Tutorial

Open Archive Initiative

Uwe Müller

Computer- und Medienservice, Humboldt-Universität zu Berlin

u.mueller@cms.hu-berlin.de

Dr. Heinrich Stamerjohanns

Institute for Science Networking, Universität Oldenburg

stamer@uni-oldenburg.de

Thanks

- Some of the slides presented here are our own!
- Many of them have been kindly donated by (taken from!):
  - Andy Powell
  - Herbert Van de Sompel
  - Carl Lagoze
  - Hussein Suleman
  - Michael Nelson
  - Simeon Warner
  - (and other probably...)

Agenda

- Part I - History and Overview
- Part II - OAI Serviceprovider - Example
- Part III - Technical Introduction
- Part IV - Implementation of Data Provider and Service Provider
- Part V - OAI Communities

Tutorial
Open Archive Initiative

Part I
History and Overview

OAI roots…

- the roots of OAI lie in the development of eprint archives…
  - arXiv, CogPrints, NACA (NASA), RePEc, NDLTD, NCSTRL
- each offered Web interface for deposit of articles and for end-user searches
- difficult for end-users to work across archives without having to learn multiple different interfaces
- recognised need for single search interface to all archives
  - Universal Pre-print Service (UPS)

Searching vs. harvesting

- two possible approaches to building the UPS…
- cross-searching multiple archives based on protocol like Z39.50
- harvesting metadata into one or more ‘central’ services – bulk move data to the user-interface
- US digital library experience in this area (e.g. NCSTRL) indicated that cross-searching not preferred approach - distributed searching of N nodes viable, but only for small values of N
- NCSTRL: N > 100; bad

Problems of cross-searching

- collection description
  - how do you know which targets to search?
- query-language problem
  - syntax varies and drifts over time between the various nodes
- rank-merging problem
  - how do you meaningfully merge multiple result sets?
- performance
  - tends to be limited by slowest target
  - difficult to build browse interface

Universal Preprint Service

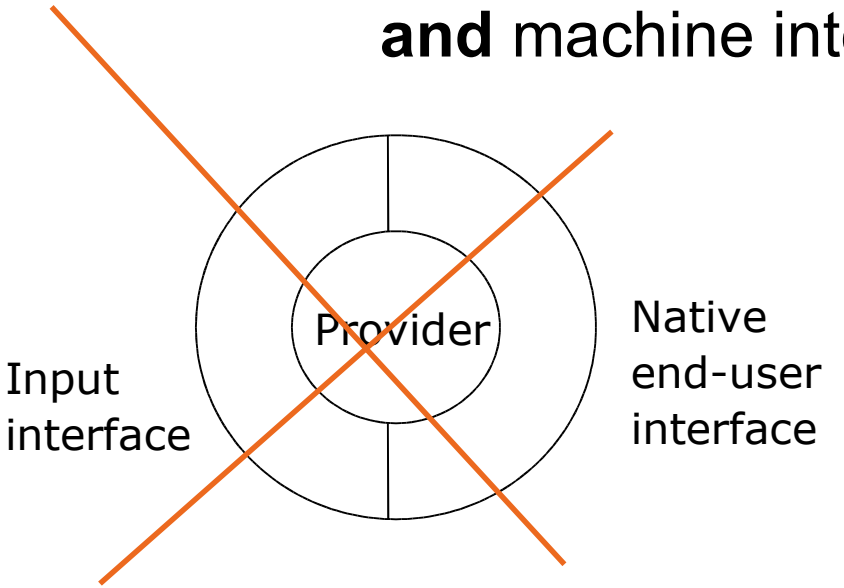
- a cross-archive DL that that provides services on a collection of metadata harvested from multiple archives
  - based on NCSTRL+; a modified version of Dienst
- demonstrated at Santa Fe NM, October 21-22, 1999
  - http://ups.cs.odu.edu/
  - D-Lib Magazine, 6(2) 2000 (2 articles)
    http://www.dlib.org/dlib/february00/02contents.html
- UPS was soon renamed the Open Archives Initiative (OAI) http://www.openarchives.org/

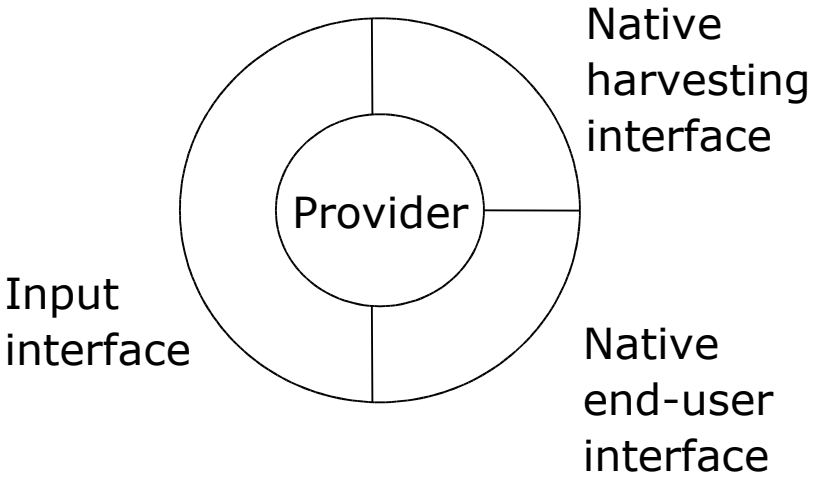
Data and Service Providers

- UPS identified two logical groups of services…
- data providers
  - handle deposit/publishing of resources in archive
  - expose metadata about resources in archive
- service providers
  - harvest metadata from data providers
  - use it to offer single user-interface across all harvested metadata
- note:
  - data provider may also be responsible for human-oriented (I.e. Web) interface to archive
  - both functions may be offered by same ‘service’

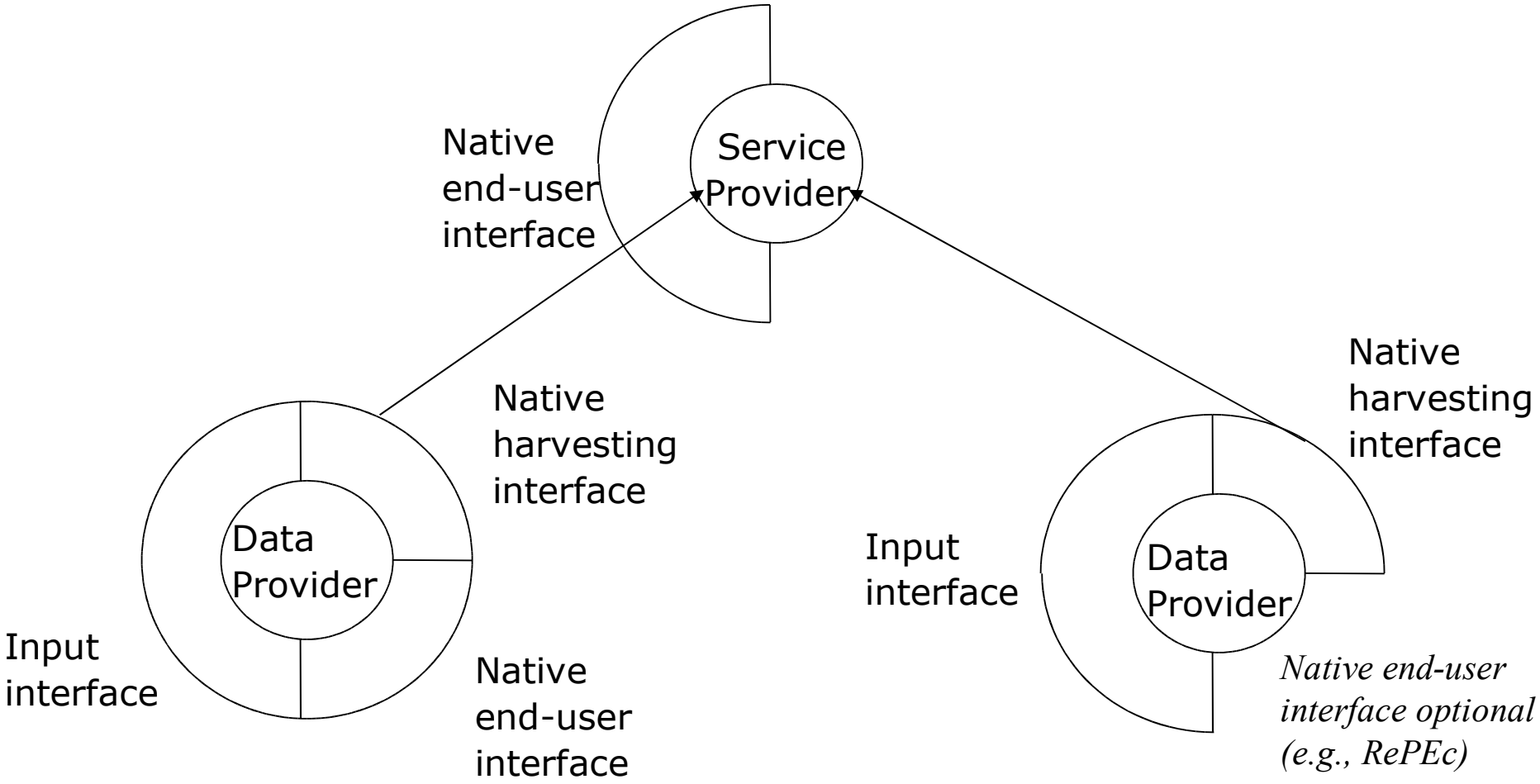
Human vs. machine interfaces

- move away from only supporting human end-user interfaces for each archive…
- …to supporting both human end-user interface **and** machine interfaces for harvesting





Service provider harvesting



Metadata harvesting requirements

- in order that the harvesting approach can work we need agreements about…
  - transport protocols – HTTP vs. FTP vs. …
  - metadata formats – DC vs. MARC vs. …
  - quality assurance – mandatory elements, mechanisms for naming of people, subjects, etc., handling duplicated records, best-practice
  - intellectual property and usage rights – who can do what with the records
- work in this area resulted in the “Santa Fe Convention”

Santa Fe Convention [02/2000]

- goal: optimize discovery of e-prints
- inputs…
  - UPS prototype
  - RePEc/SODA “data provider / service provider” model
  - Dienst protocol
  - deliberations at Santa Fe meeting [10/1999]

OAI-PMH v 1.0 [01/2001]

- goal: optimise discovery of document-like objects

- inputs…
  - Santa Fe Convention
  - various DLF meetings on metadata harvesting
  - deliberations at Cornell
  - alpha-testers of OAI-PMH v 1.0
  - recognition of DC as ‘best’ core metadata format for interoperability across multiple archives

OAI-PMH v 1.0 [01/2001]

- low-barrier interoperability specification
- metadata harvesting model: data provider / service provider
- focus on document-like objects
- autonomous protocol
- HTTP based
- XML responses
- unqualified Dublin Core
- experimental: 12-18 months

OAI timeline before v. 2.0

- October 21-22, 1999 - initial UPS meeting
- February 15, 2000 - Santa Fe Convention published in D-Lib Magazine
  - precursor to the OAI metadata harvesting protocol
- June 3, 2000 - workshop at ACM DL 2000 (Texas)
- August 25, 2000 - OAI steering committee formed, DLF/CNI support
- September 7-8, 2000 - technical meeting at Cornell University
  - defined the core of the current OAI metadata harvesting protocol
- September 21, 2000 - workshop at ECDL 2000 (Portugal)
- November 1, 2000 - Alpha test group announced (~15 organizations)
- Dezember 2000 Dini Jahrestagung in Dortmand

OAI timeline before v. 2.0

- January 23, 2001 - OAI protocol 1.0 announced, OAI Open Day in the U.S. (Washington DC)
  - purpose: freeze protocol for 12-16 months, generate critical mass
- February 26, 2001 - OAI Open Day in Europe (Berlin)
- July 3, 2001 - OAI protocol 1.1 announced
  - to reflect changes in the W3C's XML latest schema recommendation
- September 8, 2001 - workshop at ECDL 2001 (Darmstadt)

OAI-PMH v.2.0 [06/2002]

- goal: recurrent exchange of metadata about resources between systems
- inputs:
  - OAI-PMH v.1.0
  - feedback on OAI-implementers
  - deliberations by OAI-tech [09/01 - 06/02]
  - alpha test group of OAI-PMH v.2.0 [03/02 - 06/02]
  - officially released June 14, 2002

OAI-PMH v.2.0 [06/2002]

- low-barrier interoperability specification
- metadata harvesting model: data provider / service provider
- metadata about resources
- autonomous protocol
- HTTP based
- XML responses
- unqualified Dublin Core
- stable

	Santa Fe convention	OAI-PMH v.1.0/1.1	OAI-PMH v.2.0
nature	experimental	experimental	stable
verbs	Dienst	OAI-PMH	OAI-PMH
requests	HTTP GET/POST	HTTP GET/POST	HTTP GET/POST
responses	XML	XML	XML
transport	HTTP	HTTP	HTTP
metadata	OAMS	unqualified Dublin Core	unqualified Dublin Core
about	eprints	document like objects	resources
model	metadata harvesting	metadata harvesting	metadata harvesting

What's in a name?

Open Archives Initiative

the protocol is openly documented, and metadata is "exposed" to at least some peer group (note: rights management can still apply!)

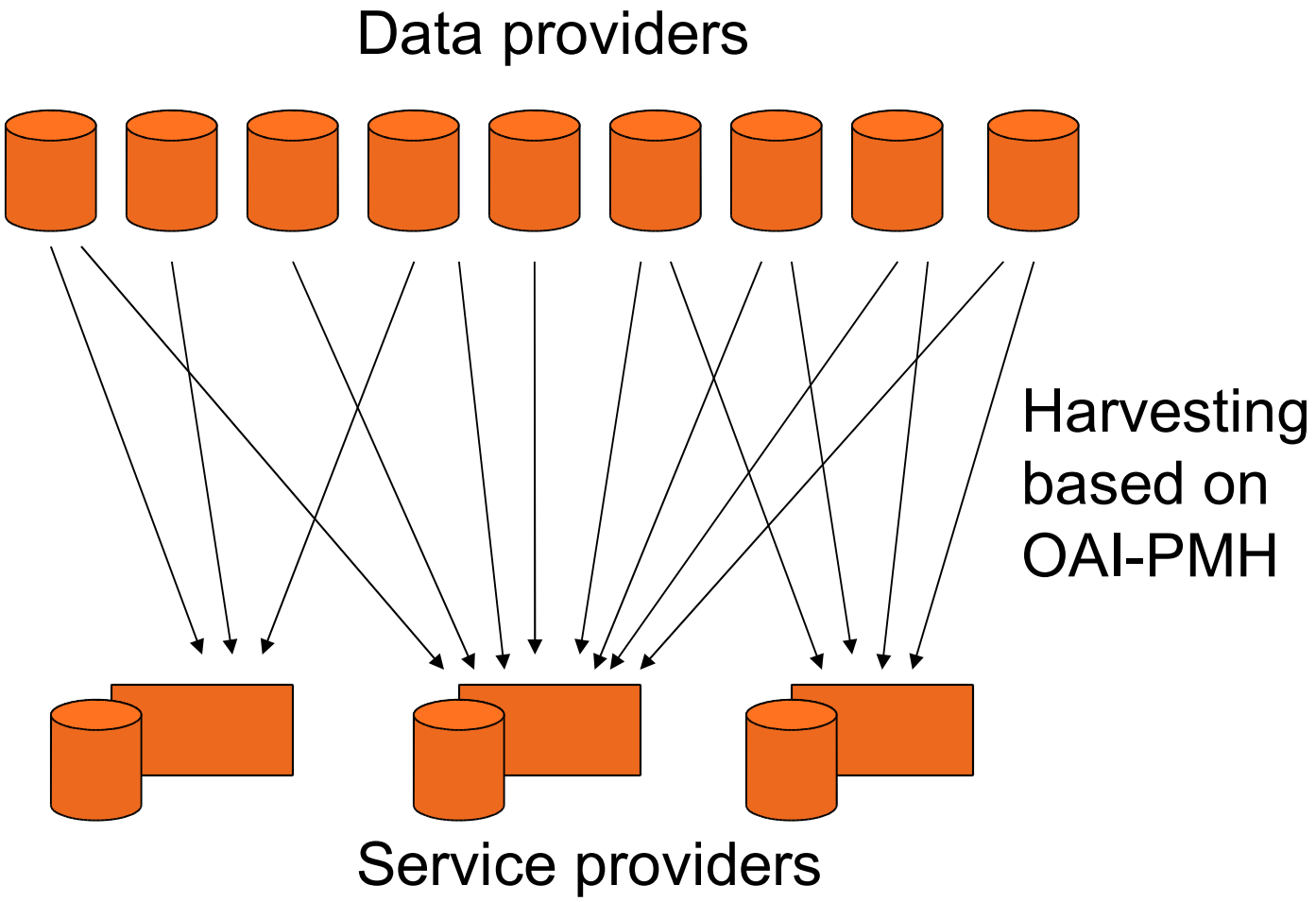
archive defined as a "collection of stuff" -- not the archivist's definition of "archive". "Repository" used in most OAI documents.

OAI is happening at break-neck speed...

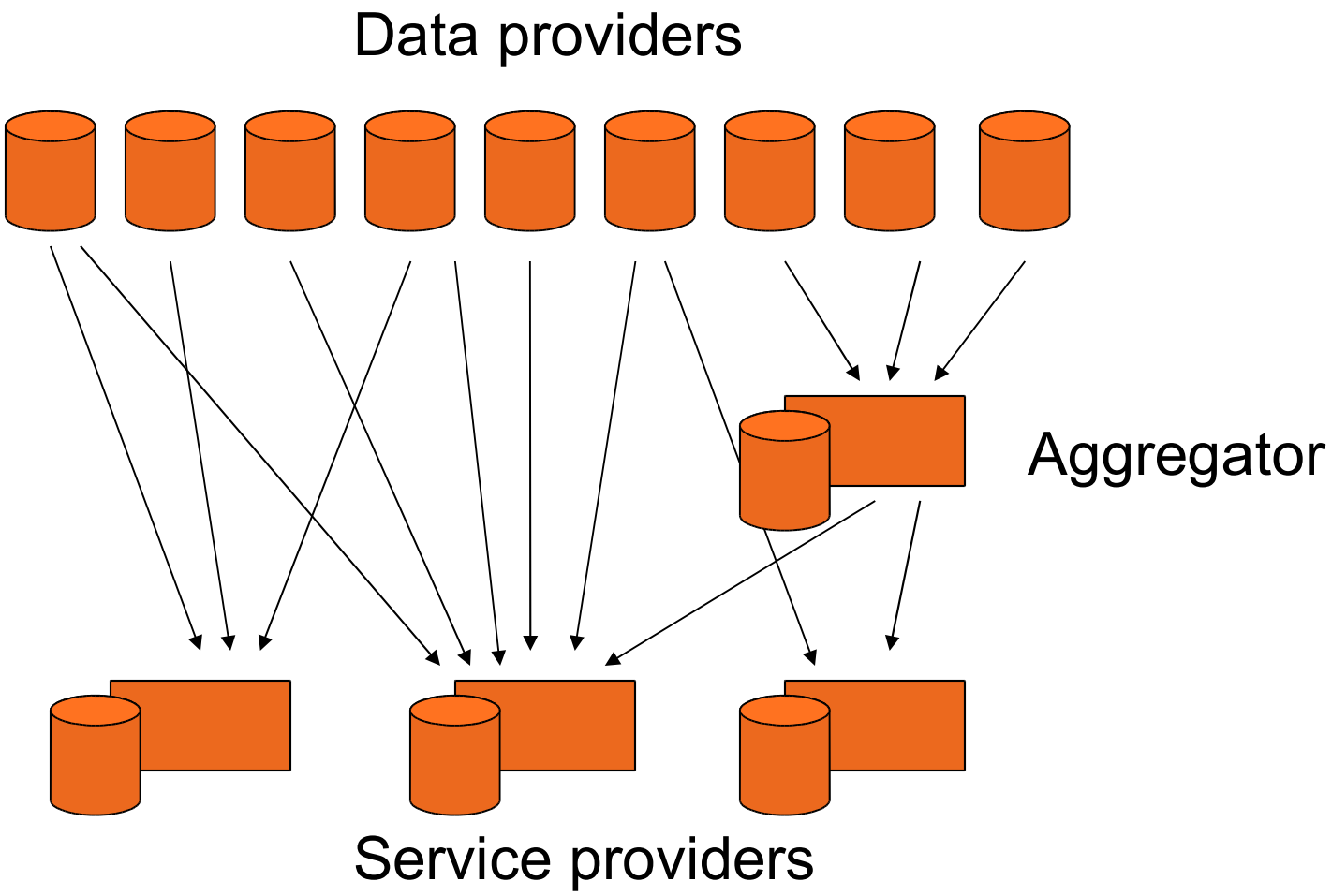
Flexible deployment

- simple protocol based on HTTP and XML allows for rapid deployment
- a number of toolkits available – see part III
- systems can be deployed in variety of configurations
- multiple service providers can harvest from multiple data providers
- aggregators can sit between data and service providers
- harvesting approach can be complemented with searching based on Z39.50 or SRW

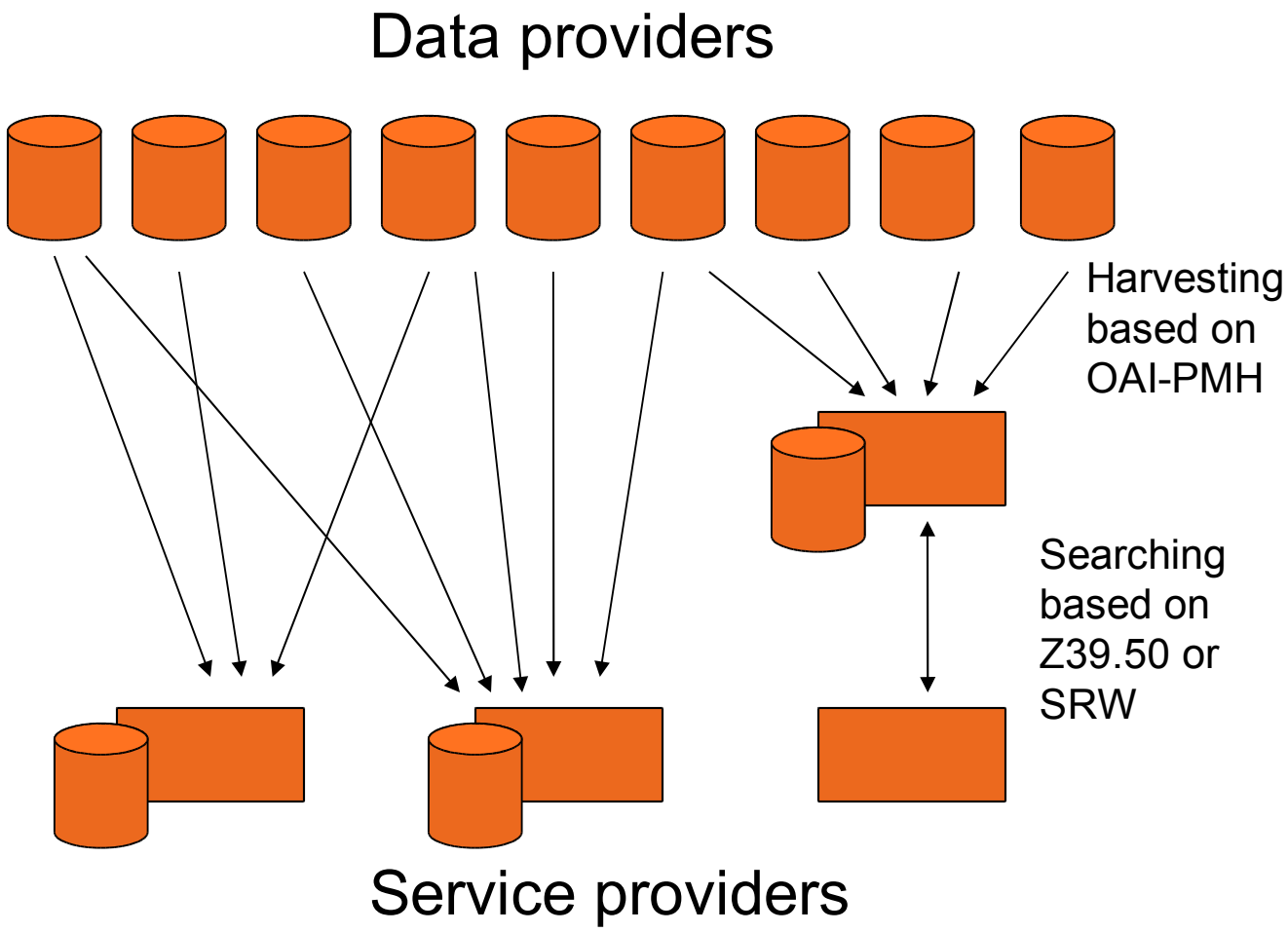
Multiple data and service p's



Aggregators



Can be mixed with x-searching



Summary

- OAI-PMH – OAI Protocol for Metadata Harvesting
- low-cost mechanism for harvesting metadata records from one system to another
  - from ‘data providers’ to ‘service providers’
- development over last 2-3 years has seen move from specific (discovery of e-prints) to generic (sharing descriptions of any resources)
- based on HTTP and XML – Web-friendly
- allows client to say ‘give me some or all of your records’ where ‘some’ is based on
- datestamps, sets, metadata formats

Summary (2)

- mandates simple DC as record format but extensible to any format encoded in XML
- OAI-PMH is **not** a search protocol
  - but use can underpin search-based services based on Z39.50 or SRW or …
- metadata and full-text typically made freely available – but not a requirement
  - OAI-PMH can be used between closed groups
- access-control and compression mechanisms based on underlying HTTP protocol
- simple protocol allows easy deployment
- systems can be combined in variety of ways

Important resources

- OAI Web site:
  - http://www.openarchives.org/
- OAI-PMH specification:
  - http://www.openarchives.org/OAI/openarchivesprotocol.html
- Implementation guidelines:
  - http://www.openarchives.org/OAI/2.0/guidelines.htm
- Discussion lists:
  - http://www.openarchives.org/mailman/listinfo/oai-general
  - http://oaisrv.nsdl.cornell.edu/mailman/listinfo/oai-implementers
- Repository explorer:
  - http://oai.dlib.vt.edu/cgi-bin/Explorer/oai2.0/testoai
- Tools: http://oai.dlib.vt.edu/cgi-bin/Explorer/oai2.0/testoai

Agenda

- Part I - History and Overview
- Part II - OAI Serviceprovider - Example
- Part III - Technical Introduction
- Part IV - Implementation of Data Provider and Service Provider
- Part V - OAI Communities

Service Provider Examples

Citation Indexing

http://icite.sissa.it

Search Engine

http://arc.cs.odu.edu

Printing on Demand Service

http://www.proprint-service.de

Value added Search Engine

http://www.myoai.com

Agenda

- Part I - History and Overview
- Part II - OAI Serviceprovider - Example
- Part III - Technical Introduction
- Part IV - Implementation of Data Provider and Service Provider
- Part V - OAI Communities

Tutorial
Open Archive Initiative

Part III

Technical Introduction

What is an „Open Archive“

- Any WWW-based system that can be accessed through the well-defined interface of the Open Archives Protocol for Metadata Harvesting.
- Is then known as an OAI-compliant archive
- No implications for:
  - Physical storage of data
  - Cost of data
  - Metadata and data formats
  - Access control to server

Reminder: Harvesting vs. Federation

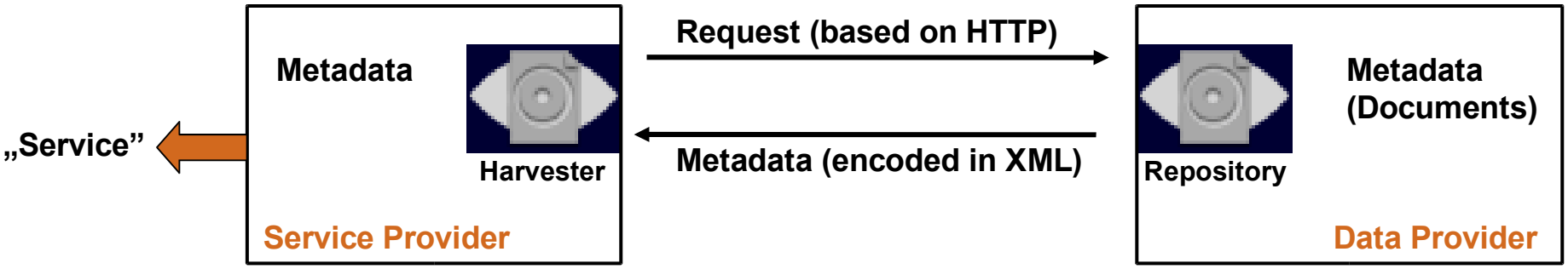
- Competing approaches to interoperability
  - Federation is when services are run remotely on remote data (e.g. Federated searching)
  - Harvesting is when data/metadata is transferred from the remote source to the destination where the services are located (e.g. Union catalogues)
- Federation requires more effort at each remote source but is easier for the local system and vice versa for harvesting
- OAI currently focuses on harvesting

Metadata vs. Data

- Data refers to digital objects or digital representations of objects
- Metadata is information about the objects (e.g. title, author, etc.)
- OAI focuses on metadata, with the implicit understanding that metadata usually contains useful links to the source digital objects

The Open Archives Initiative (OAI)

- Main ideas
  - world-wide consolidation of scholarly archives
  - free access on the archives (at least: metadata)
  - consistent interfaces for archives and service provider
  - low barrier protocol / effortless implementation
  - based on existing standards (e.g. HTTP, XML, DC)
- Basic functioning



Requirements of the protocol

Should

- be in machine readable format
- encoded in a strict format, which can be validated
  - character encoding
  - metadata encoding
- support different content models
  - metadata formats
- use existing technologies (HTTP, XML, DC)
  - easy to implement
  - easy to adjust

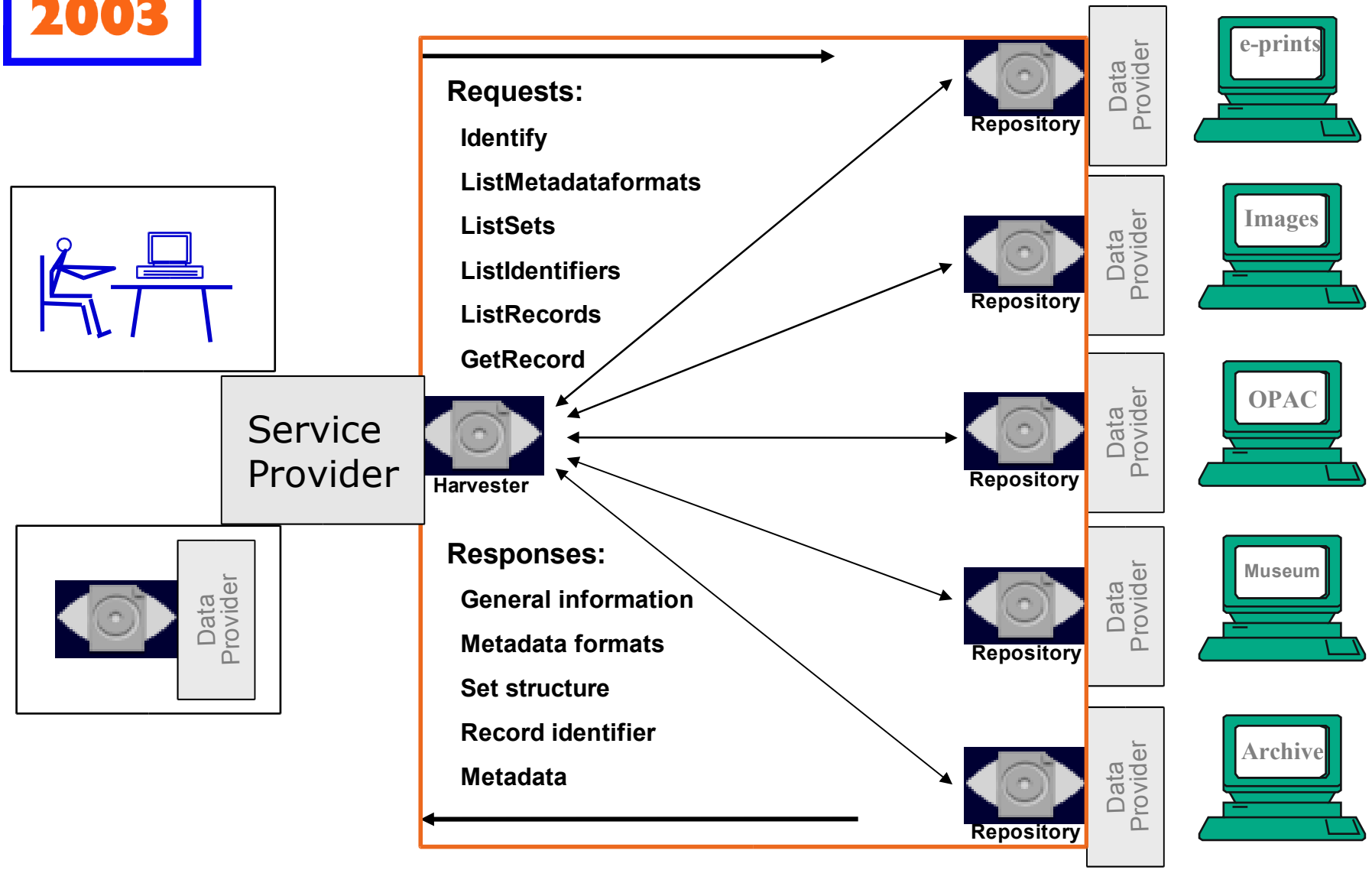
Data and Service Provider

- Data Providers refer to entities who possess data/metadata and are willing to share this with others (internally or externally) via well-defined OAI protocols (e.g. database servers)
- Service Providers are entities who harvest data from Data Providers in order to provide higher-level services to users (e.g. search engines)
- OAI uses these denotations for its client/server model (data=server, service=client)

OAI: General Assumptions

- two groups of ‘participants’
- Data Providers (Open Archives, Repositories)
  - free access of metadata
  - not necessarily: free access to full texts / resources
  - easy to implement, low barriers
- Service Providers
  - use OAI interfaces of the Data Providers
  - harvest and store metadata (no live requests!)
  - may select certain subsets from Data Providers (set hierarchy, date stamp)
  - may enrich metadata
  - offer (value-added) service on the basis of the metadata

OAI-PMH: Structure Model



OAI-PMH: Protocol Overview

- Protocol based on HTTP
  - request arguments as GET or POST parameters
  - six request types
  - e.g. http://archive.org?
    verb=ListRecords&from=2002-11-01
  - responses are encoded in XML syntax
  - supports any metadata format (at least: Dublin Core)
  - logical set hierarchy (definition: data providers)
  - datestamps (last change of metadata set)
  - error messages
  - flow control

Protocol Details: Definitions

- **Harvester**
  - client application issuing OAI-PMH requests
- **Repository**
  - network accessible server, able to process OAI-PMH requests correctly
- **Resource**
  - object the metadata is “about”, nature of resources is not defined in the OAI-PMH
- **Item**
  - component of an repository from which metadata about a resource can be disseminated
  - has an unique identifier
- **Record**
  - metadata in a specific metadata format
- **Identifier**
  - unique key for an item in a repository
- **Set**
  - optional construct for grouping items in a repository

Protocol Details: Definitions (2)

resource

item = identifier

Metadata about *David*

item

Dublin Core metadata

MARC metadata

SPECTRUM metadata

record

What is a „Record“ ?

- A record refers to an independent XML structure that may be associated with digital or physical objects
- Records are usually associated with metadata, not data
- Are the representation of an item in a specific metadata format
- OAI advocates harvesting of records, which contain metadata and additional fields to support the harvesting operation

Uniqueness and Persistence

- Each record must be uniquely addressable by a distinct identifier
  - (**`identifier + metadataPrefix`**)
- Each metadata entity should ideally be persistent to guarantee that service providers can always refer back to the source

Protocol Details: Records

- metadata of a resource in a specific format
- three parts
  - **header (mandatory)**
    - identifier (1)
    - datestamp (1)
    - setSpec elements (*)
    - status attribute for deleted item (?)
  - **metadata (mandatory)**
    - XML encoded metadata with root tag, namespace
    - repositories must support Dublin Core
  - **about (optional)**
    - rights statements
    - provenance statements

Example: OAI Record

(NOTE: Schema and Namespaces have been removed for simplicity)

```
<record>
  <header>
    <identifier>oai:physnet.de:tut1</identifier>
    <datestamp>2003-05-24</datestamp>
    <setSpec>tut</setSpec>
  </header>
  <metadata>
    <oai_dc>
      <title>OAI Tutorial at ETD 2003</tritle>
      <creator>Heinrich Stamerjohanns</creator>
      <creator>Uwe Müller</creator>
      <language>eng</language>
    </oai_dc>
  </metadata>
  <about>
    <rights>You are free to reuse this</rights>
  </about>
</record>
```

Datestamps & Harvesting

- date of last modification of the **metadata.**
- mandatory characteristic of every item
- two possible granularities: YYYY-MM-DD, YYYY-MM-DDThh:mm:ssZ
- function: information on metadata, selective harvesting (from and until arguments)
- applications: incremental update mechanisms
- modification, creating, deletion
- deletion: three support levels
  - no, persistent, transient

Protocol Details: Metadata Schemes

- OAI-PMH supports dissemination of multiple metadata formats from a repository
- properties of metadata formats
  - id string to specify the format (metadataPrefix)
  - metadata schema URL (XML schema to test validity)
  - XML namespace URI (global identifier for metadata format)
- repositories must be able to disseminate at least unqualified Dublin Core
- **arbitrary metadata formats** can be defined and transported via the OAI-PMH
- returned metadata must comply with XML schema and namespace specification

Sets

- Protocol mechanism to allow for harvesting of sub-collections
- No well-defined semantics – depends completely on local data providers
- May be defined by arrangement between data providers and service providers
- applications:
  subject gateways, dissertation search engine, …
- examples (Germany, see http://www.dini.de)
  - publication types (thesis, article, …)
  - document types (text, audio, image, …)
  - content sets, regarding DNB (Medicine, biology, …)

Protocol Details: Request format

- requests must be submitted using the **`GET`** or **`POST`** methods of HTTP
- repositories must support both methods
- at least one key=value pair: verb=[RequestType]
- additional key=value pairs depend on request type
- example for **`GET`** request: http://archive.org/oai?verb=ListRecords&metadataPrefix=oai_dc
- encoding of special characters
  e.g. “**:**” (host port separator) becomes “**%3A**”

Protocol Details: Response

- formatted as HTTP responses
- content type must be text/xml
- status codes (distinguished from OAI-PMH errors)
  e.g. 302 (redirect), 503 (service not available)
- response format: well formed XML with markup:
  1. XML declaration
     (`<?xml version="1.0" encoding="UTF-8" ?>`)
  2. root element named `OAI-PMH` with three attributes
     (`xmlns`, `xmlns:xsi`, `xsi:schemaLocation`)
  3. three child elements
     1. `responseDate` (UTC datetime)
     2. `request` (request that generated this response)
     3. a) `error` (in case of an error or exception condition)
        b) element with the name of the OAI-PMH request

Example Response

```
<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
         xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
         xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
         http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
 <responseDate>2003-03-28T14:59:21Z</responseDate>
 <request verb="ListRecords" metadataPrefix="oai_dc">
          http://physnet.uni-oldenburg.de/oai/oai2.php</request>
 <ListRecords>
  <record>
   <header>
    <identifier>oai:physdoc:http://www.ensta.fr</identifier>
    <datestamp>2002-01-25T00:00:00Z</datestamp>
   </header>
   <metadata>
     <oai_dc:dc
       xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
       xmlns:dc="http://purl.org/dc/elements/1.1/"
       xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
       xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
       http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
      <dc:title>Pole de Calcul Parallele,</dc:title>
      <dc:date>2000-01-05</dc:date>
      <dc:identifier>http://www.ensta.fr</dc:identifier>
      <dc:language>eng</dc:language>
     </oai_dc:dc>
   </metadata>
  </record>
  <record>
   <header>
    <identifier>oai:physdoc:ftp://ftp.esi.ac.at/pub/Preprints/esi955.ps</id
    <datestamp>2002-01-25T00:00:00Z</datestamp>
   </header>
   <metadata>
     <oai_dc:dc
```

Flow Control

- Flow control on two protocol levels
  - HTTP (503, retry-after)
  - OAI-PMH, Resumption-Token
- HTTP “retry-after” mechanism can be used in order delay requests of clients
- Resumption Tokens are used to return parts (incomplete lists) of the result.
- Client receive a token which can be used to issue another request, in order to receive further parts of the result.

Protocol Details: Flow Control

- four of the request types return a list of entries
- three of them may reply ‘large’ lists
- OAI-PMH supports partitioning
- decision on partitioning: repository
- response to a request includes
  - incomplete list
  - resumption token
    + expiration date, size of complete list, cursor (optional)
- new request with same request type
  - resumption token as parameter
  - all other parameters omitted!
- response includes
  - next (maybe last) section of the list
  - resumption token (empty if last section of list enclosed)

Protocol Details: Flow Control (2)

Example

Protocol Details: Errors and Exceptions

- repositories must indicate OAI-PMH errors
- inclusion of one or more **`error`** elements
- defined error identifiers
  - **`badArgument`**
  - **`badResumptionToken`**
  - **`badVerb`**
  - **`cannotDisseminateFormat`**
  - **`idDoesNotExist`**
  - **`noRecordsMatch`**
  - **`noMetaDataFormats`**
  - **`noSetHierarchy`**

Request Types

- six different request types
  1. Identify
  2. ListMetadataFormats
  3. ListSets
  4. ListIdentifiers
  5. ListRecords
  6. GetRecord
- harvester has not to use all types
- repository must implement all types
- required and optional arguments
- depend on request types

Identify

- Function
  - general information about archive
- Parameter
  - none
- Example URL
  - http://physnet.de/oai/oai2.php?verb=Identify
- Errors/Exceptions
  - **`badArgument`**
    z.B. physnet.de/oai/oai2.php?verb=Identify**&set=biology**

Request Types: Identify (2)

Responseformat

Element	*Example*	#
repositoryName	My Archive	1
baseURL	http://archive.org/oai	1
protocolVersion	2.0	1
earliestDatestamp	1999-01-01	1
deleteRecords	no, transient, persistent	1
granularity	YYYY-MM-DD, YYYY-MM-DDThh:mm:ssZ	1
adminEmail	oai-admin@archive.org	+
compression	deflate, compress, …	*
description	oai-identifier, eprints, friends, …	*

Identify – Response

```
- <OAI-PMH xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
  http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
    <responseDate>2003-03-30T14:46:08Z</responseDate>
    <request verb="Identify">http://physnet.uni-oldenburg.de/oai/oai2.php</request>
   - <Identify>
      <repositoryName>PhysNet, Oldenburg, Germany, Document Server</repositoryName>
      <baseURL>http://physnet.uni-oldenburg.de/oai/oai2.php</baseURL>
      <protocolVersion>2.0</protocolVersion>
      <adminEmail>mailto:stamer@uni-oldenburg.de</adminEmail>
      <earliestDatestamp>2000-01-01T00:00:00Z</earliestDatestamp>
      <deletedRecord>no</deletedRecord>
      <granularity>YYYY-MM-DDThh:mm:ssZ</granularity>
      <compression>gzip</compression>
     - <description>
       - <friends xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/friends/
         http://www.openarchives.org/OAI/2.0/friends.xsd">
          <baseURL>http://naca.larc.nasa.gov/oai2.0/</baseURL>
          <baseURL>http://techreports.larc.nasa.gov/ltrs/oai2.0/</baseURL>
          <baseURL>http://physnet.uni-oldenburg.de/oai/oai.php</baseURL>
          <baseURL>http://cogprints.soton.ac.uk/perl/oai</baseURL>
          <baseURL>http://ub.uni-duisburg.de:8080/cgi-oai/oai.pl</baseURL>
         - <baseURL>
            http://rocky.dlib.vt.edu/~jcdlpix/cgi-bin/OAI1.1/jcdlpix.pl
          </baseURL>
        </friends>
      </description>
    </Identify>
  </OAI-PMH>
```

ListMetadataFormats

- Function
  - list metadata formats, which are supported by archive, as well as their Schema Locations and Namespaces
- Parameter
  - identifier – for a specific record (O)
- Example URL
  - http://physnet.de/oai/oai2.php?verb=ListMetadataFormats
- Errors/Exceptions
  - **`badArgument`**
  - **`idDoesNotExist`**

    archive.org/oai-script?verb=ListMetadataFormats**&**
    identifier=really-wrong-identifier
  - **`noMetadataFormats`**

ListMetadataFormats Response

```
- <OAI-PMH xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
  http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
    <responseDate>2003-03-30T14:56:43Z</responseDate>
    <request verb="ListMetadataFormats">http://physnet.uni-oldenburg.de/oai/oai2.php</request>
   - <ListMetadataFormats>
     - <metadataFormat>
         <metadataPrefix>oai_dc</metadataPrefix>
         <schema>http://www.openarchives.org/OAI/2.0/oai_dc.xsd</schema>
         <metadataNamespace>http://www.openarchives.org/OAI/2.0/oai_dc/</metadataNamespace>
       </metadataFormat>
    </ListMetadataFormats>
  </OAI-PMH>
```

ListSets

- Function
  - hierarchical listing of Sets in which records have been organized
- Parameter
  - none
- Example URL
  - http://physnet.de/oai/oai2.php?verb=ListSets
- Errors/Exceptions
  - **badArgument**
  - **badResumptionToken**
    archive.org/oai-script?verb=ListSets**&**
    resumptionToken=any-wrong-token
  - **noSetHierarchy**

ListIdentifiers

- Function
  - retrieve headers of all Records, which comply to parameters
- Parameter
  - **`from`** – Startdate (O)
  - **`until`** – Enddate (O)
  - **`set`** – Set of which to be harvested (O)
  - **`metadataPrefix`** – metadata format, for which Identifier should be listed (R)
  - **`resumptionToken`** – flow control (X)
- Example URL
  - http://physnet.de/oai/oai2.php?verb=ListIdentifiers&metadataPrefix=oai_dc

ListIdentifiers

- Errors/Exceptions
  - **badArgument, z.B.. …&from=2002-12-01-13:45:00**
  - **badResumptionToken**
  - **cannotDisseminateFormat**
  - **noRecordsMatch**
  - **noSetHierarchy**

ListRecords

- Function
  - retrieve multiple Records
- Parameter
  - **`from`** – Startdate (O)
  - **`until`** – Enddate (O)
  - **`set`** – Set from which to be harvested (O)
  - **`metadataPrefix`** – metadata format (R)
  - **`resumptionToken`** – flow control (X)
- Example UR
  - http://physnet.de/oai/oai2.php?verb=ListRecords&metadataPrefix=oai_dc&from=2001-01-01

ListRecords

- Errors/Exceptions
  - **`badArgument`**
  - **`badResumptionToken`**
  - **`cannotDisseminateFormat`**
  - **`noRecordsMatch`**
  - **`noSetHierarchy`**

ListRecords – Response

```
<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
         xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
         xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
         http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
 <responseDate>2003-03-28T14:59:21Z</responseDate>
 <request verb="ListRecords" metadataPrefix="oai_dc">
          http://physnet.uni-oldenburg.de/oai/oai2.php</request>
 <ListRecords>
  <record>
   <header>
    <identifier>oai:physdoc:http://www.ensta.fr</identifier>
    <datestamp>2002-01-25T00:00:00Z</datestamp>
   </header>
   <metadata>
     <oai_dc:dc
       xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
       xmlns:dc="http://purl.org/dc/elements/1.1/"
       xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
       xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
       http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
      <dc:title>Pole de Calcul Parallele,</dc:title>
      <dc:date>2000-01-05</dc:date>
      <dc:identifier>http://www.ensta.fr</dc:identifier>
      <dc:language>eng</dc:language>
     </oai_dc:dc>
   </metadata>
  </record>
  <record>
   <header>
    <identifier>oai:physdoc:ftp://ftp.esi.ac.at/pub/Preprints/esi955.ps</id
    <datestamp>2002-01-25T00:00:00Z</datestamp>
   </header>
   <metadata>
     <oai_dc:dc
```

GetRecord

- Function
  - return single Record
- Parameter
  - **`identifier`** – unique ID for Record (R)
  - **`metadataPrefix`** – metadata format (R)
- Example URL
  - http://physnet.de/oai/oai2.php?verb=GetRecord &identifier=oai:test:123&metadataPrefix=oai_dc
- Errors/Exceptions
  - **`badArgument`**
  - **`cannotDisseminateFormat`**
  - **`idDoesNotExist`**

Date Ranges

Address dl/oai.pl?verb=ListIdentifiers&metadataPrefix=oai_dc&from=2001-06-26&until=2001-06-26 Go

```
<?xml version="1.0" encoding="UTF-8" ?>
- <OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
    http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
    <responseDate>2002-05-26T19:41:16Z</responseDate>
    <request verb="ListIdentifiers" metadataPrefix="oai_dc" from="2001-06-
      26" until="2001-06-26">http://rocky.dlib.vt.edu/~jcdlpix/cgi-
      bin/OAI2.0/beta2/jcdl/oai.pl</request>
  - <ListIdentifiers>
    - <header>
        <identifier>oai:JCDLPICS:200102dlb1</identifier>
        <datestamp>2001-06-26</datestamp>
        <setSpec>200102dlb</setSpec>
      </header>
    - <header>
        <identifier>oai:JCDLPICS:200102dlb2</identifier>
        <datestamp>2001-06-26</datestamp>
```

Agenda

- Part I - History and Overview
- Part II - OAI Serviceprovider - Example
- Part III - Technical Introduction
- Part IV - Implementation of Data Provider and Service Provider
- Part V - OAI Communities

Tutorial
Open Archive Initiative

Part IV
Implementation of
Data and Service Provider

Data- and Service Provider

- First questions
- Metadata
- Organisation
- Requirements of a Data-Provider
- Architecture
- Some Specialties
- Common problems
- Details for the Implementation
- Tools for Testing

General: First Questions

Data Provider

- What kind of data do I want to provide?
- (To which Service Providers will I offer my data?)

Service Provider

- What kind of service do I want to provide?
- From whom (Data Providers) do I want to collect data?
- What kind of metadata format do I want (need) to support?

Data Provider & Service Provider

- Do I need to have agreements on certain aspects?
- Metadata formats...

Metadata Mappings

- Data Provider must **map** its internal metadata to format, which it offers through OAI Interface.
- Unqualified Dublin Core is mandatory as least common denominator
  - http://dublincore.org/
  - Dublin Core Metadata Element Set has 15 Elements
  - Elements are optional, and can be repeated
  - Normally a Link to Resource is provided in the <identifier> Tag
- Source metadata formats are recommended
- Metadata formats of your own community are recommended

Organisation

- required: unqualified Dublin Core
- special subjects / communities: other metadata specifications may be required
  - describe resources in a specialised way
  - definition of an XML schema (publicly available for validation)
- define set hierarchy
  - sensible partitioning for selective harvesting
  - agreement between data providers and between data and service providers

Organisation (2)

- aggregated data providers
  - if harvested by a service provider, “sub data providers” should not be harvested by same SP (duplication ...)
- subject gateways
- selective harvesting if corresponding sets have been defined and implemented

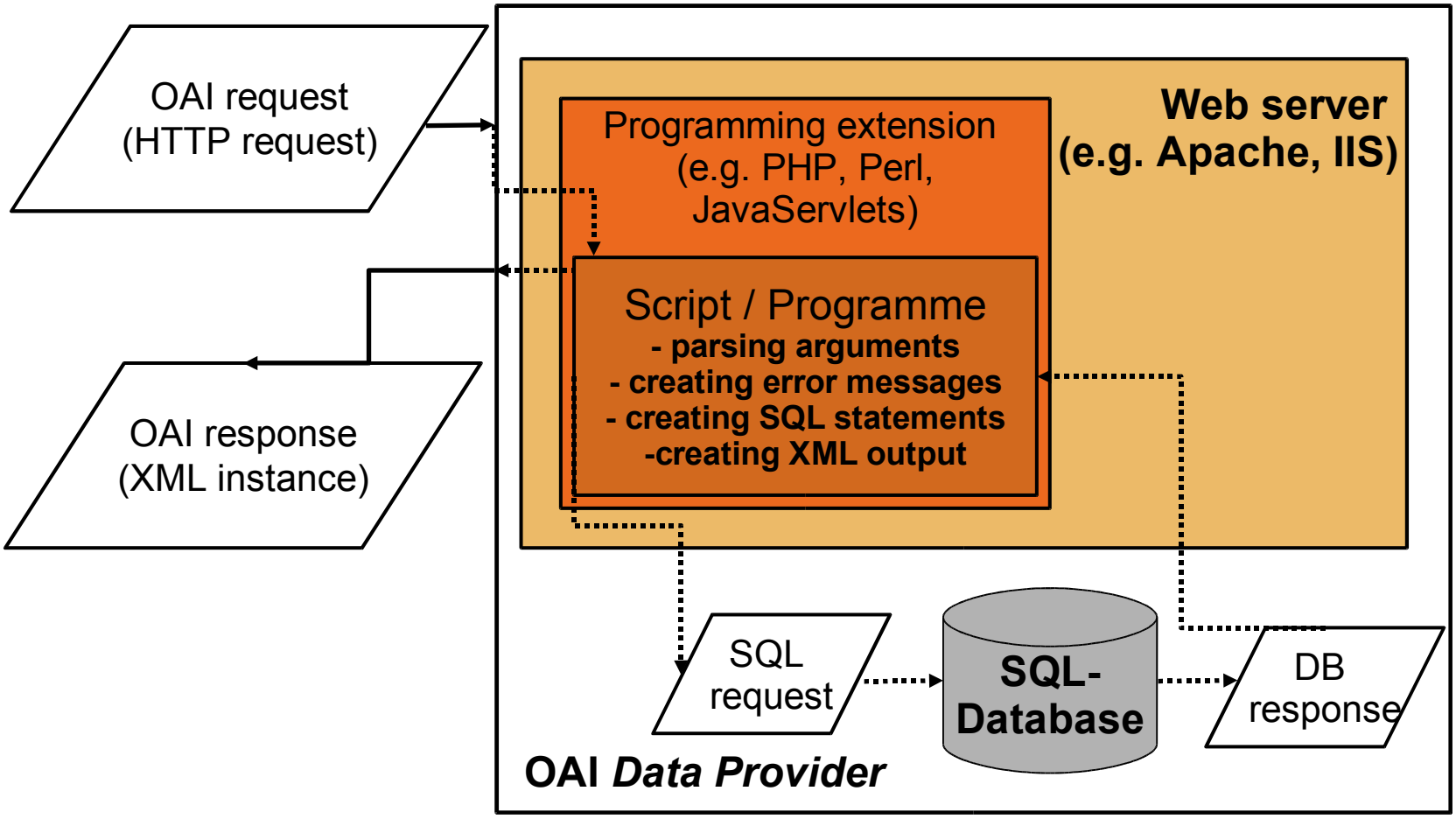
Server Technology

- WWW Server
- Protocol may be implemented in arbitrary form
  - CGI script (Perl, C++, Java)
  - Java servlet
  - PHP
- Metadata (e.g. database) access necessary
- See www.openarchives.org for list of software.

Metadata Sources

- Database in proprietary format, can be either SQL or XML databases
- Metadata collections in well-defined format(s)
  - files on disk
- Metadata can be extracted dynamically or statically from data
  - to serve XML, no storage of XML necessary
  - data from SQL database can be easily converted to XML on-the-fly

Data Provider: Architecture



Datestamps

- Needed for every record to support incremental harvesting
- Must be updated for every addition/modification/ deletion to ensure changes are correctly propagated
- Different from dates within the metadata – this date is used only for harvesting
- Can be either YYYY-MM-DD or YYYY-MM-DDThh:mm:ssZ (must be GMT timezone)

Unique Identifier

- Each record must have a unique identifier
- Identifiers must be valid URIs
- Example:
  - oai:<archiveId>:<recordId>
  - oai:etd.vt.edu:etd-1234567890
- Each identifier must resolve to a single record and always to the same record (for a given metadata format)

Deletions

- Archives may keep track of deleted records, by identifier and datestamp
- All protocol result sets can indicate deleted records
- If deletions are being tracked, this information must be stored indefinitely so as to correctly propagate to service providers with varying harvesting schedules

Details of the Implementation

- Required Tools
- Simple Program structure
- General structure
- Extensible metadata creation
- Encoding in XML
- Caching of Results
- Error handling
- Prevention of DOS (Denial-of-service)
- Creation of Resumption Tokens

Required Tools

- for new collections have a look at existing software
  - Eprints
  - Dspace
  - ETD software from VT
- to make existing collections OAI compliant
  - use web scripts
  - look for existing tools on
    - www.openarchives.org
    - http://edoc.hu-berlin.de/oai
    - http://physnet.de/oai
  - open source, easy to adapt to local needs.

Data Provider: General Structure

- Argument Parser
  - validates OAI requests
- Error Generator
  - creates XML responses with encoded error messages
- Database Query / Local Metadata Extraction
  - retrieves metadata from repository
  - according to the required metadata format
- XML Generator / Response Creation
  - creates XML responses with encoded metadata information
- Flow Control
  - realises incomplete list sequences for ‘larger’ repositories
  - uses resumption token as mechanism

Data Provider: Resumption Token

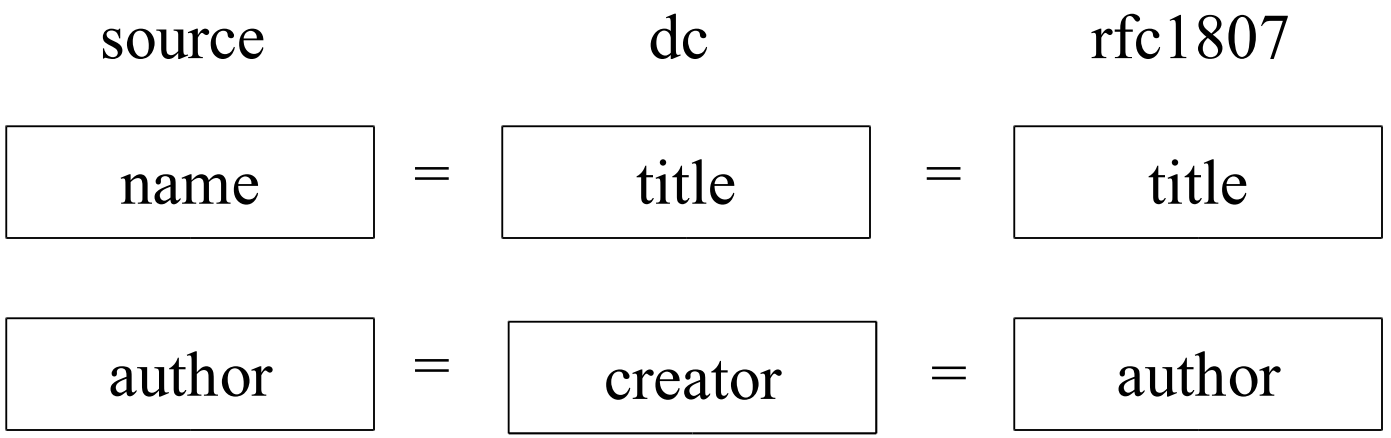
- should be implemented for “large” lists
- initiated by data provider
- store parameters (**`set`**, **`from`**, …) and number of already delivered records
- properties
  - expiration: expirationDate (optional)
  - completeListSize (optional)
  - already delivered records: cursor (optional)
  - recovery from network errors (possibility to re-issue most recent resumption token)
- problem
  - database changes
  - two possible solutions
    - duplicate data in a “request table”
    - store date of first request with the other parameters
    - use like additional until argument

Resumption Token

```
- <record>
  - <header>
    - <identifier>
        oai:physdoc:http://www.logos-verlag.de/cgi-local/buch?isbn=607
      </identifier>
      <datestamp>2002-01-25T00:00:00Z</datestamp>
    </header>
  - <metadata>
    - <oai_dc:dc xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
      http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
      - <dc:title>
          Die Natur der Naturwissenschaften historisch verstehen
        </dc:title>
        <dc:date>2001-01-29</dc:date>
        <dc:identifier>http://www.logos-verlag.de/cgi-local/buch?isbn=607</dc:identifier>
      </oai_dc:dc>
    </metadata>
  </record>
  <resumptionToken expirationDate="2003-03-27T00:01:10Z" completeListSize="319"
  cursor="0">664850492</resumptionToken>
</ListRecords>
</OAI-PMH>
```

Metadata Creation

- Approaches:
  - Map from source to each metadata format
  - Use crosswalks (maybe XSLT) to generate additional formats



Data Provider: Data Representation

- use recommended data representation
  - dates
    - ■ 2002-12-05
    - ✘ 2002-xx-xx, 2002, 05.12.2002
  - language code
    - ■ eng, ger, ...
    - ✘ en, de, english, german
- multi values: use own XML element for each entity
  - author
    - ■ <dc:creator>Smith, Adam</dc:creator> <dc:creator>Nash, John</dc:creator>
    - ✘ <dc:creator>Smith, Adam; Nash, John </dc:creator>

Encoding data for XML

- Special XML Characters must be escaped.
- Convert to UTF-8 (Unicode)
- Convert entities
- Remove unneccessary spaces
- Convert CR/LF for paragraphs
- URLs
  - /?#=&:;+ must be encoded as escape sequence

Data Provider: Compression

- method to reduce traffic and enhance performance
- optional for both sides: data and service providers
- handled on HTTP level
- harvesters may include an Accept-Encoding header in their requests –specifying preferences
- harvesters without Accept-Encoding header always receive uncompressed data
- repositories must support HTTP identity encoding
- repositories should specify supported encodings by including compression elements in the identify response

Error Handling

- All protocol errors are in XML format
  - **`badVerb`**
    illegal verb requested
  - **`badArgument`**
    illegal parameter values or combinations
  - **`badResumptionToken`**
    `cannotDisseminateFormat`
    `idDoesNotExist`
    parameters are in right format but are not legal under current conditions
  - **`noRecordsMatch`**
    `noMetadataFormats`
    `noSetHierarchy`
    empty response exception

Errors and Exceptions

```
- <OAI-PMH xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
  http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
    <responseDate>2003-03-26T00:06:56Z</responseDate>
    <request>http://physnet.uni-oldenburg.de/oai/oai2.php</request>
   - <error code="badVerb">
       The verb 'ListeAlles' provided in the request is illegal.
    </error>
  </OAI-PMH>
```

Prevention of Denial-of-Service

- Return only partial results and issue a resumption token for more
- Use 503 retry-after HTTP errors to have clients try again after a specified back-off time
- Use access control lists to limit who may access the archive
- Invoke an explicit delay before sending back results

Common Problems

- No unique identifiers !
- No datestamps !
- Incomplete information in database
- New metadata format
- XML responses not validating

No Unique Identifiers

- Create an independent identifier mapping
- Use row numbers for a database
- Use filenames for data in files
- Use a hash from other fields
- E.g. author+year+first word in title

No Datestamps

- Ignore the datestamp parameters and stamp all records with the current date
- Create a date table with the current date for all old entries and update dates for new entries
- Most Important: Any harvesting algorithm that is interoperably stable for an archive with real dates should be stable for an archive with synthesized dates

Incomplete Information

- Synthesize metadata fields based on a priori knowledge of the data
  - Example: publisher and language may be hard-coded for many archives
  - Omit fields that cannot be filled in correctly – better to have less information than incorrect information !

New Metadata Format

- Find the description, namespace and formal name of the standard
- Find an XML Schema description of the data format
  - If none exists, write one (consult other OAI people for assistance)
  - Create the mapping and test that it passes XML schema validation

Not Validating XML

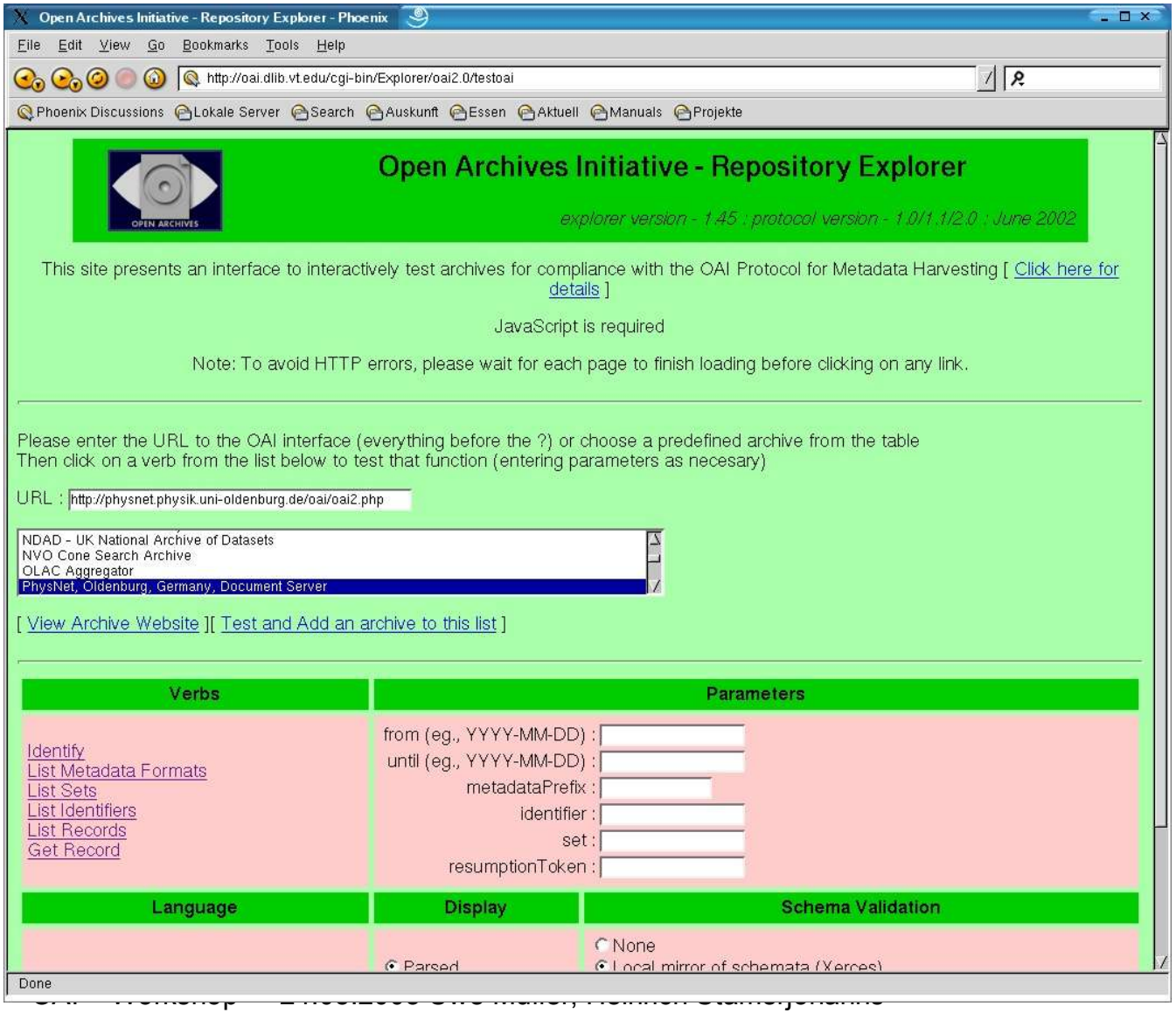
- Check namespaces and schema
- Use Repository Explorer in non-validating mode to check structure of XML, without looking at namespaces or schemata
- Validate schema by itself if it is non-standard
- Look at XML produced by other repositories
- Watch out for common character encoding issues (iso8859-1 --> utf-8)

Tools for Testing

- Repository Explorer
  - Interactive Browsing
  - Testing of parameters
  - Multiple views of data
  - Multilingual support
  - Automatic test suite
- OAI Registry
- XML Schema Validator

ETD 2003

Repository Explorer: Interactive Browsing



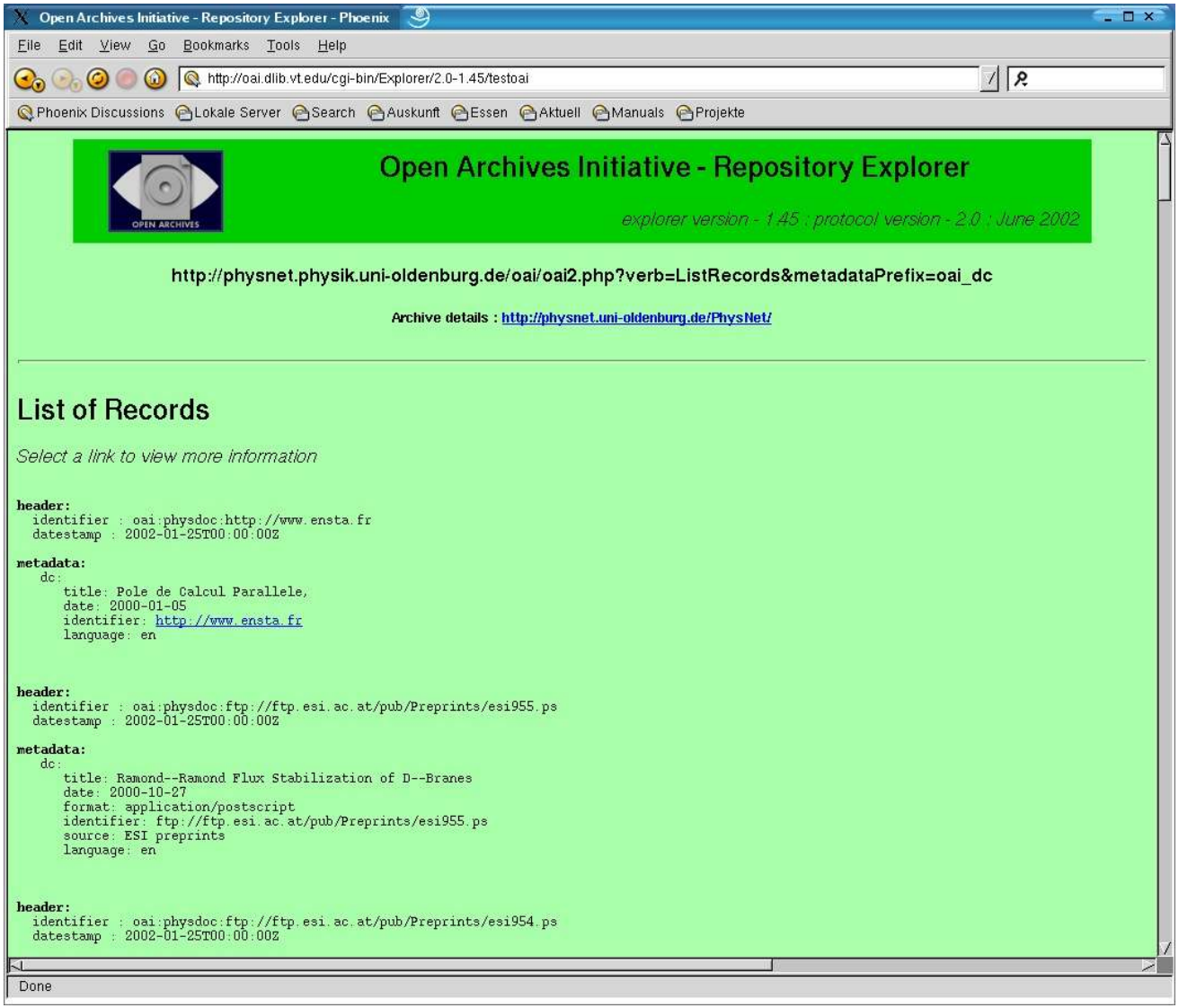
Repository Explorer: Parameter Test

Verbs	Parameters
Identify List Metadata Formats List Sets List Identifiers List Records Get Record	from (eg., YYYY-MM-DD) : until (eg., YYYY-MM-DD) : metadataPrefix : identifier : set : resumptionToken :

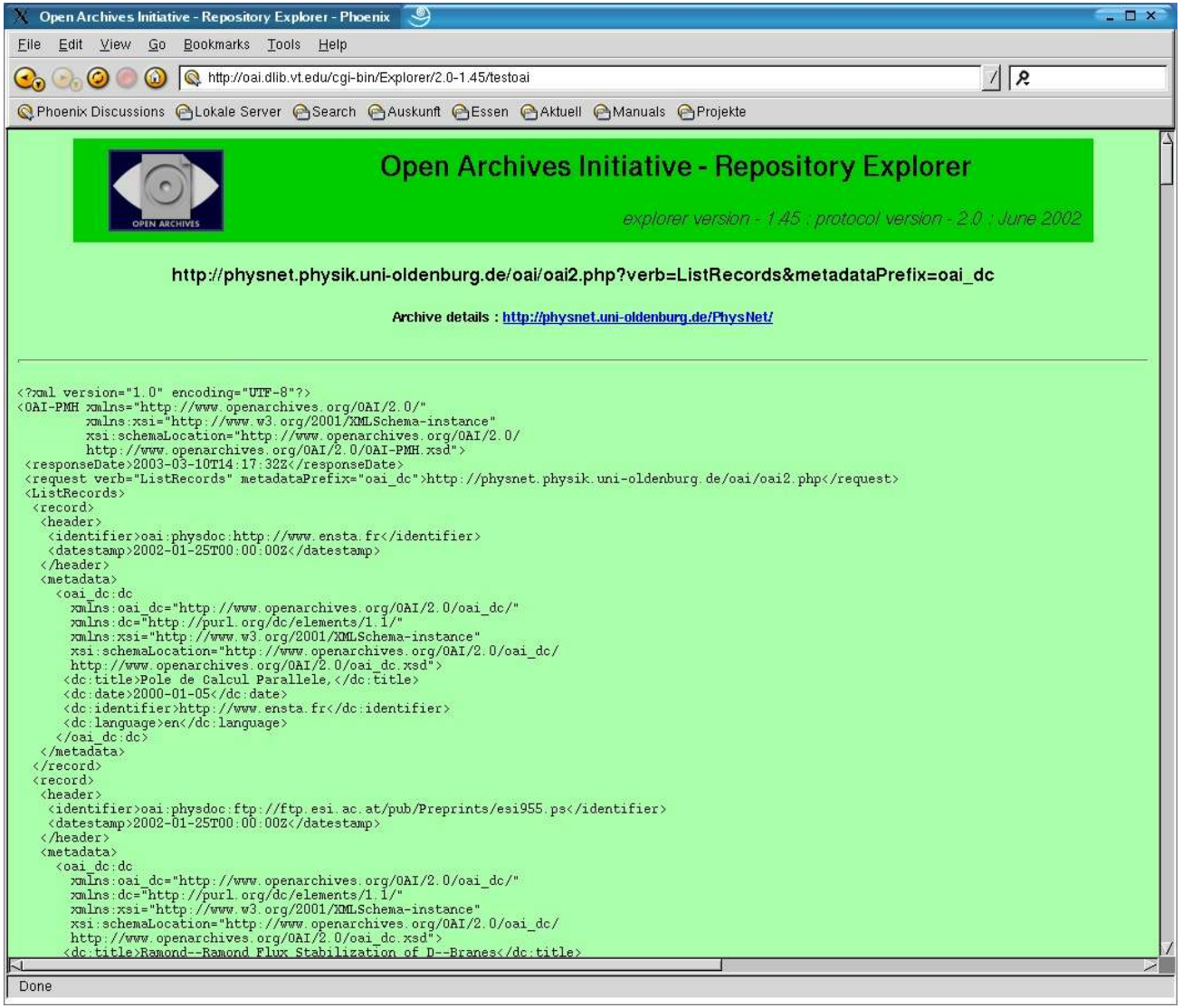
Language	Display	Schema Validation
English	(•) Parsed ( ) Raw XML ( ) Both	( ) None (•) Local mirror of schemata (Xerces) ( ) Online schemata (Xerces) ( ) Local mirror of schemata (XSV) ( ) Online schemata (XSV)

home about Send all comments to hussein@vt.edu --- Digital Library Research Laboratory@Virginia Tech

Repository Explorer: Browsing



RE: Presentation of XML



OAI Registry

The Open Archives Initiative

Registering as a Data Provider

Data providers who support the OAI protocol may choose to list their repository in the OAI registry. The goals of the registry are:

- Provide a publicly accessible list of OAI conformant repositories, making it easy for service providers to discover repositories from which metadata can be harvested.
- Provide a mechanism for data providers to ensure their conformance with the OAI protocol specification.
- Provide a means for the OAI to monitor use of the protocol and plan future activities and strategies.

This page allows you to register your repository by entering your BASE-URL in the text box at the bottom of this page. *Before* doing that, please read all of this instruction page so you understand what registration means and the choices you have.

Consequences of Registration
Protocol Testing
  Conformance Testing

XSV Schema Validator

W3C XML LTG

Validator for XML Schema REC (20010502) version

XSV version: XSV 2.3-1 of 2003/02/14 09:39:35

NOTICE: This is an Beta Test of a service for a **approved recommendation**. This version is for schema documents with the namespace URI `http://www.w3.org/2001/XMLSchema` and is being actively developed: see XSV for XML Schema 20000922 version for the no longer maintained previous version, for schema documents with the namespace URI `http://www.w3.org/2000/10/XMLSchema`, and XSV for XML Schema 200004007 version for the no longer maintained even earlier version, for schema documents with the namespace URI `http://www.w3.org/1999/XMLSchema`.

Use this form for checking a schema which is accessible via the Web, and/or schema-validating an instance with a schema of your own.

Address(es):

☐ Show warnings ☐ Keep Going ☐ Contribute

Service Provider

- Requirements
- Structure
- Architecture
- Harvesting
- Harvest Policies
- Intermediate systems
- Tools

Service Provider: Requirements

- internet connected server
- database system (relational or XML)
- programming environment
  - can issue HTTP requests to web servers
  - can issue database requests
  - XML parser

Service Provider: Structure (1)

Archive Management

- selection of archives to be harvested
- enter entries manually or
- automatically add / remove archives using the official registry

Request Component

- creates HTTP requests and sends them to OAI archives (data provider)
- demands metadata using the allowed verbs of the OAI-PMH
- possibly selective harvesting (set parameter)

Service Provider: Structure (2)

Scheduler

- realises timed and regular retrieval of the associated archives
- simplest case: manual initiation of the jobs
- else: e.g. cron job …

Flow Control

- resumption token: partitioning of the result list into incomplete sections – anew request to retrieve more results
- HTTP error 503 (service not available) – analysis of response to extract “retry-after” period

Service Provider: Structure (3)

Update Mechanism

- realises consolidation of metadata which have been harvested earlier (merge old and new data)
- easiest case: always delete all ‘old’ metadata of an archive before harvesting it
- reasonable: incremental update (from parameter) – insert new metadata and overwrite changed / deleted metadata (assignment using the unique identifiers)

XML Parser

- analyses the responses received from the archives
- validation: using the XML schema
- transforms the metadata encoded in XML into the internal data structure

Service Provider: Structure (4)

Normaliser and Mapper

- transforms data into a homogenous structure (different metadata formats)
- harmonises representation (e.g. date, author, language code)
- maps / translates different languages

Database

- mapping the XML structure of the metadata into a relational database (multi values …)
- or: use an XML database

Service Provider: Structure (5)

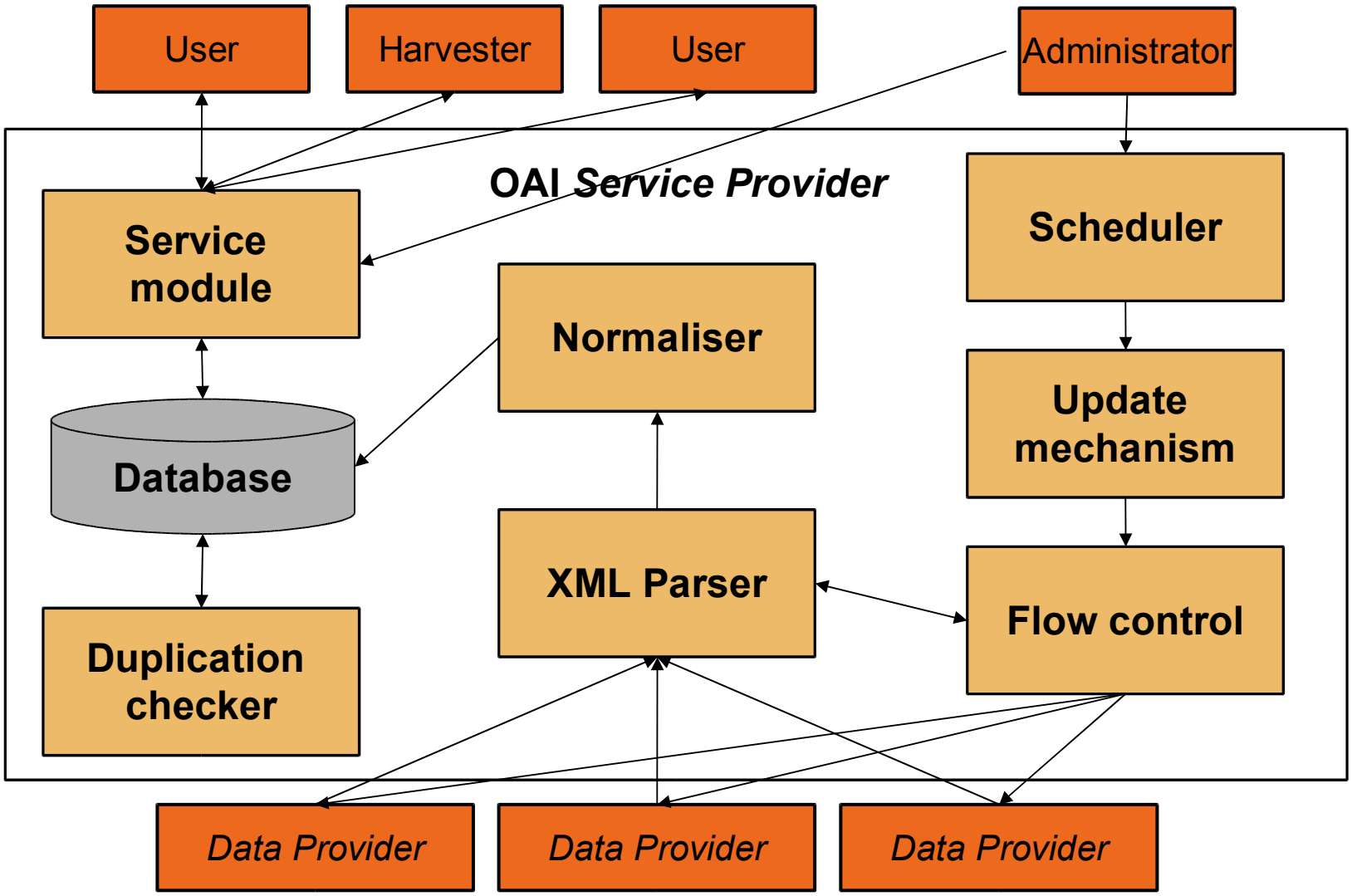
Duplication Checker

- merges identical records from different data providers
- possibility: unique identifier for the item (e.g. URN, …)
- but: often not easily practicable and not risk / error free

Service Module

- provides the actual service to the ‘public’
- basis: harvested and stored records of the associated archives
- uses only local database for requests etc.

Service Provider: Architecture



How to Harvest

- **`Identify`** to get basic information
- **`ListIdentifiers`**, followed by **`ListMetadataFormats`** for each record and then **`GetRecord`** for each id/metadata combination
  - No. of short HTTP requests = 1+n+n x m
    n=no. of identifiers, m=no. of metadata formats
- **`ListRecords`** for each metadata format required
- No. of long HTTP requests = m
  m=no. of metadata formats

Harvest Policies

- Use schedule for harvesting regularly
- Store date when last harvested (before you start)
- Use a two day overlap (or one day if your archive uses proper UTC datestamps)
  - New items may be added for the current day
  - Timezones create up to a day of lag if you ignore them
  - If the source uses correct UTC datestamps and second granularity then only 1 second of overlap is needed!
- Each time a record is encountered, erase previous instances

Intermediate Systems

- Both a data provider and service provider
- All harvested data must have the datestamps updated to the date on which the harvesting was done
- Identifiers retain their original values
- Note: Consistency in the source archive propagates, but so does inconsistency!

Tools

- Check OAI website for sample code
- XML parsers – depending on platform – check W3C
- XML Schema validators
  - Very few available – the reference version works but may not be easy to install
  - Ignore validation if you can trust the source
  - Sample data providers – check the OAI website for a list of conformant public archives

Agenda

- Part I - History and Overview
- Part II - OAI Serviceprovider - Example
- Part III - Technical Introduction
- Part IV - Implementation of Data Provider and Service Provider
- Part V - OAI Communities

Tutorial
Open Archive Initiative

Part V
OAI Communities

OAI Communities

- Shared Metadata Formats
- Shared semantics
- Closed OAI networks
- OAI within Digital Libraries

Shared Metadata Formats

- Use metadata formats accepted within a community to convey more specific information
- Examples
  - E-Print format (under development)
  - ETD-MS for theses and dissertations
  - VRA Core for multimedia
  - IMS Metadata for educational material

Shared Semantics

- Develop a shared understanding for the meanings of fields and sets
- Examples
  - Developing controlled vocabularies for fields
  - Using specific fields for external links (OAI recommends using identifier in DC for this)
  - Choosing from among existing standards (like language names)

Closed OAI Networks

- Data providers need not go public !
- Within an organization, OAI can be used for data transfer among heterogeneous systems
- More control over use, making global optimizations possible (like harvesting schedules and choice of metadata formats)

OAI within Digital Libraries

- OAI protocol may be used as basis for components to communicate
- Examples
  - Search Engines could use dynamic sets to correspond to search results
  - Browsing can be directed by sets
  - Reviews and Annotations can each be independent OAI data providers
- Open Digital Libraries project to investigate this approach:
  - http://oai.dlib.vt.edu/odl

Links

- Open Archives Initiative
  http://www.openarchives.org
- OAI Metadata Harvesting Protocol
  http://www.openarchives.org/OAI/openarchivesprotocol.htm
- Virginia Tech DLRL OAI Project
  http://www.dlib.vt.edu/projects/OAI/
- Repository Explorer
  http://purl.org/net/oai_explorer
- NDLTD
  http://www.ndltd.org

More Links

- ARC Cross-Archive Search Service
  http://arc.cs.odu.edu/
- XML Schema Validator
  http://www.w3.org/2001/03/webdata/xsv
- Dublin Core Metadata Initiative
  http://www.dublincore.org
- E-Prints DL-in-a-box
  http://www.eprints.org
- XML Tools at W3C
  http://www.w3.org/XML/#software

Summary

During today's tutorial we hope that you have

- gained an overview of the history behind the OAI-PMH and an overview of its key features
- been given a deeper technical insight into how the protocol works
- learned something about some of the main implementation issues
- found some useful starting points and hints that will help you as implementers

Thanks

Andy Powell, and Hussein Suleman whose Tutorials have been used as a base for this one.

Thank you.