# Building Interoperable Digital Libraries: A Practical Guide to creating Open Archives

Hussein Suleman, <u>hussein@vt.edu</u> Digital Library Research Laboratory Virginia Tech







### 1. Introduction



**\*** What is the OAI?

**\*** Motivation

# General System Strategy

**\*** History

Case study: NDLTD

# 1.1. What is the OAI?



\* What is the Open Archives Initiative (OAI)?

- Organization dedicated to solving problems of digital library interoperability by defining simple protocols, most recently for the exchange of metadata.
- **\*\*** What is the Protocol for Metadata Harvesting?
  - Protocol to transfer metadata from a source archive to a destination archive

## 1.2. Motivation



\* Existence of some established but independent archives

------

- \* Need for cross-archive services (like search engines)
- \* Lack of low-cost interoperability technology
- \* Experience from past projects such as Dienst



- - - <u>j</u> - - - - <u>j</u> - - - - - <u>j</u> - - - - <u>-</u> - <u>j</u> - - - - - <u>j</u> - - - - - <u>-</u> - <u>j</u> - - - - - - <u>-</u> - <u>j</u> - - - - - - <u>-</u> - <u>-</u> - - <u>-</u> - <u>-</u>



**Services** 

#### **Metadata Harvesting**

#### **Document Model**

**JCDL 2002** 

Slide 5

## 1.4. History



Santa Fe Meeting – October 1999

Santa Fe Convention, January 2000

- \* Workshops (ACM-DL 2000, ECDL 2000)
- \* Structuring of the OAI
  - Steering Committee
  - Technical Committee
- ✤ Open Days US/Europe
  - Protocol for Metadata Harvesting v1.0, January 2001
- ✗ Minor Update: v1.1 − July 2001
- **\*** Version 2.0 June 2002

# 1.5. Case Study: NDLTD

- -- -<u>ġ</u>- -- -<u>ġ</u>- -- -- <u>-</u>ġ- -- --



\* Networked Digital Library of Theses and Dissertations

\* Multiple independent university-based collections of electronic documents



**JCDL 2002** 

Slide 7

# 2. Definitions / Concepts



#### **Basic Principles**

- What is an Open Archive?
- Harvesting vs. Federation
- Metadata vs. Data
- Data and Service Providers
- Underlying Technology
  - HTTP and XML
  - XML, XML Namespaces and Schema
- \* Protocol Policies
  - Uniqueness and Persistence
  - What is a record?
  - Multiplicity of Metadata
  - Sets
  - Datestamp, Harvesting and Flow Control JCDL 2002

2.1. 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

\* ... a.k.a. OAI-Compliant Repository

**\*** No implications for:

- Physical storage of data
- Cost of data
- Metadata and data formats
- Access control to server

## 2.2. Harvesting vs. Federation

— — -**ķ**- — — -**ķ**- — — -**ķ**- — — -**ķ**- — --**ķ**-



\* 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

### 2.3. 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

# 2.4. Data and Service Providers



\* 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 higherlevel services to users (e.g. search engines)
- \* OAI uses these denotations for its client/server model (data=server, service=client)

# 2.5. HTTP and XML



- Metadata Harvesting Protocol is an almost stateless request/response protocol
- \* Requests and responses are sent via the HTTP protocol

— — -ġ- — -- -ġ- -- -- -ġ- -- -- -ġ-

- \* Requests are encoded as GET/POST operations
- \* Responses are well-formed XML documents

2.6. XML Namespaces and Schema



Consistency and data quality is ensured by using XML Schema descriptions for each possible response

\* XML Namespaces are used where necessary to clearly define which parts of the responses are actual metadata and which support the Metadata Harvesting Protocol

# 2.7. Uniqueness and Persistence



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

#### 2.8. 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
- \*\* OAI advocates harvesting of records, which contain metadata and additional fields to support the harvesting operation

# 2.9. Sample OAI Record



(note: schema and namespaces have been left out for clarity)

```
<record>
   <header>
       <identifier>oai:jcdl2002.org:tut1</identifier>
       <datestamp>2002-02-03</datestamp>
       <setSpec>tut</setSpec>
    </header>
    <metadata>
       \langle dc \rangle
          <title>OAI Tutorial at JCDL</title>
          <creator>Hussein Suleman</creator>
          <language>English</language>
       </dc>
    </metadata>
    <about>
       <metadataID>oai:jcdl2002.org:tut1md</metadataID>
    </about>
</record>
```

**JCDL 2002** 

# 2.10. Multiplicity of Metadata



\_\_\_\_

- Multiple formats of metadata allowed
- Dublin Core is mandatory
- \* Any other format allowed as long as it has an XML encoding
- \* E.g. MARC (Libraries), IMS (Education), ETDMS (Theses/Dissertations), RFC1807 (Bibliographies)

### 2.11. Sets



- \* Protocol mechanism to allow for harvesting of sub-collections
- \* No well-defined semantics depends completely on local data providers

\_\_\_\_\_i

- May be defined by arrangement between data providers and service providers
- \* E.g. Subject areas, years, author names, search queries

# 2.12. Datestamps & Harvesting



\* Each record needs a datestamp that indicates its date of creation or modification

\* Dates are used to allow for harvesting by date range, thus allowing incremental and continuous transfer of metadata from a data provider to a service provider

## 2.13. Flow Control



# HTTP "retry-after" mechanism can be leveraged to support server-side delaying of a client's request

\_\_\_\_<u>i</u>\_\_\_\_<u>i</u>\_\_\_\_\_i

Resumption Tokens can be used to return partial results – the client is issued with a token which may be presented to the server to receive more results



- 3. Requirements to be a Data Provider
- **\*** Source of metadata
- Server technology
- **\*** Datestamps
- **\*** Deletions
- **\*** Unique identifiers
- Metadata mappings

## 3.1. Source of Metadata



\* Database in proprietary format

- Collection of metadata records in well-defined format/s
  - Files on disk
- Metadata may be dynamically or statically extracted from data
- **\*** Synthetic collection

# 3.2. Server Technology



**\* WWW Server** 

\* Protocol may be implemented in many forms

• CGI Script (Perl, C++, Java)

Java Servlet

• PHP

\* Metadata (e.g. database) access mechanism required

See <u>www.openarchives.org</u> for list of publicly available software templates

\* See <u>www.dlib.vt.edu</u> for VT experimental software

## 3.3. 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)

# 3.4. Unique Identifiers



\* Each record must have a unique identifier

\_\_\_\_\_i\_\_\_\_\_

- # 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)

## 3.5. Deletions



\* Archives may keep track of deleted records, by identifier and datestamp

-<u>i</u>- -- -<u>i</u>- -- -<u>i</u>-

\* 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

# 3.6. Metadata Mappings



\* Data provider must map its metadata to the formats it chooses to provide through its OAI interface

- # Unqualified Dublin Core required
  - Best practice is to include a link in the <identifier> tag to the actual digital resource or at least a humanreadable web page
- \* Native formats recommended
- Community-based formats recommended



#### 4. Metadata Harvesting Protocol

#### **\*** Service Requests

- Identify
- ListMetadataFormats
- ListSets
- GetRecord
- ListIdentifiers
- ListRecords
- \* Metadata Multiplicity
- **\*** Date Ranges
- **\*** Resumption Tokens
- **\*** Error and Exceptions

JCDL 2002

# 4.1. Identify



\* Purpose

Return general information about the archive and its policies

**\*\*** Parameters

None

#### **\* Sample URL**

http://www.anarchive.org/cgi-bin/OAI?verb=Identify

- -- -<u>i</u>- -- -<u>i</u>- -- -<u>i</u>- -- -<u>i</u>- -- -<u>i</u>-

# 4.2. Identify - Response



Address 🛃 http://rocky.dlib.vt.edu/~jcdlpix/cgi-bin/0AI2.0/beta2/jcdl/oai.pl?verb=Identify

#### <?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>2002-05-26T19:28:31Z</responseDate> <responseDate>2002-05-26T19:28:31Z</responseDate> <request verb="Identify">http://rocky.dlib.vt.edu/~jcdlpix/cgibin/OAI2.0/beta2/jcdl/oai.pl</request>
  - <Identify>

<repositoryName>JCDL 2001 Picture Archive</repositoryName> <baseURL>http://rocky.dlib.vt.edu/~jcdlpix/cgi-

bin/OAI2.0/beta2/jcdl/oai.pl</baseURL>

<protocolVersion>2.0b2</protocolVersion>

<adminEmail>jcdlpix@rocky.dlib.vt.edu</adminEmail> <earliestDatestamp>1970-01-01T00:002</earliestDatestamp> <deletedRecord>no</deletedRecord>

<granularity>YYYY-MM-DD</granularity>

- + <description>
- ~/Identify>
- </OAI-PMH>

**JCDL 2002** 

## 4.3. ListMetadataFormats



\* Purpose

•List metadata formats supported by the archive as well as their schema locations and namespaces

— — -<u>ģ</u>. — - -<u>ģ</u>. — - -<u>ģ</u>. — - -<u>ģ</u>.

#### **\*\*** Parameters

identifier – for a specific record (O)

#### **\*\*** Sample URL

http://www.anarchive.org/cgi-bin/OAI?verb=ListMetadataFormats

#### 4.4. ListMetadataFormats - Response



Address 🔄 http://rocky.dlib.vt.edu/~jcdlpix/cgi-bin/0AI2.0/beta2/jcdl/oai.pl?verb=ListMetadataFormats

- <?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>2002-05-26T19:27:25Z</responseDate> <request verb="ListMetadataFormats">http://rocky.dlib.vt.edu/~jcdlpix/cgibin/OAI2.0/beta2/jcdl/oai.pl</request>
  - <ListMetadataFormats>
    - + <metadataFormat>
    - <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>

### 4.5. ListSets



\* Purpose

 Provide a hierarchical listing of sets in which records may be organized

#### **\*\*** Parameters

None

#### **\* Sample URL**

http://www.anarchive.org/cgi-bin/OAI?verb=ListSets

# 4.6. ListSets – Response



Address 🛃 http://rocky.dlib.vt.edu/~jcdlpix/cgi-bin/0AI2.0/beta2/jcdl/oai.pl?verb=ListSets



- <?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>2002-05-26T19:29:52Z</responseDate> <request verb="ListSets">http://rocky.dlib.vt.edu/~jcdlpix/cgibin/OAI2.0/beta2/jcdl/oai.pl</request>
  - <ListSets>
    - < <set>
      - <setSpec>200105dle</setSpec>
      - <setName>JCDL Day Four</setName>
      - <setDescription>
        - <oaidc:dc

JCDL 2002

## 4.7. GetRecord



\* Purpose

 Returns the metadata for a single identifier in the form of an OAI record

**\*\*** Parameters

identifier – unique id for record (R)

metadataPrefix – metadata format (R)

#### **\*** Sample URL

http://www.anarchive.org/cgi-bin/OAI?
 verb=GetRecord&identifier=oai:test:123&metadataPrefix=oai\_dc
#### 4.8. GetRecord - Response



Address 🙋 http://rocky.dlib.vt.edu/~jcdlpix/cgi-bin/OAI2.0/beta2/jcdl/oai.pl?verb=GetRecord&metadataPrefix=oai\_dc&i 💌

\_\_\_\_\_

```
i 🔽 🖉
```

<responseDate>2002-05-26T19:32:54Z</responseDate> <request verb="GetRecord" metadataPrefix="oai\_dc" identifier="oai:JCDLPICS:200105dle1">http://rocky.dlib.vt.edu/~jcdlpix/cgibin/OAI2.0/beta2/jcdl/oai.pl</request>

- <GetRecord>
  - <record>
    - <header>
      - <identifier>oai:JCDLPICS:200105dle1</identifier>
      - <datestamp>2001-06-27</datestamp>
      - <setSpec>200105dle</setSpec>
      - </header>
    - <metadata>
      - <oaidc:dc xmlns="http://purl.org/dc/elements/1.1/"
         xmlns:oaidc="http://www.openarchives.org/OAI/2.0/oai\_dc/"
         xmlns:xsi="http://www.openarchives.org/OAI/2.0/oai\_dc/
         http://www.openarchives.org/OAI/2.0/oai\_dc.xsd">
         title>O1ale1</title>
         <creator>Hussein Suleman</creator>
         <subject>JCDL Day Four</subject>
         <description>Jim French and Unmil Karadkar over lunch</description>
         <publisher>JCDL</publisher>
         <date>2001-06-27</date>
         <tt><tvpe>image</tvpe>



# 4.9. ListIdentifiers



\* Purpose

- List headers for all records corresponding to the specified parameters
- \* Parameters
  - from start date (O)
  - until end date (O)
  - set set to harvest from (O)
  - metadataPrefix metadata format to list identifiers for (R)
  - resumptionToken flow control mechanism (X)

✷ Sample URL

 http://www.anarchive.org/cgi-bin/OAI? verb=ListIdentifiers&metadataPrefix=oai\_dc

**JCDL 2002** 

## 4.10. ListIdentifiers - Response



Address 🛃 http://rocky.dlib.vt.edu/~jcdlpix/cgi-bin/0A12.0/beta2/jcdl/oai.pl?verb=ListIdentifiers&m 💌

#### • 🖉

#### <?xml version="1.0" encoding="UTF-8" ?>

- - <ListIdentifiers>
    - <header>
      - <identifier>oai:JCDLPICS:200105dle1</identifier>
      - <datestamp>2001-06-27</datestamp>
      - <setSpec>200105dle</setSpec>
      - </header>
    - <header>
      - <identifier>oai:JCDLPICS:200105dle2</identifier>
      - <datestamp>2001-06-27</datestamp>
      - <setSnec>200105dle</setSnec>
    - **JCDL 2002**

# 4.11. ListRecords



\* Purpose

- Retrieves metadata for multiple records
- \* Parameters
  - from start date (O)
  - until end date (O)
  - set set to harvest from (O)
  - resumptionToken flow control mechanism (X)

\_\_\_\_\_i\_\_\_\_\_\_

metadataPrefix – metadata format (R)

#### ₭ Sample URL

http://www.anarchive.org/cgi-bin/OAI?
 verb=ListRecord&metadataprefix=oai\_dc&from=2001-01-01

#### 4.12. ListRecords - Response



Address 🙋 du/~jcdlpix/cgi-bin/OAI2.0/beta2/jcdl/oai.pl?verb=ListRecords&metadataPrefix=oai\_dc 💌 🔗 Go

i∻Go ∐Links

xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd"> <responseDate>2002-05-26T19:37:37Z</responseDate> <request verb="ListRecords" metadataPrefix="oai\_dc">http://rocky.dlib.vt.edu/~jcdlpix/cgibin/OAI2.0/beta2/jcdl/oai.pl</request> - <ListRecords> - <record> - <header> <identifier>oai:JCDLPICS:200105dle1</identifier> <datestamp>2001-06-27</datestamp> <setSpec>200105dle</setSpec> </header> - <metadata> - <oaidc:dc xmlns="http://purl.org/dc/elements/1.1/"</p> xmlns:oaidc="http://www.openarchives.org/OAI/2.0/oai\_dc/" xmlns:xsi="http://www.w3.org/2001/XMLSchemainstance<sup>\*</sup> xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai dc/ http://www.openarchives.org/OAI/2.0/oai\_dc.xsd">

<title>01dle</title>

<creator>Hussein Suleman</creator>

<subject>JCDL Day Four</subject>

<description>Jim French and Unmil Karadkar over

lunch</description>

# 4.13. Metadata Multiplicity



- <record>
  - <header>
    - <identifier>oai:VTETD:etd-3123162539751141</identifier><datestamp>1997-04-22</datestamp>

- </header>
- <metadata>
  - <rfc1807 xmlns="http://info.internet.isi.edu:80/in-</p>
    - notes/rfc/files/rfc1807.txt"
    - xsi:schemaLocation="http://info.internet.isi.edu:80/innotes/rfc/files/rfc1807.txt
    - http://www.openarchives.org/OAI/rfc1807.xsd">
    - <bib-version>1</bib-version>
    - <id>etd-3123162539751141</id>
    - <entry>1997-04-22</entry>
    - <organization>Virginia Polytechnic Institute and State
      - University</organization>
    - <title>SMA-Induced Deformations In general Unsymmetric
      - Laminates</title>
    - <tvne>Thesis/Dissertation</tvne>

### 4.14. Date Ranges



Address 🛃 dl/oai.pl?verb=ListIdentifiers&metadataPrefix=oai\_dc&from=2001-06-26&until=2001-06-26 💌 🦷

i 🤗 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/cgibin/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>

```
<cotGnoc>200102dlb/cotGnoc>
```

# 4.15. Resumption Token



Address 🧧 dl/oai.pl?verb=ListIdentifiers&metadataPrefix=oai\_dc&from=2001-06-26&until=2001-06-26 🔽

<identifier>oai:JCDLPICS:200101dla9</identifier>
<datestamp>2001-06-26</datestamp>
<setSpec>200101dla</setSpec>

- </header>
- <header>
  - <identifier>oai:JCDLPICS:200101dla10</identifier>
  - <datestamp>2001-06-26</datestamp>
  - <setSpec>200101dla</setSpec>
  - </header>

<resumptionToken cursor="0" completeListSize="35">!2001-06-26! 2001-06-26!oai\_dc!30</resumptionToken>

- </ListIdentifiers>
- </OAI-PMH>

### 4.16. Errors and Exceptions



Address 🧧 dl/oai.pl?verb=ListIdentifiers&metadataPrefix=oai\_dc&from=2001-06-28&until=2001-06-28 🝷



#### <?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>2002-05-26T19:43:59Z</responseDate>
 </responseDate>2002-05-26T19:43:59Z</responseDate>
 </request verb="ListIdentifiers" metadataPrefix="oai\_dc" from="2001-06 28" until="2001-06-28">http://rocky.dlib.vt.edu/~jcdlpix/cgi bin/OAI2.0/beta2/jcdl/oai.pl</request>
 </requests verb="noRecordsMatch">The combination of the values of
 arguments results in an empty set</representation of the values of
 arguments results in an empty set</representation.com
 </representation="normalized"></representation="normalized"></representation="normalized"></representation="normalized"></representation="normalized"></representation="normalized"></representation="normalized"></representation="normalized"></representation="normalized"></representation="normalized"></representation="normalized"></representation="normalized"></representation="normalized"></representation="normalized"></representation="normalized"></representation="normalized"</representation="normalized"></representation="normalized"></representation="normalized"></representation="normalized"></representation="normalized"></representation="normalized"></representation="normalized"</representation="normalized"></representation="normalized"></representation="normalized"</representation="normalized"></representation="normalized"</representation="normalized"></representation="normalized"></representation="normalized"></representation="normalized"</representation="normalized"></representation="normalized"></representation="normalized"</representation="normalized"</representation="normalized"></representation="normalized"</representation="normalized"</representation="normalized"></representation="normalized"</representation="normalized"</representat

# 5. Implementation Details



- **\*** Tools Required
- Basic program layout
- Object-oriented approaches
- \* Extensible metadata generation
- \* Data cleaning
- \* Caching of results
- \* Error handling
- Denial-of-service prevention
- Creating resumption tokens

## 5.1. Tools Required



Code templates if available (available for many languages)

**\*** Basic programming environment

— — —**i**- — — —**i**- — — —**i**-

- **\*** XML generators (for non-trivial encoding)
- \* Database access libraries/drivers (e.g. DBI, ODBC, JDBC)

## 5.2. Basic program layout

-<u><u><u>i</u></u>----<u><u>i</u></u>-----<u>i</u>-</u>



parse WWW request to extract parameters if (verb='Identify') ProcessIdentify; else if (verb='ListMetadataFormats') ProcessListMetadataFormats; else if (verb='ListSets') ProcessListSets: else if (verb='GetRecord') ProcessGetRecord; else if (verb='ListIdentifiers') ProcessListIdentifiers: else if (verb='ListRecords') ProcessListRecords; else ReportError ('badVerb');

#### **JCDL 2002**

## 5.3. Object-Oriented Approaches



\* Cleaner separation of protocol, database access and metadata generation

- Example approaches
  - Each service request is handled by a object
    - Simpler incremental development
  - Protocol, Database and Metadata are objects
    - Greater portability of code

Inheritance from a basic OAI data provider

#### 5.4. Metadata Generation **\*** Approaches • Map from source to each metadata format •Use crosswalks (maybe XSLT) to generate additional formats rfc1807 dc source



# 5.5. Data Cleaning



\* Escape special XML characters

- Convert to UTF-8 version of Unicode
- **\*** Convert entity references
- Remove extraneous whitespace
- Convert CR/LF for paragraphs

**# URLs** 

•/?#=&:;+ must be encoded as escape sequences

# 5.6. Result Caching



\* For multiple requests from many clients or to handle partial result sets

- Keep temporary tables/files
- Expire temporary data when no longer needed
  Is this necessary to handle date-range requests where new items are added to the result set while harvesting is in progress?

# 5.7. 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

## 5.8. Denial-of-Service Prevention

----<u>-</u>



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

5.9. Creating resumptionTokens



\* Combine from/until/metadataPrefix/set and a record number indicator with delimiters into a sequential token For example:

from!until!metadataPrefix!set!recordnumber

•2000-01-01!2001-01-01!!All!100

\* Use a session manager with automatic expiry For example:

•vtetd14june10amsession12

# 6. Common Problems



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

# 6.1. 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

#### 6.2. No datestamps



\* Ignore the datestamp parameters and stamp all records with the current date

\_\_\_\_<u>i</u>\_\_\_\_i<u>\_\_</u>\_\_\_i\_\_

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 6.3. Incomplete information



Synthesize metadata fields based on a priori knowledge of the data

 Example: publisher and language may be hardcoded for many archives

\* Omit fields that cannot be filled in correctly – better to have less information than incorrect information !

#### 6.4. New metadata format



\* Find the description, namespace and formal name of the standard

\_ \_ \_ġ\_ \_ \_ \_ \_ġ\_ \_ \_ \_ \_ġ\_ \_ \_ \_ \_ \_ \_ \_ \_

- \* 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

### 6.5. XML not validating



\* 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 character encoding issues

#### 7. Tools for Testing **\*** Repository Explorer Interactive Browsing Testing of parameters Multiple views of data Multilingual support Automatic test suite **\*** OAI Registry **\*** XML Schema Validator



**JCDL 2002** 

#### 7.1. RE Interactive Browsing





#### **Open Archives Initiative - Repository Explorer**

explorer version - 1.44 : protocol version - 1.0/1.1/2.0b2 : May 2002

This site presents an interface to interactively test archives for compliance with the OAI Protocol for Metadata Harvesting [ <u>Click here for details</u> ]

JavaScript is required

Note: To avoid HTTP errors, please wait for each page to finish loading before clicking on any link.

Please enter the URL to the OAI interface (everything before the ?) or choose a predefined archive from the table Then click on a verb from the list below to test that function (entering parameters as necesary)

URL :

Humboldt University Berlin, Document Server JCDL Picture Album {1.1} A Celebration of Women Writers {1.1} AISRI (American Indian Studies Research Institute)

[ View Archive Website ][ Test and Add an archive to this list ]

**JCDL 2002** 





| Verbs  | Parameters  |  |  |
|--|---|--|--|
| <u>Identify</u><br><u>List Metadata Formats</u><br><u>List Sets</u><br><u>List Identifiers</u><br><u>List Records</u><br><u>Get Record</u> | from (eg., YYYY-MM-DD) :<br>until (eg., YYYY-MM-DD) :<br>metadataPrefix :<br>identifier :<br>set :<br>resumptionToken : |  |  |
| Language   | Display   | Schema Validation  |  |
| English  | ⊙ Parsed<br>⊂ Raw XML<br>⊂ Both   | <ul> <li>None</li> <li>Local mirror of schemata (Xerces)</li> <li>Online schemata (Xerces)</li> <li>Local mirror of schemata (XSV)</li> <li>Online schemata (XSV)</li> </ul> |  |
| home about Send all comments to hussein@vt.edu Digital Library Research Laboratory@Virginia Tech   |   |  |  |

**JCDL 2002** 

Slide 64

#### 7.3. RE Browsing



#### **Archive Self-Description**

| Repository Name            | JCDL 2001 Picture Archive   |  |
|----------------------------|---|--|
| Base URL                   | http://rocky.dlib.vt.edu/~jcdlpix/cgi-bin/OAI2.0/beta2/jcdl/oai.pl  |  |
| Protocol Version           | 2.0b2   |  |
| Admin Email                | jcdlpix@rocky.dlib.vt.edu   |  |
| Earliest Datestamp         | 1970-01-01T00:002   |  |
| Deleted Record<br>Handling | no  |  |
| Granularity                | YYYY-MM-DD  |  |
| Other Information          | <pre>description:<br/>toolkit:<br/>title: VTOAI Perl Data Provider<br/>author:<br/>name: Hussein Suleman<br/>email: hussein@vt.edu<br/>institution: Virginia Tech<br/>version: 3.04<br/>URL: http://www.dlib.vt.edu/projects/OAI/</pre> |  |

### 7.4. RE Browsing



#### List of Sets

Click on the link to list the contents

JCDL Day Four

set description:

dc:

description: Pictures taken during JCDL Day Four

JCDL Banquet

set description:

dc:

description: Pictures taken during JCDL Banquet

JCDL Day Three

JCDL 2002

Slide 66

### 7.5. RE Browsing



#### **List of Record Identifiers**

\_ \_\_ \_\_ \_**\_** \_\_ \_

```
Select a link to view more information
```

#### header:

identifier : oai:JCDLPICS:200105dle1
datestamp : 2001-06-27
setSpec : 200105dle

[display record in Dublin Core] [display metadata formats]

```
header:
   identifier : oai:JCDLPICS:200105dle2
   datestamp : 2001-06-27
   setSpec : 200105dle
```

[display record in Dublin Core] [display metadata formats]

JCDL 2002

Slide 67

# 7.6. RE Browsing



#### List of Metadata Formats

Click on the link to view schema

Prefix=[dc2] NameSpace=[http://www.openarchives.org/OAI/2.0/oai\_dc/] Schema=[http://www.openarchives.org/OAI/2.0/oai\_dc.xsd]

[Not a standard OAI metadata name] [display record]

Prefix=[oai\_dc] NameSpace=[http://www.openarchives.org/OAI/2.0/oai\_dc/] Schema=[<u>http://www.openarchives.org/OAI/2.0/oai\_dc.xsd]</u>

[display record]

**JCDL 2002** 

#### 7.7. RE Browsing

ġ<mark>→ --- -- ġ</mark>→ --- -- -- ġ→ --- -- -- ġ→ -



#### List of Fields

#### header:

identifier : oai:JCDLPICS:200105dle1
datestamp : 2001-06-27
setSpec : 200105dle

#### metadata:

#### dc:

```
title: 01dle1
creator: Hussein Suleman
subject: JCDL Day Four
description: Jim French and Unmil Karadkar over lunch
publisher: JCDL
date: 2001-06-27
type: image
format: image/jpeg
identifier: <u>http://rocky.dlib.vt.edu/~jcdlpix/pictures/200105dle/01dle1.jpg</u>
language: en-us
relation: <u>http://www.jcdl.org</u>
rights: unrestricted
```

## 7.8. RE Multiple views of data

— — -<u>İ</u> — — -<u>İ</u> — — -<u>İ</u> — — -<u>İ</u> — — -<u>İ</u> — — -- -<u>İ</u>



#### Raw XML Output

```
<?xml version="1.0" encoding="UTF-8"?>
```

<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi='

```
<responseDate>2002-05-26T19:59:35Z</responseDate>
<request verb="GetRecord" metadataPrefix="oai dc" identifier="oa:
```

```
<GetRecord>
<record>
<header>
<identifier>oai:JCDLPICS:200105dle1</identifier>
<datestamp>2001-06-27</datestamp>
<setSpec>200105dle</setSpec>
</header>
<metadata>
<oaidc:dc xmlns="http://purl.org/dc/elements/1.1/" xmlns:oaidc="}
<title>01dle1</title>
<creator>Hussein Suleman</creator>
<subject>JCDL Day Four</subject>
```

**JCDL 2002** 

## 7.9. RE Multilingual Support



OPEN ARCHIVES

"Explorer" Version - 1.44 : Protokollversion - 1.0/1.1/2.0b2 : M

Hier koennen Sie interaktiv die OAI-Kompatibilitaet ihrer Archive verifizieren. [Bitte hier klicken um D

JavaScript is required

Bemerkung: Um HTTP-Fehler zu vermeiden, warten Sie bitte bis eine Seite fertig geladen ist, bevor Sie klicken

Bitte geben Sie die URL zu dem OAI Interface ein (alles vor dem "?") oder waehlen sie ein vordefinierte: Tabelle

#### 7.10. RE Automatic Test Suite





#### **Open Archives Initiative - Repository Explorer**

explorer version - 1.44 : protocol version - 2.0b2 : May 2002

Open Archives Initiative :: Protocol for Metadata Harvesting v2.0b2 RE Protocol Tester 1.44 :: Virginia Tech DLRL :: May 2002

```
(1) Testing : Identify
URL : http://rocky.dlib.vt.edu/~jcdlpix/cgi-bin/OAI2.0/beta2/jcdl/oai.pl?verb=Identify
Test Result : OK
---- [ Repository Name = JCDL 2001 Picture Archive ]
---- [ Protocol Version = 2.0b2 ]
---- [ Base URL = http://rocky.dlib.vt.edu/~jcdlpix/cgi-bin/OAI2.0/beta2/jcdl/oai.pl ]
---- [ Admin Email = jcdlpix@rocky.dlib.vt.edu ]
---- [ Granularity = YYYY-MM-DD ]
```

```
(2) Testing : Identify (illegal_parameter)
URL : http://rocky.dlib.vt.edu/~jcdlpix/cgi-bin/OAI2.0/beta2/jcdl/oai.pl?verb=Identify&
Test Result : OK
```

```
(3) Testing : ListMetadataFormats
URL : http://rocky.dlib.vt.edu/~jcdlpix/cgi-bin/OAI2.0/beta2/jcdl/oai.pl?verb=ListMetada
Test Result : OK
---- [ Sample Metadata Format = dc2 ]
```
#### 7.11. RE Error in Response



#### **Archive Self-Description**

| Repository<br>Name   | Virginia Tech Electronic Thesis and Dissertation Collection   |  |
|--|---|--|
| Base URL   | http://oai.dlib.vt.edu:80/~hussein/cgi-bin/NDLTDErr1/VTETD.p1   |  |
| Protocol<br>Version  | 1.0   |  |
| Error: Missing field : <identify>/<adminemail></adminemail></identify> |   |  |
| Other<br>Information   | <pre>description:<br/>oai-identifier:<br/>scheme: oai<br/>repositoryIdentifier: VTETD<br/>delimiter: :<br/>sampleIdentifier: oai:VTETD:etd-171110282975860<br/>description:<br/>eprints:<br/>content:<br/>text: Theses and Dissertations produced by students at Virginia<br/>metadataPolicy:<br/>text: Theses and Dissertations produced by students at Virginia<br/>metadataPolicy:<br/>text: Metadata may be used by commercial and non-commercial user<br/>dataPolicy:<br/>text: Full texts are individually tagged and the rights statemer</pre> |  |

#### 7.12. RE Error in XML





explorer version - 1.1 : protocol version - 1.0 : April 2001

http://oai.dlib.vt.edu/~hussein/cgi-bin/NDLTDErr1/VTETD.pl?verb=Identify

#### XSD Schema/Instance Validation Error !

Errors in XML instance

```
<?xml version='1.0'?>
<xsv docElt='{http://www.openarchives.org/OAI/1.0/OAI_Identify)Identify' instanceAsses
<importAttempt URI='http://oai.dlib.vt.edu/OAI/1.0/OAI_Identify.xsd' namespace='http://
<importAttempt URI='http://oai.dlib.vt.edu/OAI/eprints.xsd' namespace='http://www.open
<invalid char='4' code='cvc-complex-type.1.2.4' line='11' resource='file:///tmp/file2V
<fsm>
<node id='1'>
<edge dest='2' label='{http://www.openarchives.org/OAI/1.0/OAI_Identify):responseDate'
</node>
<node id='2'>
<edge dest='3' label='{http://www.openarchives.org/OAI/1.0/OAI_Identify):requestURL'/>
</node>
<node id='3'>
```



# 7.13. 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 <u>BASE-URL</u> 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

**JCDL 2002** 

#### 7.14. OAI Registry





#### The Open Archives Initiative

List of Registered, OAI-Conformant Repositories

| This application allows you to browse the current list of<br>OAI conforming repositories. Currently there are 29 such<br>repositories. The table may be sorted either by the <u>OAI</u><br><u>Repository Identifier</u> or by the <u>Repository Name</u> .   | Sort repositories by:<br>Repository Name<br>OAI identifier |
|--|--|
| You may retrieve information about an OAI repository by<br>selecting one of the rows in the following table. You may<br>view the registration record from the database;<br>alternatively, if your browser can render XML, you may<br>issue the <u>Identify request</u> to the selected repository and<br>receive the current XML response. | • view registration record<br>• issue Identify request     |

| OAI Repository Identifier | Repository Name  |
|---------------------------|--|
| C celebration             | A Celebration of Women Writers                               |
| C anic                    | Alaska Native Language Center                                |
| O arXiv                   | arXiv  |
| O CDLCIAS                 | California International and Area Studies Digital Repository |
| C                         | 141, CICTTD  |
|                           |  |

#### 7.15. XSV Schema Validator





#### Validator for XML Schema 20000922 version, XML Output

XSV version: XSV 1.176/1.87 of 2001/02/16 16:38:43

**NOTICE:** This is an <u>ALPHA TEST</u> of a service for a <u>work-in-progress specification</u>. This version is for schema documents with the namespace URI http://www.w3.org/2000/10/XMLSchema and is being actively developed: see <u>XSV for XML Schema 200004007 version</u> for the no longer maintained previous 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):

🗖 Chaussianan 🗖 Kaan Caina 🗖 Cantributa



## 8. Service Providers



₭ How to Harvest

**\*** Policies

**\*** Intermediate systems

**\*** Tools

★ Case Study: ARC

Case Study: NDLTD

#### 8.1. 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

#### 8.2. 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

## 8.3. Intermediate Systems



\_\_\_\_\_i\_\_\_\_\_\_i\_\_\_

**\*** 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!

#### 8.4. 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

#### 8.5. Case Study: ARC





# 8.6. Case Study: NDLTD





#### 9. OAI Communities

\*\* Shared Metadata Formats
\*\* Shared semantics
\*\* Layering over OAI
\*\* Closed OAI networks
\*\* OAI within the DL



9.1. 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

#### 9.2. Shared Semantics



Develop a shared understanding for the meanings of fields

— — -**ķ- — — -ķ- — - - - - - - - - - - - - -**

\* 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)

## 9.3. Layering over OAI

— — -<u>ė</u>- — — -<u>è</u>- — — -<u>è</u>-



- Convert OAI records into more standard formats like MARC communications format
- Collapse multiple requests into one to make harvesting easier
- \*\* Name authority system (developed at OCLC) piggybacks name resolution over the OAI protocol

# 9.4. Closed OAI networks

— — -<u>è</u>- — — -<u>è</u>- — — -<u>è</u>-



\* 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)

# 9.5. OAI within the DL



Set the OAI protocol as the 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

# 10. Now What ? \* 1-2-3 Recipe \* Future of Metadata Harvesting Protocol \* Future of OAI \* Links





#### 10.1. 1-2-3 Recipe **\*** DO I REALLY WANT TO DO THIS? \* Do I have an accessible metadata source? \* Do I have a server to host the OAI script/program? \* Can I satisfy the requirements to be a data provider? \* Can I write the code or modify a template or hire a programmer to do either?

**JCDL 2002** 

#### 10.2. Future of Protocol



★ Version 2.0

Released ~ June 2001

SOAP versionEnvisioned for near future

#### 10.3. Future of OAI



\* Advocacy for easier access to information
\* New protocols/tools to support this mission
\* Research projects to test theory underlying current architecture e.g. Cyclades

#### 10.4. Links



Open Archives Initiative
 <u>http://www.openarchives.org</u>
 OAI Metadata Harvesting Protocol

\_\_\_\_\_i\_\_\_\_i\_\_

http://www.openarchives.org/OAI/openarchivesprotocol.htm

\* Virginia Tech DLRL OAI Projects

http://www.dlib.vt.edu/projects/OAI/

**\*** Repository Explorer

http://purl.org/net/oai\_explorer

\* NDLTD

http://www.ndltd.org

## 10.5. More Links



\*\* ARC Cross-Archive Search Service
 <u>http://arc.cs.odu.edu/</u>
 \*\* XML Schema Validator
 <u>http://www.w3.org/2001/03/webdata/xsv</u>
 \*\* 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



#### That's All Folks !

**JCDL 2002** 

Slide 97