

OAI Protocol for Metadata Harvesting



hussein suleman
uct cs honours 2009

What is the OAI ?

- What is the Open Archives Initiative (OAI)?
 - Organisation dedicated to solving problems of digital library interoperability by defining simple protocols, most recently for the exchange of metadata among repositories.
- What is the Protocol for Metadata Harvesting?
 - Protocol to transfer metadata from a source archive to a destination archive.

Protocol Specifics

□ Service Requests

- Identify
- ListMetadataFormats
- ListSets
- GetRecord
- ListIdentifiers
- ListRecords

□ Metadata Multiplicity

□ Date Ranges

□ Resumption Tokens

□ Error and Exceptions

Identify

□ Purpose

- Return general information about the archive and its policies

□ Parameters

- None

□ Sample URL

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

Identify - Response

Address  <http://scholar.lib.vt.edu/theses/OAI2/?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>2003-10-14T20:01:48Z</responseDate>
  <request verb="Identify">http://scholar.lib.vt.edu/theses/OAI2/</request>
- <Identify>
  <repositoryName>VT Electronic Thesis and Dissertation Archive</repositoryName>
  <baseURL>http://scholar.lib.vt.edu/theses/OAI2/</baseURL>
  <protocolVersion>2.0</protocolVersion>
  <adminEmail>mailto:webmaster@scholar.lib.vt.edu</adminEmail>
  <earliestDatestamp>1970-01-01T00:00:00Z</earliestDatestamp>
  <deletedRecord>no</deletedRecord>
  <granularity>YYYY-MM-DD</granularity>
+ <description>
- <description>
  - <eprints xmlns="http://www.openarchives.org/OAI/1.1/eprints"
    xsi:schemaLocation="http://www.openarchives.org/OAI/1.1/eprints
    http://www.openarchives.org/OAI/1.1/eprints.xsd">
    - <content>
      <text>Theses and Dissertations produced by students</text>
    </content>
    - <metadataPolicy>
      <text>Metadata may be used by commercial and non-commercial
        users</text>
    </metadataPolicy>
```

ListMetadataFormats

□ Purpose

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

□ Parameters



- identifier – for a specific record (O)

□ Sample URL

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

ListMetadataFormats - Response

Address  <http://scholar.lib.vt.edu/theses/OAI2/?verb=ListMetadataFormats>

 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>2003-10-14T20:02:51Z</responseDate>
  <request verb="ListMetadataFormats">http://scholar.lib.vt.edu/theses/OAI2/</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>
- <metadataFormat>
  <metadataPrefix>oai_marc</metadataPrefix>
  <schema>http://www.openarchives.org/OAI/1.1/oai_marc.xsd</schema>

  <metadataNamespace>http://www.openarchives.org/OAI/1.1/oai_marc</metadataNamespace:
</metadataFormat>
- <metadataFormat>
  <metadataPrefix>oai_rfc1807</metadataPrefix>
  <schema>http://www.openarchives.org/OAI/1.1/rfc1807.xsd</schema>
  <metadataNamespace>http://info.internet.isi.edu:80/in-
  notes/rfc/files/rfc1807.txt</metadataNamespace>
</metadataFormat>
```

ListSets

□ Purpose

- Provide a hierarchical listing of sets in which records may be organised


□ Parameters

- None

□ Sample URL

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

ListSets – Response

Address  <http://scholar.lib.vt.edu/theses/OAI2/?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>2003-10-14T20:03:45Z</responseDate>
  <request verb="ListSets">http://scholar.lib.vt.edu/theses/OAI2/</request>
- <ListSets>
  - <set>
    <setSpec>All</setSpec>
    <setName>All theses and dissertations</setName>
  </set>
</ListSets>
</OAI-PMH>
```

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`

GetRecord - Response

Address  holar.lib.vt.edu/theses/OAI2/?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:VTETD:etd-3345131939761081  

```
<?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>2003-10-14T20:07:39Z</responseDate>
  <request verb="GetRecord" metadataPrefix="oai_dc" identifier="oai:VTETD:etd-
    3345131939761081">http://scholar.lib.vt.edu/theses/OAI2/</request>
- <GetRecord>
- <record>
  - <header>
    <identifier>oai:VTETD:etd-3345131939761081</identifier>
    <timestamp>1997-03-31</timestamp>
    <setSpec>All</setSpec>
  </header>
- <metadata>
  - <oai_dc:dc xmlns="http://purl.org/dc/elements/1.1/"
    xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
    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">
    <title>Conceptual Development and Empirical Testing of an Outdoor
      Recreation Experience Model: The Recreation Experience Matrix (REM)
    </title>
    <creator>Walker, Gordon James</creator>
    <subject>Forestry</subject>
    <description>This dissertation examines four issues, including: (a) whether
```

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`

ListIdentifiers - Response

Address  http://scholar.lib.vt.edu/theses/OAI2/?verb=ListIdentifiers&metadataPrefix=oai_dc

```
<?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>2003-10-14T20:05:14Z</responseDate>
  <request verb="ListIdentifiers"
    metadataPrefix="oai_dc">http://scholar.lib.vt.edu/theses/OAI2/</request>
- <ListIdentifiers>
  - <header>
    <identifier>oai:VTETD:etd-3345131939761081</identifier>
    <datestamp>1997-03-31</datestamp>
    <setSpec>All</setSpec>
  </header>
  - <header>
    <identifier>oai:VTETD:etd-171110282975860</identifier>
    <datestamp>1997-03-13</datestamp>
    <setSpec>All</setSpec>
  </header>
  - <header>
    <identifier>oai:VTETD:etd-05012000-14030054</identifier>
    <datestamp>2000-05-01</datestamp>
```

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)
- metadataPrefix - metadata format (R)

□ Sample URL

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

ListRecords - Response

Address  http://scholar.lib.vt.edu/theses/OAI2/?verb=ListRecords&metadataPrefix=oai_dc

```
<?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>2003-10-14T20:09:30Z</responseDate>
  <request verb="ListRecords"
    metadataPrefix="oai_dc">http://scholar.lib.vt.edu/theses/OAI2/</request>
- <ListRecords>
  + <record>
  + <record>
  + <record>
  - <record>
    - <header>
      <identifier>oai:VTETD:etd-3621112139711101</identifier>
      <datestamp>1997-04-18</datestamp>
      <setSpec>All</setSpec>
    </header>
  - <metadata>
    - <oai_dc:dc xmlns="http://purl.org/dc/elements/1.1/"
      xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
      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">
      <title>Fermion Quantum Field Theory In Black-hole Spacetimes</title>
      <creator>Ahmad, Syed Alwi B.</creator>
      <subject>Physics</subject>
```

Metadata Multiplicity

Address  ir.lib.vt.edu/theses/OAI2/?verb=GetRecord&metadataPrefix=oai_etdms&identifier=oai:VTETD:etd-3345131939761081  

```
- <record>
- <header>
  <identifier>oai:VTETD:etd-3345131939761081</identifier>
  <timestamp>1997-03-31</timestamp>
  <setSpec>All</setSpec>
</header>
- <metadata>
- <oai_etdms:thesis
  xmlns="http://www.ndltd.org/standards/metadata/etdms/1.0/"
  xmlns:oai_etdms="http://www.ndltd.org/standards/metadata/etdms/1.0/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.ndltd.org/standards/metadata/etdms/1.0/
  http://www.ndltd.org/standards/metadata/etdms/1.0/etdms.xsd">
  <title>Conceptual Development and Empirical Testing of an Outdoor
    Recreation Experience Model: The Recreation Experience Matrix (REM)
  </title>
  <creator>Walker, Gordon James</creator>
  <subject>outdoor recreation</subject>
  <subject>recreation experience preference scales</subject>
  <subject>recreation experience matrix</subject>
  <subject>recreation opportunity spectrum</subject>
  <description>This dissertation examines four issues, including: (a) whether
    outdoor recreation experiences not included in the Recreation Experience
    Preference (REP) scales exist; (b) whether these experiences can be
    categorized using a framework called the Recreation Experience Matrix
    (REM); (c) how well the Recreation Opportunity Spectrum (ROS) variables
    of activity, setting, and expertise explain the types of experiences outdoor
```


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>
```

Errors and Exceptions

Address  http://scholar.lib.vt.edu/theses/OAI2/?verb=GetRecord&metadataPrefix=oai_dc

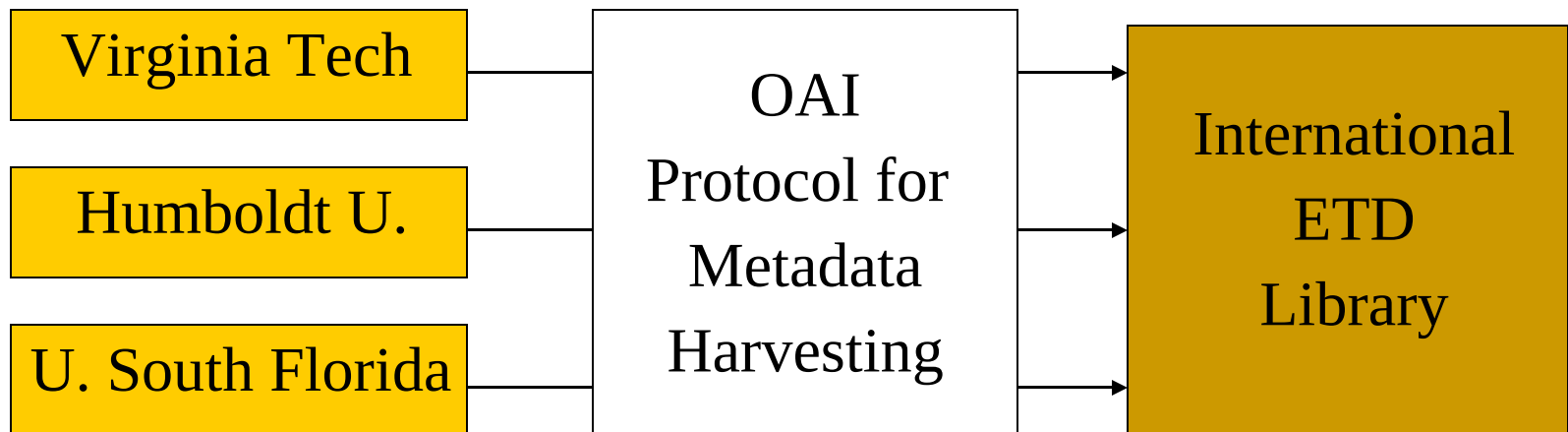
```
<?xml version="1.0" encoding="UTF-8" ?>
- <OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
  http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2003-10-14T20:11:31Z</responseDate>
  <request>http://scholar.lib.vt.edu/theses/OAI2/</request>
  <error code="badArgument">missing identifier parameter</error>
</OAI-PMH>
```

Motivation

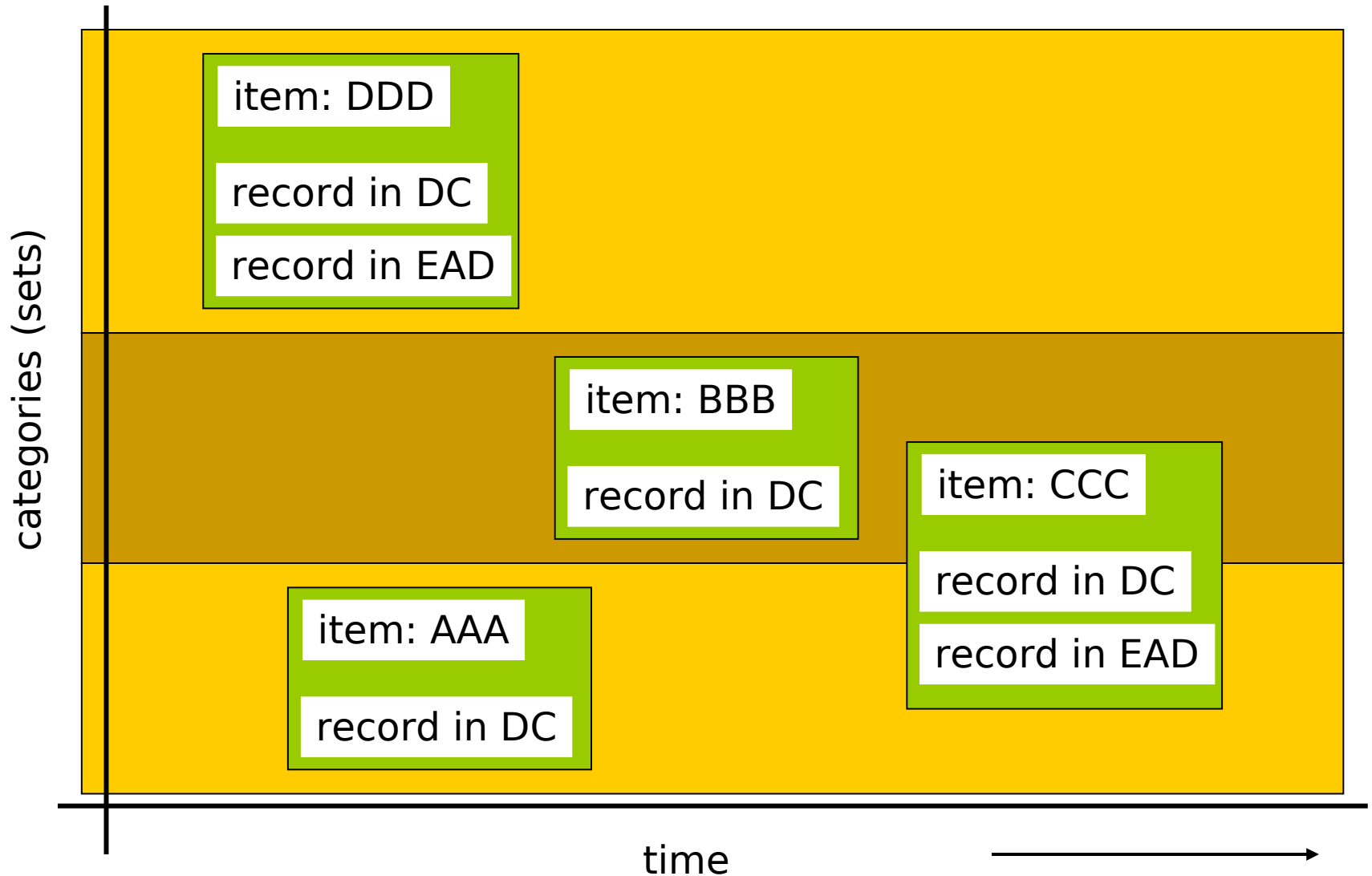
- ❑ Existence of some established but independent archives.
- ❑ Need for cross-archive services (like search engines).
- ❑ Lack of low-cost interoperability technology.
- ❑ Experience from past projects such as Dienst.

Case Study: NDLTD

- Networked Digital Library of Theses and Dissertations
- Made up of multiple independent university-based collections of electronic documents.



Multi-dimensional Data Model



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

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

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.

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

HTTP and XML

- Protocol for Metadata Harvesting is an almost stateless request/response protocol.
- Requests and responses are sent via the HTTP protocol.
- Requests are encoded as GET/POST operations.
- Responses are well-formed XML documents.

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 Protocol for Metadata Harvesting.

Uniqueness and Persistence

- Each record must be uniquely addressable by a distinct 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).

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.

Sample OAI Record

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

```
<record>
  <header>
    <identifier>oai:jcd12002.org:tut1</identifier>
    <datestamp>2002-02-03</datestamp>
    <setSpec>tut</setSpec>
  </header>
  <metadata>
    <dc>
      <title>Oldie-but-goodie example</title>
      <creator>Hussein Suleman</creator>
      <language>English</language>
    </dc>
  </metadata>
  <about>
    <metadataID>oai:jcd12002.org:tut1md</metadataID>
  </about>
</record>
```

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)

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.
- ❑ E.g. Subject areas, years, author names, search queries

Datestamps & Harvesting

- ❑ Each record needs a datestamp that indicates its date of creation/modification/deletion.
- ❑ 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)
- ❑ 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.

Flow Control

- HTTP “retry-after” mechanism can be leveraged to support server-side delaying of a client’s request.
- 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.

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.

Implementation Details

- ❑ Basic requirements
- ❑ Basic program layout
- ❑ Object-oriented approaches
- ❑ Extensible metadata generation
- ❑ Data cleaning
- ❑ Caching of results
- ❑ Error handling
- ❑ Denial-of-service prevention
- ❑ Creating resumption tokens

Basic Requirements

- You need a 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 www.openarchives.org for list of publicly available software templates.

Basic Program Layout

```
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');
```

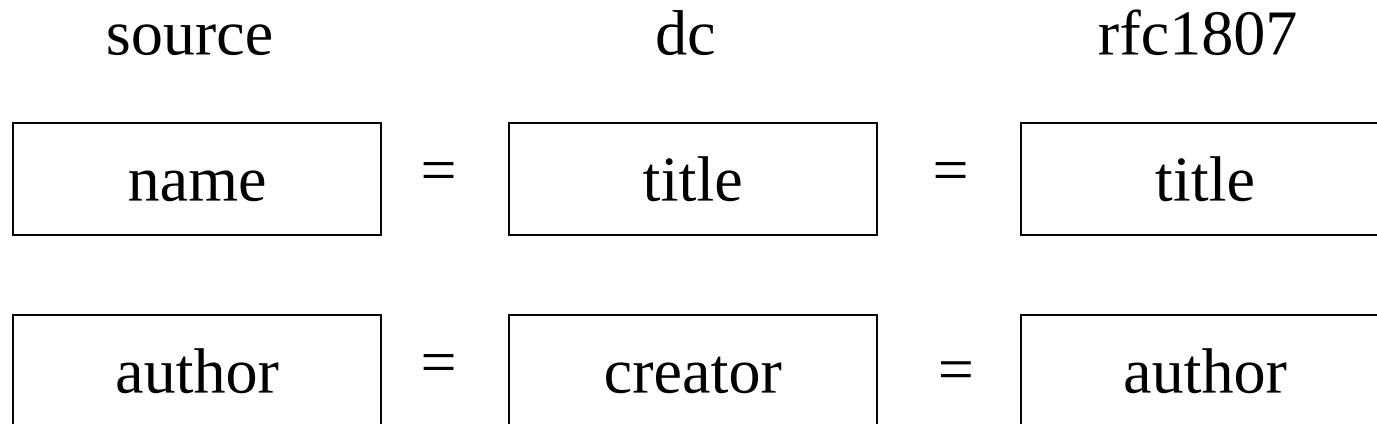
Object-Oriented Approaches

- 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

Metadata Generation

□ Approaches

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



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

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?

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

Denial-of-Service Prevention

- ❑ 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.

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

Tools for Testing

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

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 [[Click here for details](#)]

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 necessary)

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](#)]

RE Parameter Testing

Verbs		Parameters
Identify List Metadata Formats List Sets List Identifiers List Records Get Record	from (eg., YYYY-MM-DD) : <input type="text"/> until (eg., YYYY-MM-DD) : <input type="text"/> metadataPrefix : <input type="text"/> identifier : <input type="text"/> set : <input type="text"/> resumptionToken : <input type="text"/>	
Language	Display	Schema Validation
<input type="text" value="English"/>	<input checked="" type="radio"/> Parsed <input type="radio"/> Raw XML <input type="radio"/> Both	<input type="radio"/> None <input checked="" type="radio"/> Local mirror of schemata (Xerces) <input type="radio"/> Online schemata (Xerces) <input type="radio"/> Local mirror of schemata (XSV) <input type="radio"/> Online schemata (XSV)

[home](#) [about](#) Send all comments to hussein@vt.edu --- [Digital Library Research Laboratory@Virginia Tech](#)

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	1970-01-01T00:00:00Z
Deleted Record Handling	no
Granularity	YYYY-MM-DD
Other Information	<pre>description: toolkit: title: VTOAI Perl Data Provider author: name: Hussein Suleman email: hussein@vt.edu institution: Virginia Tech version: 3.04 URL: http://www.dlib.vt.edu/projects/OAI/</pre>

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](#)

RE Browsing

List of Record Identifiers

Select a link to view more information

header:

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

[\[display record in Dublin Core\]](#) [\[display metadata formats\]](#)

header:

identifier : oai:JCPLPICS:200105dle2
datestamp : 2001-06-27
setSpec : 200105dle

[\[display record in Dublin Core\]](#) [\[display metadata formats\]](#)

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=[http://www.openarchives.org/OAI/2.0/oai_dc.xsd]

[[display record](#)]

RE Browsing

List of Fields

header:

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

metadata:

dc:

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

RE Multiple views of data

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:JC DLPICS:200105dle1</identifier>
<datestamp>2001-06-27</datestamp>
<setSpec>200105dle</setSpec>
</header>
<metadata>
<oaidc:dc xmlns="http://purl.org/dc/elements/1.1/" xmlns:oaidc="1
  <title>0idle1</title>
  <creator>Hussein Suleman</creator>
  <subject>JC DL Day Four</subject>
```

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 vordefiniertes
Tabelle

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=ListMetadataFormats>

Test Result : OK

---- [Sample Metadata Format = dc2]

RE Error in Response

Archive Self-Description

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

RE Error in XML



Explorer

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/oai-identifier.xsd' namespace='http://w
<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'>
```

OAI Registry



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](#)

OAI Registry



The Open Archives Initiative

List of Registered, OAI-Conformant Repositories

This application allows you to browse the current list of OAI conforming repositories. Currently there are 29 such repositories. The table may be sorted either by the [OAI Repository Identifier](#) or by the [Repository Name](#).

Sort repositories by:

Repository Name

OAI identifier

You may retrieve information about an OAI repository by selecting one of the rows in the following table. You may view the registration record from the database; alternatively, if your browser can render XML, you may issue the [Identify request](#) to the selected repository and receive the current XML response.

- view registration record
- issue Identify request

OAI Repository Identifier	Repository Name
<input type="radio"/> celebration	A Celebration of Women Writers
<input type="radio"/> anlrc	Alaska Native Language Center
<input type="radio"/> arXiv	arXiv
<input type="radio"/> CDLCIAS	California International and Area Studies Digital Repository
<input type="radio"/>

Service Providers

- How to Harvest
- Policies
- Intermediate systems
- Case Study: ARC
- Case Study: NDLTD

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 = $1+n+n \times m$
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

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 timestamps).
 - 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 timestamps 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 timestamps 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!

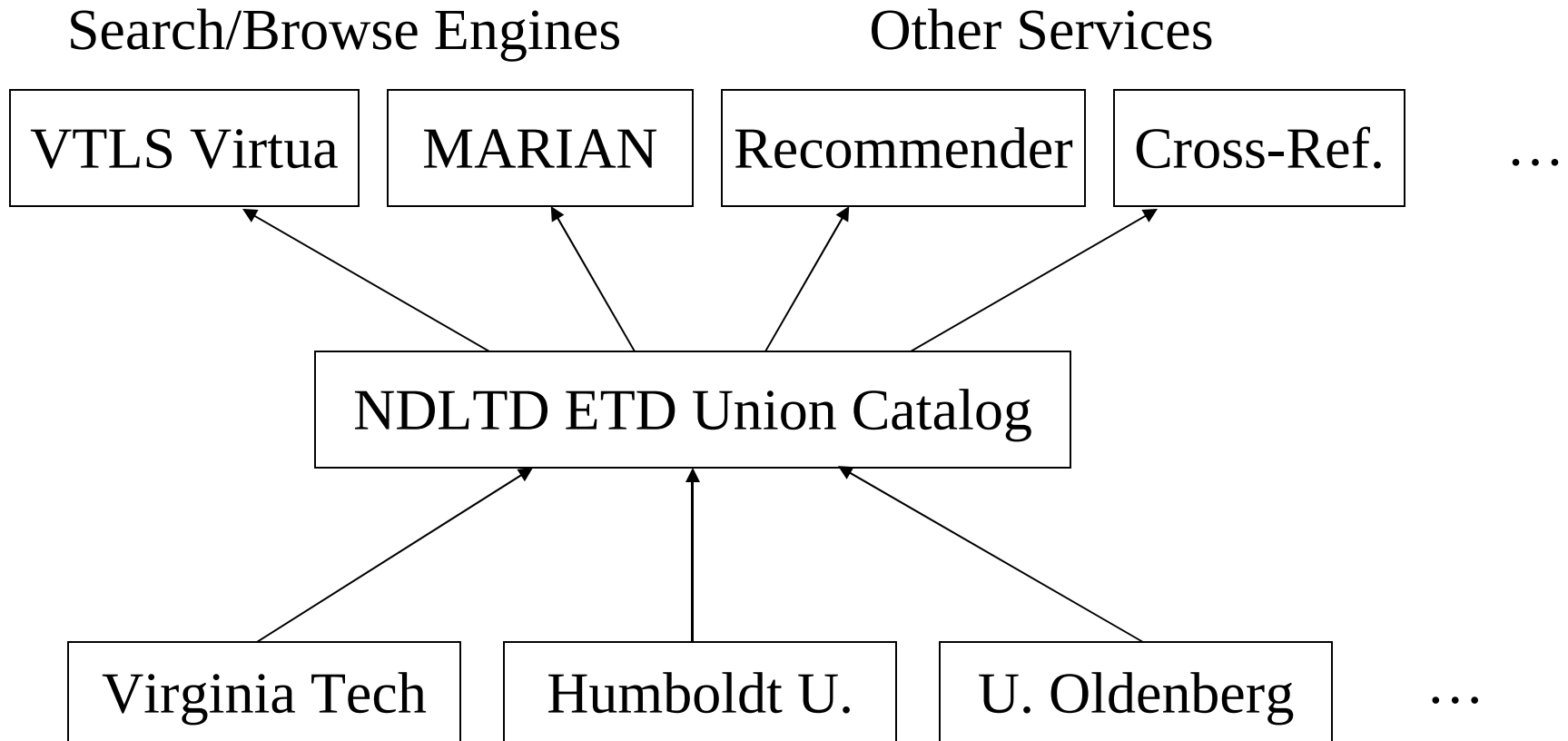
Case Study: ARC

The screenshot shows a Netscape browser window titled "ARC - Netscape". The address bar is empty. The page content includes the ARC logo (a red square with the letters "arc" in white), the text "Cross Archive Searching Service", and a logo for "OPEN ARCHIVES" (a stylized eye with a circular center). Below the logo is a navigation bar with "Simple search", "Advanced Search", and "Help" links. A message states "This is page 1, hits (1--10) of total 66 hits." Below this is a pagination link "Results Pages: 1 2 3 4 5 6 7". A black header bar contains the text "SEARCH RESULTS". The main content area displays a search result with the following fields:

Title	<i>Graphical Encoding for Information Visualization: Using Icon Color, Shape, and Size to Convey Nominal and Quantitative Data</i>
Creators	Nowell, Lucille Terry
Description	centerH2Graphical Encoding for Information Visualization: Using Icon Color, Shape, and Size To Convey Nominal and Quantitative Data/h2/center centerH3Lucille Terry Nowell/H3/centerBCenterABSTRACT/Center/BpIn producing a user interface design to visualize search results for a digital library called En
Archive	NDLTD
Document ID	oai:VTETD:etd-111897-163723

At the bottom of the page, a red banner contains the text: "This prototype is based on the UPS project and the NCSTRL+ based digital library developed by Old Dominion University". The browser's status bar at the bottom shows "Document: Done" and a taskbar with various system icons.

Case Study: ND LTD



References

- ❑ Lagoze, C., and Herbert Van de Sompel (2001) The open archives initiative: building a low-barrier interoperability framework, in *Proceedings of JCDL 2001*, 24-28 June, Roanoke, VA, USA, ACM Press, 54-62. Available <http://www.openarchives.org/documents/jcdl2001-oai.pdf>
- ❑ Lagoze, Carl, Herbert Van de Sompel, Simeon Warner and Michael Nelson (2001) The Open Archive Initiative Protocol for Metadata Harvesting, Open Archives Initiative. Available <http://www.openarchives.org/OAI/openarchivesprotocol.htm>
- ❑ NDLTD (2006) Website <http://www.ndltd.org>
- ❑ Old Dominion University (2006) ARC Cross-Archive Search Service. Website <http://arc.cs.odu.edu/>
- ❑ Open Archives Initiative (2006) Website <http://www.openarchives.org>
- ❑ Suleman, H (2006) Repository Explorer. Website http://purl.org/net/oai_explorer
- ❑ Thompson, Henry S., and Richard Tobin (2005) XML Schema Validator. Website <http://www.w3.org/2001/03/webdata/xsv>