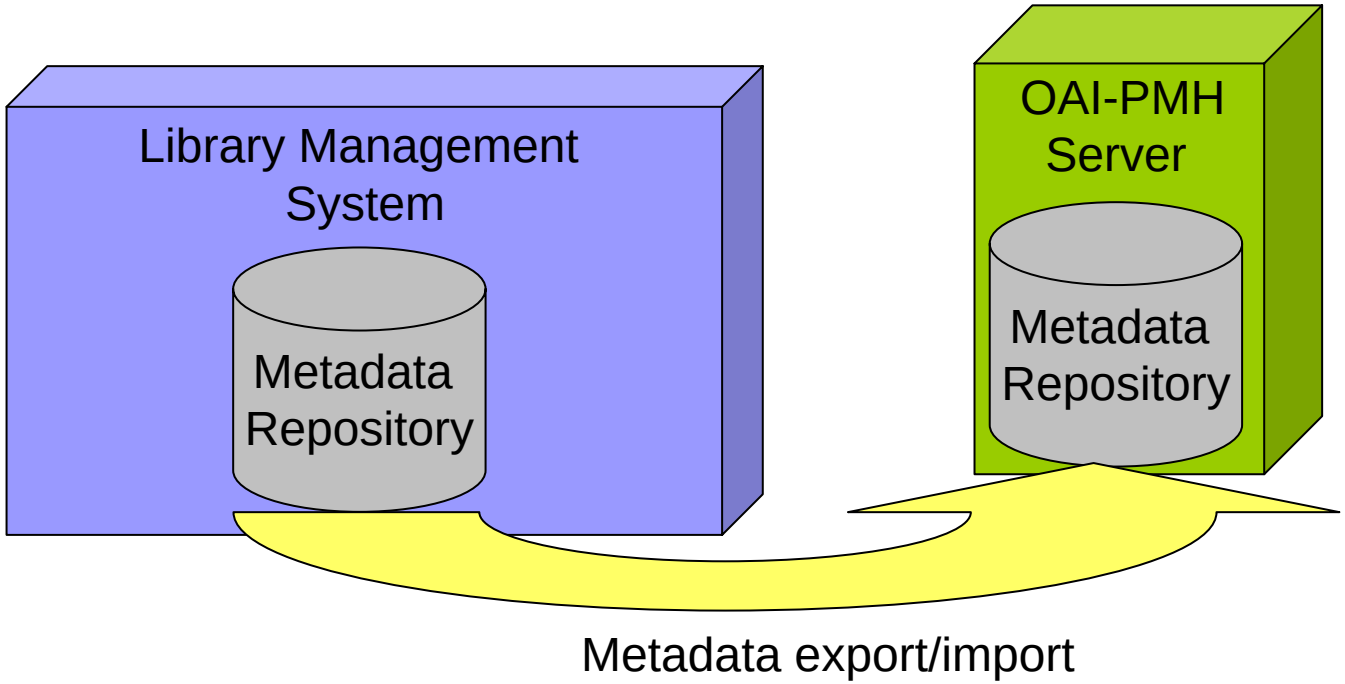
The three most common scenarios of OAI-PMH implementation

- In-house OAI-PMH server development



The three most common scenarios of OAI-PMH implementation

- Standalone OAI-PMH server



The three most common scenarios of OAI-PMH implementation

- Metadata crosswalks
  - All three scenarios require metadata to be converted to other formats:
    - To the TEL application profile
      - Recommended for sharing data with TEL
    - To simple Dublin Core
      - Minimum requirements for TEL
      - Useful for sharing your data with other service providers

Library Management System OAI module provided by the vendor

- Main tasks
  - Inquire the vendor for the availability of the module
    - Many vendors don’t have an OAI module for their products
  - The OAI module should support metadata crosswalks
    - Required for sharing your data according to the TEL application profile
  - Develop metadata crosswalk specifications for your data format
    - The vendor will implement them on the OAI module

Library Management System OAI module provided by the vendor

- Advantages
  - Data available by OAI-PMH is always up to date
  - No in-house developers needed
  - Low maintenance costs
- Disadvantages
  - No all vendors offer an OAI module for their products
  - Costs of the software

In-house OAI-PMH server development

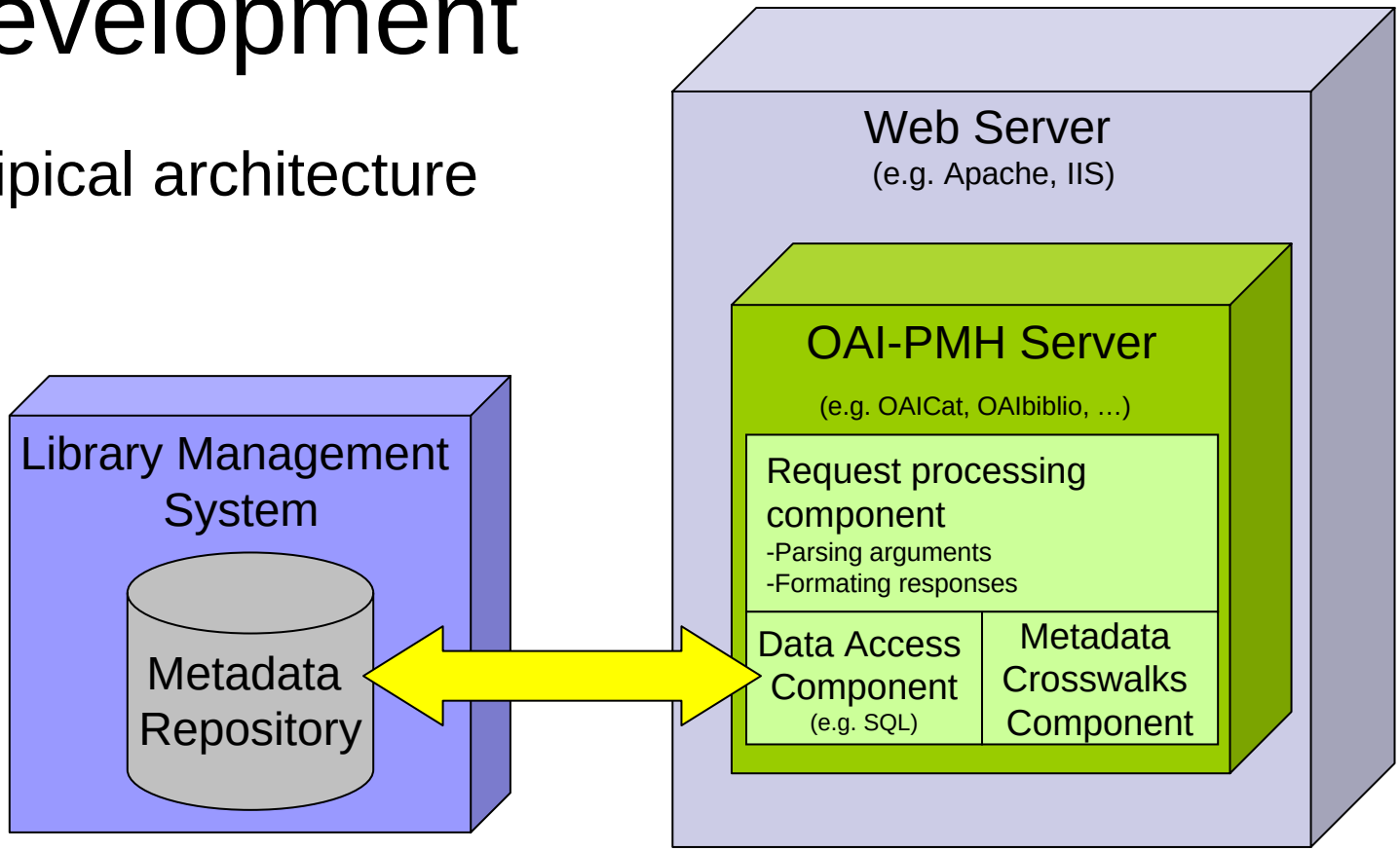
- Main tasks
  - Check with your IT team to see if access to the metadata repository of the LMS is possible.
    - Check also which programming languages can be used
  - Check that the LMS has unique identifiers and date stamps on the records
  - Choose an open source OAI-PMH implementation
    - Taking in consideration the programming language to use, based also on the knowledge of the software developer
  - Develop metadata crosswalk specifications for your format
  - Test the server

In-house OAI-PMH server development

- Choose an open source OAI-PMH implementation
  - It's not necessary to develop your own OAI software from start
    - a number of software tools are available, many of them under open source license (or similar) terms.
  - The OAI maintains a list of OAI software tools at http://www.openarchives.org/tools/
  - Among the many tools available, you will find general OAI-PMH implementations in several languages to plug-in into your library management system
    - OCLCs Java implementation
    - Perl implementations
    - OAIbiblio PHP Data Provider
    - others…

In-house OAI-PMH server development

Tipical architecture



Data Provider: prerequisites

- Other implementation issues
  - Unique identifiers for each item
  - Organization of records in sets
  - Keeping information on deleted records
  - Flow control with resumption tokens
- References:
  - The Open Archives Forum online tutorial section 4 - Implementing OAI-PMH http://www.oaforum.org/tutorial/english/page4.htm
  - OAI-PMH v2.0 specification: http://www.openarchives.org/OAI/openarchivesprotocol.html

In-house OAI-PMH server development

- Advantages
  - Data available by OAI-PMH is always up to date
  - Requires only free open source software
  - Low maintenance costs
  - A good option for those libraries who are using an in-house developed LMS or repository
- Disadvantages
  - No all libraries have personnel with the required knowledge
    - Outsourcing is possible but may be hard to find companies with OAI experience
  - May be the most time consuming scenario
    - although in some cases this option may be done in just 2-5 days work

Standalone OAI-PMH server

- Main tasks
  - Choose an open source standalone OAI-PMH server
  - Install and configure the server
  - Develop metadata crosswalk specifications for your format
  - Implement an procedure in the library to transfer the metadata records from the LMS to the OAI-PMH server
  - Test the server

Standalone OAI-PMH server

- Choosing an open source standalone OAI-PMH server
  - REPOX
    - Soon to be made available
    - Integrating special features for TEL members (specially crosswalks to TEL-AP)
  - DLESE OAI software http://www.dlese.org/Metadata/tool/index.htm
    - Very easy to deploy
  - OCLC OAICat http://www.oclc.org/research/software/oai/cat.htm
    - Not so easy and flexible to deploy as DLESE
  - More servers at http://www.openarchives.org/tools/

Standalone OAI-PMH server

- Implementing an procedure in the library to transfer the metadata records from the LMS to the OAI-PMH server
  - Export metadata records from the LMS
    - Typically in ISO2709 or MarcXchange
    - If possible, just those records that changed since the last update
  - Execute the metadata conversion
    - Generating xml records in TEL and DC
  - Import the records into the OAI-Server
- This process should be automated as much as possible
  - Ideally it should be fully automatic

Standalone OAI-PMH server

- Advantages
  - Very easy to deploy
    - It may take just a few hours
  - Requires only free open source software
  - No in-house developers needed
- Disadvantages
  - Highest maintenance of all scenarios
    - In those cases where the metadata transfer can’t be fully automated

Other scenarios

- In certain scenarios other kind of tools may be used.
- When building a repository that serves more purposes than just OAI, some full featured repository software may be used.
  - If you are setting up an e-print archive then you may want to consider using the GNU EPrints software package,
  - DSpace provides a digital asset management framework that includes preservation considerations

Testing

- When you think your implementation is ready to run, create some OAI-PMH requests, send them to your OAI interface and check the results.
- You can use the Repository Explorer at Vermont University to do this (http://oai.dlib.vt.edu/cgi-bin/Explorer/oai2.0/testoai/) by browsing through your repository

Registration

- Once your implementation is working as expected, you can register it at the official registration site for Data Providers.
  (http://www.openarchives.org/data/registerasprovider.html)
  - You provide the base URL for your Data Provider implementation.
  - OAI then performs an extensive conformance test (including tests for error conditions, among others
  - information on incorrect behaviour (if any) that was found will be notified to you
  - In the case of conformance, your Data Provider implementation will be added to the official list.
  - OAI performs regular checks on registered Data Providers to confirm that all is well.