Coordination Seminar of EMWIS
NATIONAL FOCAL POINTS

Collecting Metadata for the bibliographical references

Luis Mosteiro Ramírez.

CEDEX. Mº Fomento de España

Technical Unit of EMWIS

Amman (Jordan)

September 28, 2005

SEMIDE EMWIS

What the OAI-PMH?

- Open Archive Initiative-Protocol for Metadata Harvesting
- Archive
- Opened

to promote and facilitate the interoperability standards for the effective diffusion of the contents on the Internet

SEMIDE EMWIS

History

- arXiv.org
- CogPrints
- Networked Computer Science Technical Library Reference
- Digital Networked Library off Theses and Essays
- Research Papers in Economics
- E-Revist@s

with Santa Fe (New Mexico, October 1991)

Projects based in Open Access

- Central BioMed (BMC, 2001)
- Public The Library off Science (PLOS, 2000)The Scholarly Publishing and Academic Resources (SPARC, 1998)
- Scientific Electronic Library Online (SciELO, 2002)
- Directory off Open Access Journals (DOAJ, 2003)
- E-revist@s (CINDOC, 2004)

SEMIDE EMWIS

Logical architecture General

Preprint Service and OAI

- Provider of Meta data (data provider)
- Provider of Services (service provider)

SEMIDE EMWIS

Levels of the system (architecture 3 tiers)

- Presentation Level: navigator Web
- Logical/application level: to put into practice protocol OAI-PMH for the communication with files XML. Applications: Java, C, PHP, etc.
- Data Level: Meta data of Dublin Core. One recommends open source databases, such as MySQL, etc.

Application for SEMIDE- EMWIS

National Focal points

data Provider

International Focal Point

Services provider

Objectives for EMWIS

- Diffusion for all of the information collected by all the actors of the water sector: libraries, NFP, projects, etc
- Establish a system of quality standardized and validated for the bibliographical references.
- Architecture open to the deposit of bibliographical references offered by any organization working in the water sector in the Mediterranean
- To create an interface using protocol OAI-PMH: multiple data bases (NFP) and international catalogue (TU).
- Services: alarms, personalization, statistics, housing of reviews.
- Contribution on the initiative international "Open Access"

SEMIDE
EMWIS

OPEN ACCESS

Rationale:

- Internet has modified the way of information diffusion
- Scientists, technicians and professional do not have a free access to their own contents.
- Violations of the copyright (software, deposits, etc...).
- Barriers imposed by the leading monopolies.

Definition of Open Access

"Free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself"

(*Budapest Open Access Initiative 2002*)

Objectives of Open Access:

- To increase the speed of distribution of work by electronic means
- To reduce the costs of the scientific publications, and to increase the visibility of works published.
- To reduce the phenomenon of "lost science", i.e. which does not manage to have an important impact due to its absence in the major scientific data bases and directories.
- Access to information for all the citizens

Meta data collection for bibliographical references

Various implementations:

- Free access to the full information.
- Access restricted for some sectors (universities, developing country, etc).
- Free access to the articles having been published more than 6 months ago
- Free access to the reviews in digital format, but paper versions being sold.
- Etc

Meta data collection for bibliographical references

Self archiving

- The authors deposit their documents in a warehouse of open electronic archives.
- On the Web: Metadata and full text.
- The author is the person who requires the editor to include the complete text and requests the right to correct the data necessary if a direct access is not provided.
- Solution with prospects for growth.

Meta data collection for bibliographical references

Debate:

- Viability: “Open” different from “free”.
- Same problems as in the world of open source software: Licenses LPG, copyleft, creative commons.
- Advantages: broad diffusion, low costs, digital format.
- Barriers: economic interests, losses of quality? And of prestige?

STRUCTURE AND PROTOCOL OAI-PMH

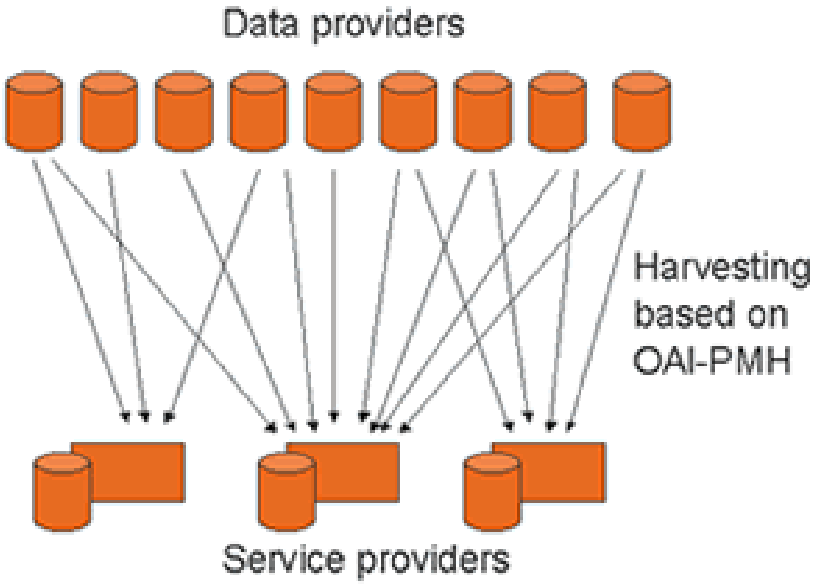
Meta data collection for bibliographical references

Introduction

- OAI-PHM: Open Archives Initiative Protocol for Metadata Harvesting.
- System requirements:
  - Integration among the open Archives systems
  - Support for a distributed architecture
  - Implementation of protocol OAI-PMH, for the Metadata exchange
  - Implementation of a “harvester” for collecting metadata
  - Creation of “Data Providers”

Logical architecture

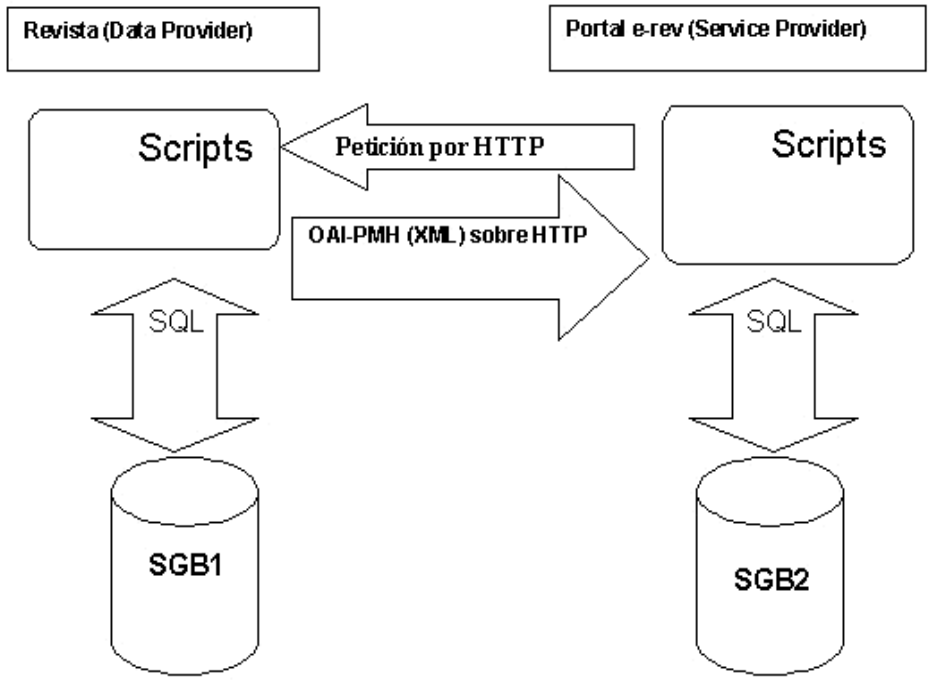
- The system logical architecture is based on UPS (Universal Prepint Service) and its derivation suggested by OAI (Open Archives Initiative)



Meta data collection for bibliographical references

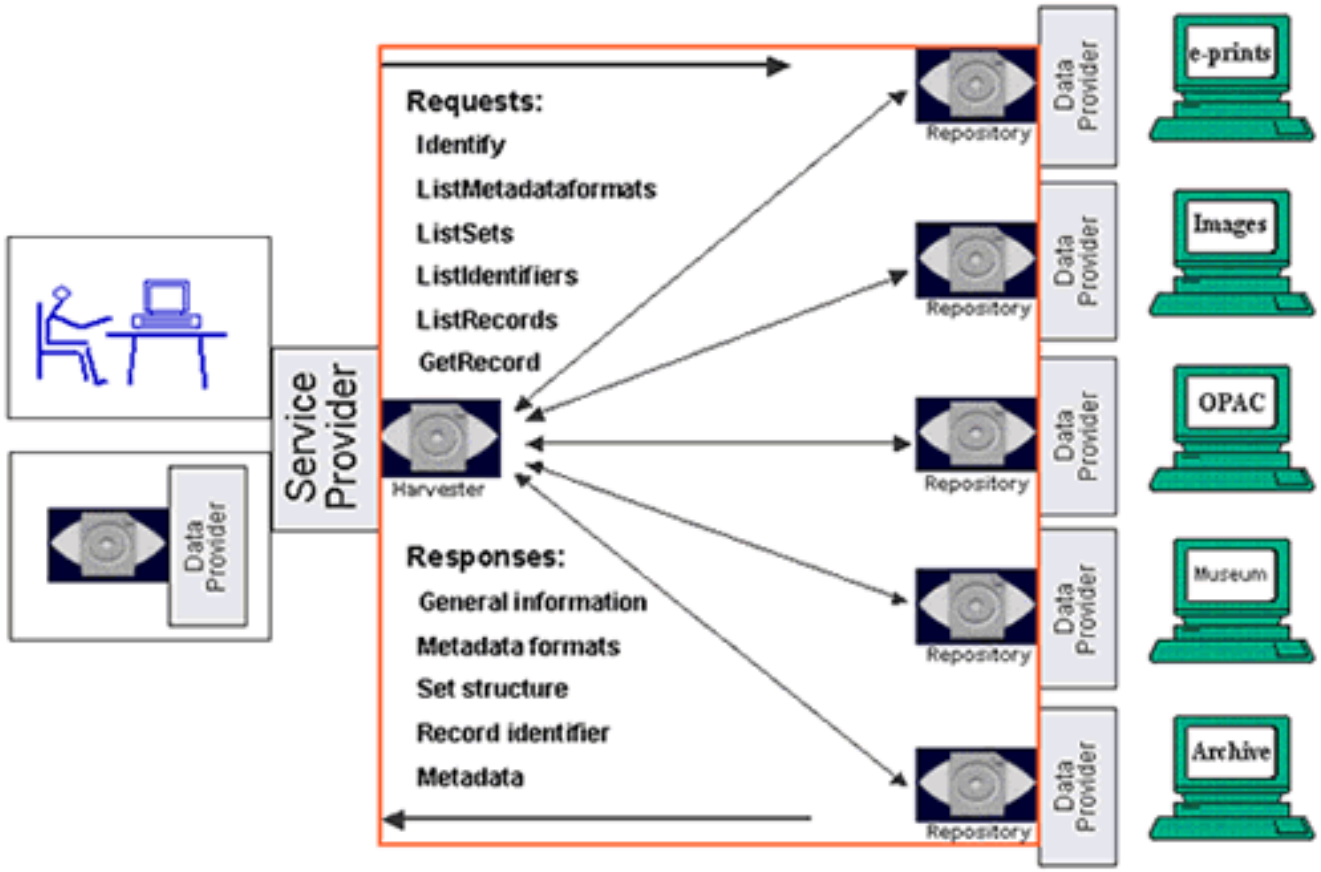
Multi-tiers system

- Presentation level
- Logical level (application, business logic)
- Data level

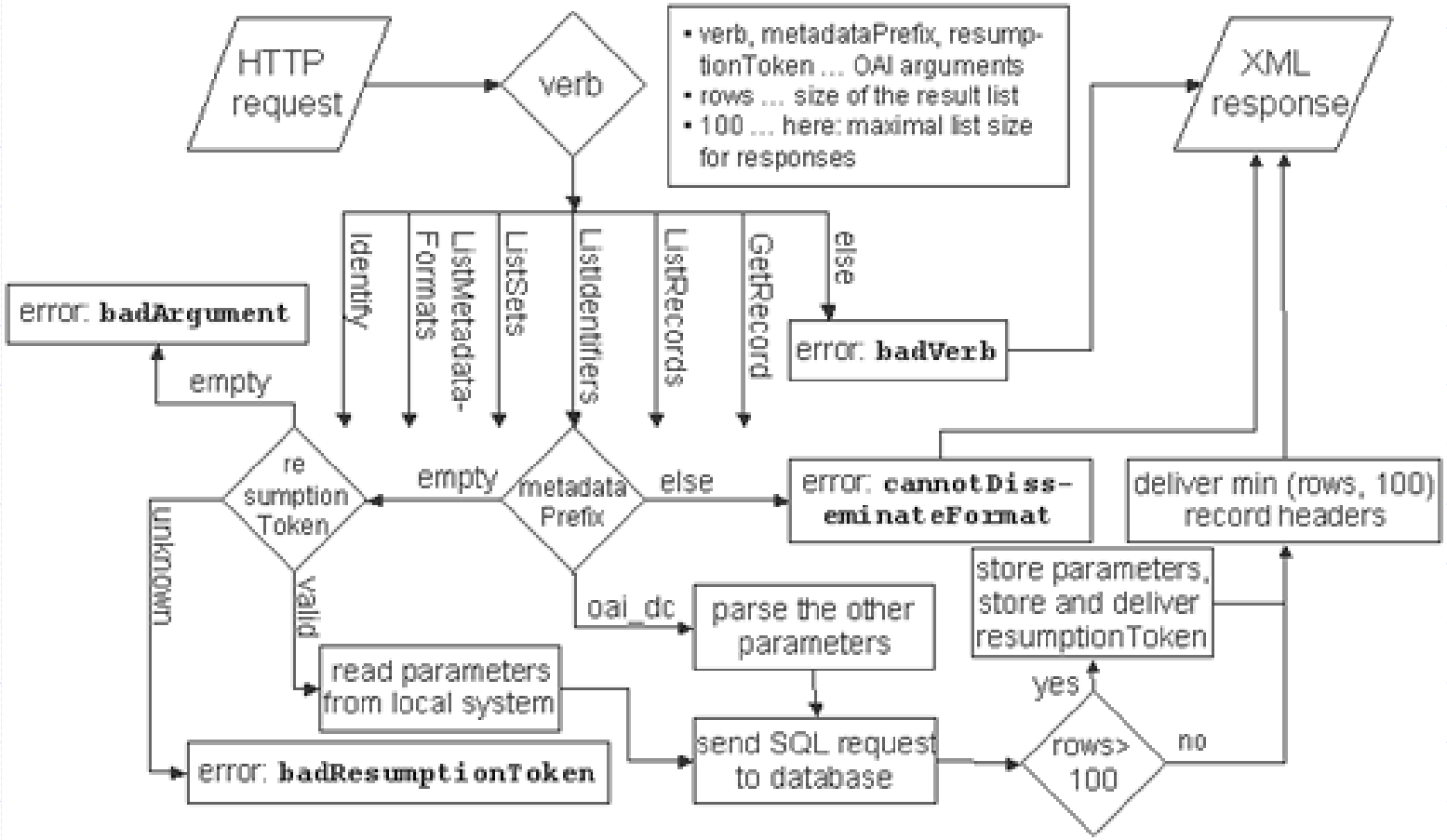


Meta data collection for bibliographical references

System Structure and Data flow (1)



System Structure and Data flow(2)



Requirements for the data Provider

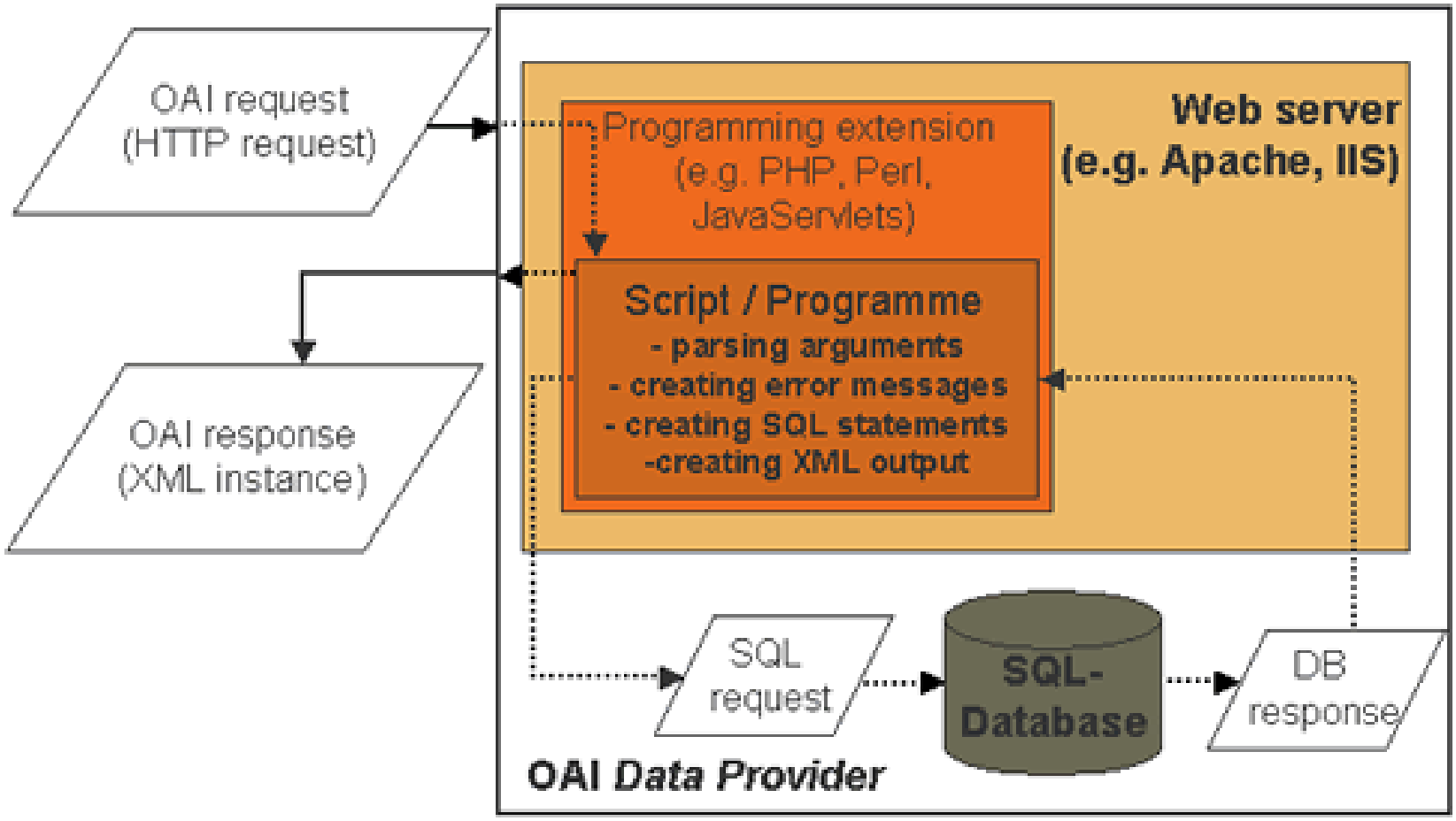
- Storage medium for Metadata
- Support of web standards
- Maintaining an API to implement the protocol
- URL
- URI
- Formats of Metadata
- Date-stamps
- Logical support of a hierarchy of sets
- Control flow (resumption token)

Meta data collection for bibliographical references

Components and architecture of a data provider

- Requests analyzer
- Generator of error messages
- Consultation application
- Generator of XML responses
- Flow control

Meta data collection for bibliographical references



Meta data collection for bibliographical references

Requirements for a Service Provider

- Storage medium for the metadata
- Support of web standards
- Maintaining an API to implement the protocol

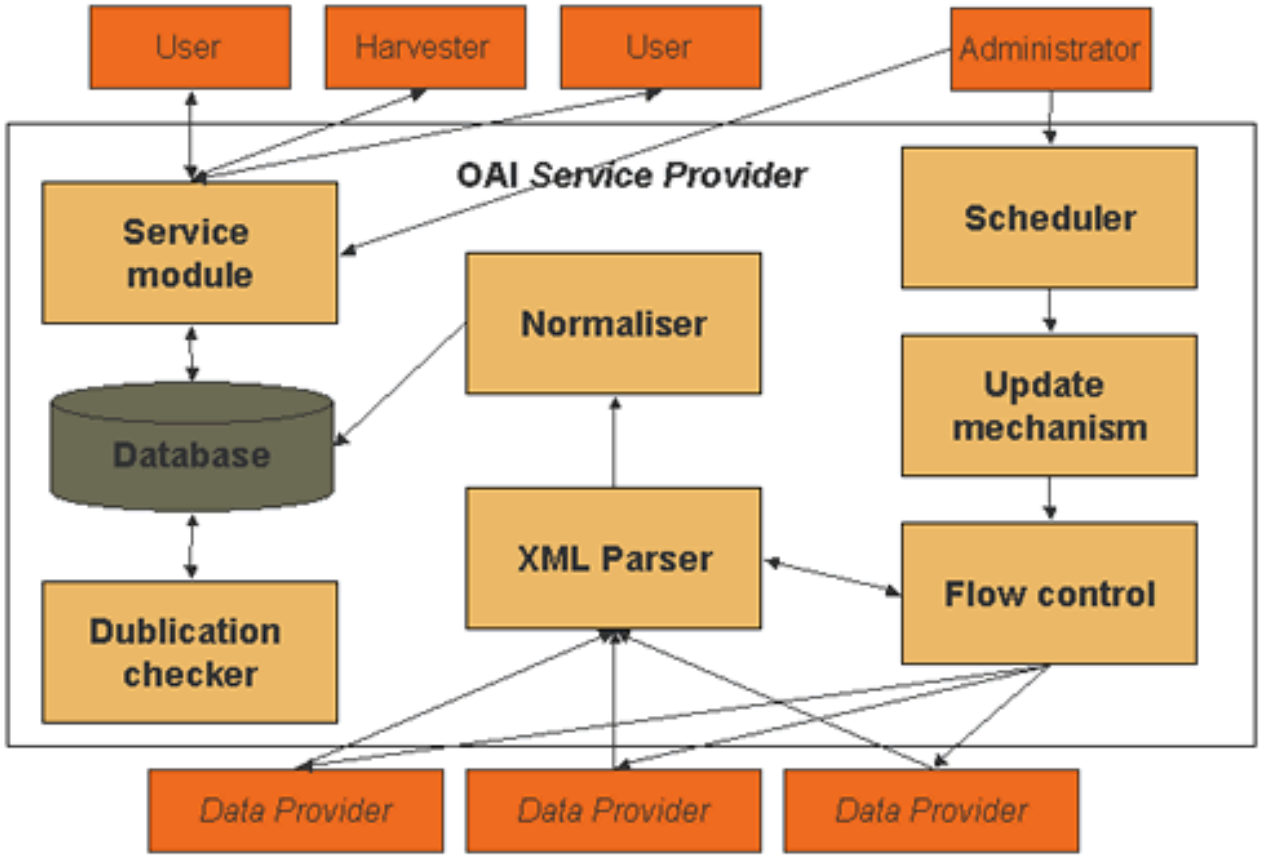
Meta data collection for bibliographical references

Components and architecture for the Service Provider

- Files management system
- Creator of requests (to access the metadata of “data providers”)
- Timer (Scheduler)
- Flow control
- Updating mechanism
- XML parser
- Normaliser (format)
- Data bases
- Duplication checker
- Service module

Meta data collection for bibliographical references

Components and architecture of the Service Provider



Meta data collection for bibliographical references

Description of protocol OAI-PMH

- Based on HTTP
- Responses are in XML
- Requests by GET and POST
- Requests key words = value
- There are 6 standard requests:
  - Identify
  - GetRecord
  - ListIdentifiers
  - ListRecords
  - ListSets
  - ListMetadataFormats

DUBLIN CORE

Meta data collection for bibliographical references

Introduction

- Main exchange format of metadata in the framework of the OIA-PMH protocol.
- The Dublin Core Metadata Initiative (DCMI) is dedicated to the promotion and the diffusion of the metadata interoperability standards and the development of metadata vocabularies specialized for the description of resources allowing systems to access information in an intelligent way.

Meta data collection for bibliographical references

Definition:

XML schema (XSD) with the following fields:

- Title
- Creator
- Subject
- Description
- Publisher
- Contributor
- Date
- Type of resource
- Format
- Identifier of the Resource
- Source
- Language
- Relation
- Coverage
- Rights

Meta data collection for bibliographical references

Definition:

- The elements have descriptive names which claim to transmit semantic by themselves
- Each element is optional and can be repeated. The elements can appear in any order
- Classification of the elements:
  - Elements on the contents of the resource: Title, Subject, Description, Source, Language, Relation, Coverage.
  - Elements on the intellectual property of the resource: Creator, Publisher, Contributor, Rights
  - Elements related to the definition of the resource: Date, Type, Format, Identifier

Meta data collection for bibliographical references

XSD: http://dublincore.org/schemas/xmls/simpledc20021212.xsd

```
<element name="dc" type="oai_dc:oai_dcType"/>

<complexType name="oai_dcType">
    <choice minOccurs="0" maxOccurs="unbounded">
        <element ref="dc:title"/>
        <element ref="dc:creator"/>
        <element ref="dc:subject"/>
        <element ref="dc:description"/>
        <element ref="dc:publisher"/>
        <element ref="dc:contributor"/>
        <element ref="dc:date"/>
        <element ref="dc:type"/>
        <element ref="dc:format"/>
        <element ref="dc:identifier"/>
        <element ref="dc:source"/>
        <element ref="dc:language"/>
        <element ref="dc:relation"/>
        <element ref="dc:coverage"/>
        <element ref="dc:rights"/>
    </choice>
</complexType>
```

Meta data collection for bibliographical references

XSD:

Each elements is of type “String” with the additional attribute of the language:

```
<xs:complexType name= " elementType " >
  <xs:simpleContent>
    <xs:extension base= " xs:string " >
   <xs:attribute ref= " xml:Lang “use= " optional "/>
    </xs:extension>
  </xs:simpleContent>
</xs:complexType>
```