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

Hussein Suleman, hussein@vt.edu

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 Metadata Harvesting Protocol?
  - 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

1.3. General System Strategy

Services

Metadata Harvesting

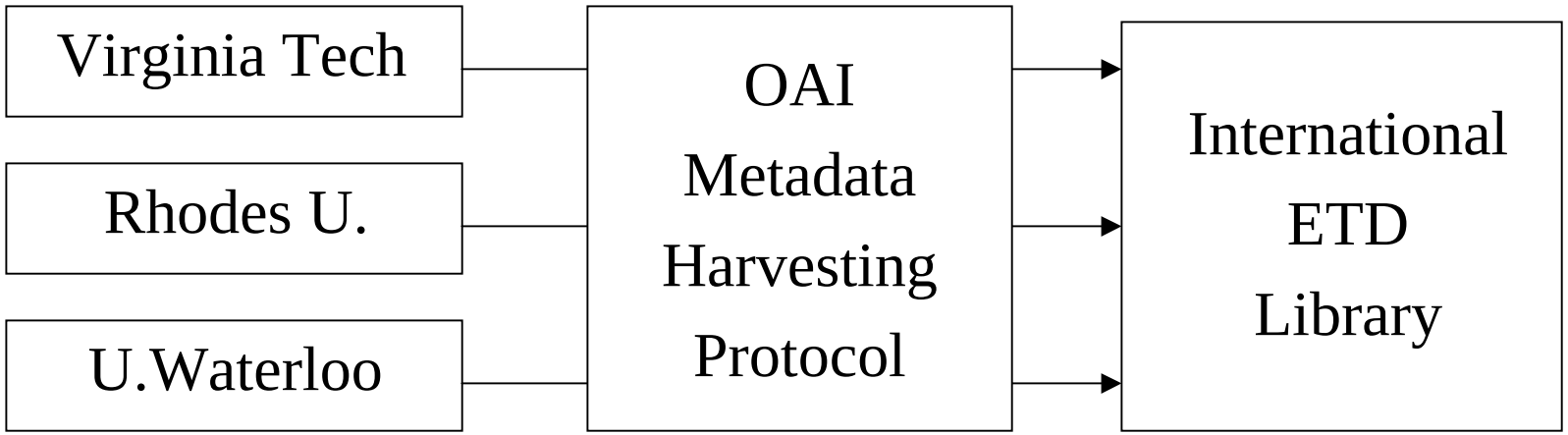
Document Model

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
  - Metadata Harvesting Protocol, January 2001

1.5. Case Study: NDLTD

- Networked Digital Library of Theses and Dissertations
- Multiple independent university-based collections of electronic documents



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

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
- … aka 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

- 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 higher-level 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
- 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 must 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

```
<record>
   <header>
      <identifier>oai:jcdl:tut3</identifier>
      <datestamp>2001-02-03</datestamp>
   </header>
   <metadata>
      <dc>
         <title>OAI Tutorial at JCDL</title>
         <creator>Hussein Suleman</creator>
         <language>English</language>
      </dc>
   </metadata>
   <about>
      <metadataID>oai:jcdl:tut3md</metadataID>
   </about>
</record>
```

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
- 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
- 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 www.openarchives.org for list of publicly available software templates
- See www.dlib.vt.edu 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

3.4. Unique Identifiers

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

3.5. Deletions

- Archives must keep track of deleted records, by identifier and datestamp
- All protocol result sets can indicate deleted records
- Deletions 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 to a human-readable page in the <identifier> tag
- 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

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

4.2. Identify - Response


```
<?xml version="1.0" encoding="UTF-8" ?>
- <Identify xmlns="http://www.openarchives.org/OAI/1.0/OAI_Identify"
    xmlns:xsi="http://www.w3.org/2000/10/XMLSchema-instance"
    xsi:schemaLocation="http://www.openarchives.org/OAI/1.0/OAI_Identify
    http://www.openarchives.org/OAI/1.0/OAI_Identify.xsd">
    <responseDate>2001-06-14T15:09:40-05:00</responseDate>
    <requestURL>http://oai.dlib.vt.edu:80/~hussein/cgi-bin/NDLTD/VTETD.pl?
      verb=Identify</requestURL>
    <repositoryName>Virginia Tech Electronic Thesis and Dissertation
      Collection</repositoryName>
    <baseURL>http://oai.dlib.vt.edu:80/~hussein/cgi-
      bin/NDLTD/VTETD.pl</baseURL>
    <protocolVersion>1.0</protocolVersion>
    <adminEmail>mailto:hussein@vt.edu</adminEmail>
  - <description>
    - <oai-identifier xmlns="http://www.openarchives.org/OAI/oai-identifier"
        xsi:schemaLocation="http://www.openarchives.org/OAI/oai-identifier
        http://www.openarchives.org/OAI/oai-identifier.xsd">
        <scheme>oai</scheme>
        <repositoryIdentifier>VTETD</repositoryIdentifier>
        <delimiter>:</delimiter>
        <sampleIdentifier>oai:VTETD:etd-171110282975860</sampleIdentifier>
      </oai-identifier>
    </description>
  - <description>
    + <eprints xmlns="http://www.openarchives.org/OAI/eprints"
        xsi:schemaLocation="http://www.openarchives.org/OAI/eprints
        http://www.openarchives.org/OAI/eprints.xsd">
    </description>
  </Identify>
```

4.3. 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

4.4. ListMetadataFormats - Response


```
  <?xml version="1.0" encoding="UTF-8" ?>
- <ListMetadataFormats
    xmlns="http://www.openarchives.org/OAI/1.0/OAI_ListMetadataFormats"
    xmlns:xsi="http://www.w3.org/2000/10/XMLSchema-instance"
    xsi:schemaLocation="http://www.openarchives.org/OAI/1.0/OAI_ListMetadataFormats
    http://www.openarchives.org/OAI/1.0/OAI_ListMetadataFormats.xsd">
    <responseDate>2001-06-14T15:12:53-05:00</responseDate>
    <requestURL>http://oai.dlib.vt.edu:80/~hussein/cgi-bin/NDLTD/VTETD.pl?
      verb=ListMetadataFormats</requestURL>
  - <metadataFormat>
      <metadataPrefix>oai_rfc1807</metadataPrefix>
      <schema>http://www.openarchives.org/OAI/rfc1807.xsd</schema>
      <metadataNamespace>http://info.internet.isi.edu:80/in-
        notes/rfc/files/rfc1807.txt</metadataNamespace>
    </metadataFormat>
  - <metadataFormat>
      <metadataPrefix>oai_dc</metadataPrefix>
      <schema>http://www.openarchives.org/OAI/dc.xsd</schema>
      <metadataNamespace>http://purl.org/dc/elements/1.1/</metadataNamespace>
    </metadataFormat>
  + <metadataFormat>
  + <metadataFormat>
  </ListMetadataFormats>
```

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


```
  <?xml version="1.0" encoding="UTF-8" ?>
- <ListSets xmlns="http://www.openarchives.org/OAI/1.0/OAI_ListSets"
    xmlns:xsi="http://www.w3.org/2000/10/XMLSchema-instance"
    xsi:schemaLocation="http://www.openarchives.org/OAI/1.0/OAI_ListSets
    http://www.openarchives.org/OAI/1.0/OAI_ListSets.xsd">
    <responseDate>2001-06-14T15:14:30-05:00</responseDate>
    <requestURL>http://oai.dlib.vt.edu:80/~hussein/cgi-bin/NDLTD/VTETD.pl?
      verb=ListSets</requestURL>
  - <set>
      <setSpec>All</setSpec>
      <setName>All theses and dissertations</setName>
    </set>
  </ListSets>
```

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


```
  <?xml version="1.0" encoding="UTF-8" ?>
- <GetRecord xmlns="http://www.openarchives.org/OAI/1.0/OAI_GetRecord"
    xmlns:xsi="http://www.w3.org/2000/10/XMLSchema-instance"
    xsi:schemaLocation="http://www.openarchives.org/OAI/1.0/OAI_GetRecord
    http://www.openarchives.org/OAI/1.0/OAI_GetRecord.xsd">
    <responseDate>2001-06-14T15:16:09-05:00</responseDate>
    <requestURL>http://oai.dlib.vt.edu:80/~hussein/cgi-bin/NDLTD/VTETD.pl?
      verb=GetRecord&identifier=oai:VTETD:etd-
      3123162539751141&metadataPrefix=oai_dc</requestURL>
  - <record>
    - <header>
        <identifier>oai:VTETD:etd-3123162539751141</identifier>
        <datestamp>1997-04-22</datestamp>
      </header>
    - <metadata>
      - <dc xmlns="http://purl.org/dc/elements/1.1/"
          xsi:schemaLocation="http://purl.org/dc/elements/1.1/
          http://www.openarchives.org/OAI/dc.xsd">
          <title>SMA-Induced Deformations In general Unsymmetric
            Laminates</title>
          <creator>Dano, Marie-Laure</creator>
          <subject>Engineering Science and Mechanics</subject>
          <description>General unsymmetric laminates exhibit large natural
            curvatures at room temperature. Additionally, inherent to most
            unsymmetric laminates is the presence of two stable configurations.
            Multiple configurations and stability issues arise because of the
            geometric nonlinearities associated with the large curvatures. The
            laminate can be changed from one stable configuration to the other by
```

4.9. ListIdentifiers

- Purpose
  - List all unique identifiers corresponding to records in the repository
- Parameters
  - from – start date (O)
  - until – end date (O)
  - set – set to harvest from (O)
  - resumptionToken – flow control mechanism (X)
- Sample URL
  - http://www.anarchive.org/cgi-bin/OAI?verb=ListIdentifiers&set=All

4.10. ListIdentifiers - Response


```
  <?xml version="1.0" encoding="UTF-8" ?>
- <ListIdentifiers xmlns="http://www.openarchives.org/OAI/1.0/OAI_ListIdentifiers"
    xmlns:xsi="http://www.w3.org/2000/10/XMLSchema-instance"
    xsi:schemaLocation="http://www.openarchives.org/OAI/1.0/OAI_ListIdentifiers
    http://www.openarchives.org/OAI/1.0/OAI_ListIdentifiers.xsd">
    <responseDate>2001-06-14T15:17:32-05:00</responseDate>
    <requestURL>http://oai.dlib.vt.edu:80/~hussein/cgi-bin/NDLTD/VTETD.pl?
      verb=ListIdentifiers</requestURL>
    <identifier>oai:VTETD:etd-3345131939761081</identifier>
    <identifier>oai:VTETD:etd-171110282975860</identifier>
    <identifier>oai:VTETD:etd-05012000-14030054</identifier>
    <identifier>oai:VTETD:etd-3621112139711101</identifier>
    <identifier>oai:VTETD:etd-133422039701091</identifier>
    <identifier>oai:VTETD:etd-23281533974920</identifier>
    <identifier>oai:VTETD:etd-123322282975860</identifier>
    <identifier>oai:VTETD:etd-255314202974780</identifier>
    <identifier>oai:VTETD:etd-335713312971890</identifier>
    <identifier>oai:VTETD:etd-104722369631841</identifier>
    <identifier>oai:VTETD:etd-831102339731121</identifier>
    <identifier>oai:VTETD:etd-454016449701231</identifier>
    <identifier>oai:VTETD:etd-3034112939721181</identifier>
    <identifier>oai:VTETD:etd-522014589642481</identifier>
    <identifier>oai:VTETD:etd-274210359611541</identifier>
    <identifier>oai:VTETD:etd-3210192049721391</identifier>
    <identifier>oai:VTETD:etd-0521318109613220</identifier>
    <identifier>oai:VTETD:etd-310141259631631</identifier>
    <identifier>oai:VTETD:etd-12164379662151</identifier>
```

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)
  - 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


```
  <?xml version="1.0" encoding="UTF-8" ?>
- <ListRecords xmlns="http://www.openarchives.org/OAI/1.0/OAI_ListRecords"
    xmlns:xsi="http://www.w3.org/2000/10/XMLSchema-instance"
    xsi:schemaLocation="http://www.openarchives.org/OAI/1.0/OAI_ListRecords
    http://www.openarchives.org/OAI/1.0/OAI_ListRecords.xsd">
    <responseDate>2001-06-14T15:19:19-05:00</responseDate>
    <requestURL>http://oai.dlib.vt.edu:80/~hussein/cgi-bin/NDLTD/VTETD.pl?
      verb=ListRecords&metadataPrefix=oai_dc</requestURL>
  + <record>
  + <record>
  - <record>
    - <header>
        <identifier>oai:VTETD:etd-05012000-14030054</identifier>
        <datestamp>2000-05-01</datestamp>
      </header>
    - <metadata>
      - <dc xmlns="http://purl.org/dc/elements/1.1/"
          xsi:schemaLocation="http://purl.org/dc/elements/1.1/
          http://www.openarchives.org/OAI/dc.xsd">
          <title>An Examination of Race and Recurrent Substance Problems in the
            United States</title>
          <creator>Bell, Tannisha D.</creator>
          <subject>Sociology</subject>
          <description>Several studies show that African-Americans are less likely
            than whites to use alcohol or drugs. However, if African-Americans use
            drugs then they are more likely to become heavy and persistent users.
            African-Americans are also more likely to have a current substance
            abuse disorder. There is not much in the literature to explain this
```

4.13. Metadata Multiplicity


```
  <?xml version="1.0" encoding="UTF-8" ?>
- <GetRecord xmlns="http://www.openarchives.org/OAI/1.0/OAI_GetRecord"
    xmlns:xsi="http://www.w3.org/2000/10/XMLSchema-instance"
    xsi:schemaLocation="http://www.openarchives.org/OAI/1.0/OAI_GetRecord
    http://www.openarchives.org/OAI/1.0/OAI_GetRecord.xsd">
    <responseDate>2001-06-14T15:20:41-05:00</responseDate>
    <requestURL>http://oai.dlib.vt.edu:80/~hussein/cgi-bin/NDLTD/VTETD.pl?
      verb=GetRecord&identifier=oai:VTETD:etd-
      3123162539751141&metadataPrefix=oai_rfc1807</requestURL>
  - <record>
    - <header>
        <identifier>oai:VTETD:etd-3123162539751141</identifier>
        <datestamp>1997-04-22</datestamp>
      </header>
    - <metadata>
      - <rfc1807 xmlns="http://info.internet.isi.edu:80/in-
          notes/rfc/files/rfc1807.txt"
          xsi:schemaLocation="http://info.internet.isi.edu:80/in-
          notes/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>
          <type>Thesis/Dissertation</type>
```

4.14. Date Ranges


```
<?xml version="1.0" encoding="UTF-8" ?>
- <ListIdentifiers xmlns="http://www.openarchives.org/OAI/1.0/OAI_ListIdentifiers"
    xmlns:xsi="http://www.w3.org/2000/10/XMLSchema-instance"
    xsi:schemaLocation="http://www.openarchives.org/OAI/1.0/OAI_ListIdentifiers
    http://www.openarchives.org/OAI/1.0/OAI_ListIdentifiers.xsd">
    <responseDate>2001-06-14T15:21:37-05:00</responseDate>
    <requestURL>http://oai.dlib.vt.edu:80/~hussein/cgi-bin/NDLTD/VTETD.pl?
      verb=ListIdentifiers&from=2000-11-24&until=2000-12-01</requestURL>
    <identifier>oai:VTETD:etd-11212000-155513</identifier>
    <identifier>oai:VTETD:etd-11242000-130040</identifier>
    <identifier>oai:VTETD:etd-11272000-115149</identifier>
    <identifier>oai:VTETD:etd-11162000-19160016</identifier>
    <identifier>oai:VTETD:etd-11222000-095443</identifier>
    <identifier>oai:VTETD:etd-11142000-16540027</identifier>
    <identifier>oai:VTETD:etd-11282000-110022</identifier>
    <identifier>oai:VTETD:etd-11152000-13100048</identifier>
    <identifier>oai:VTETD:etd-11272000-114011</identifier>
    <identifier>oai:VTETD:etd-11182000-10350010</identifier>
    <identifier>oai:VTETD:etd-11272000-214847</identifier>
    <identifier>oai:VTETD:etd-11182000-16030010</identifier>
  </ListIdentifiers>
```

4.15. Resumption Token


```
  <identifier>oai:VTETD:etd-254122839711171</identifier>
  <identifier>oai:VTETD:etd-4524171049761291</identifier>
  <identifier>oai:VTETD:etd-3156151139751001</identifier>
  <identifier>oai:VTETD:etd-424817300974290</identifier>
  <identifier>oai:VTETD:etd-13514459731541</identifier>
  <identifier>oai:VTETD:etd-2047101569611961</identifier>
  <identifier>oai:VTETD:etd-5414132139711101</identifier>
  <identifier>oai:VTETD:etd-3132141279612241</identifier>
  <identifier>oai:VTETD:etd-3123162539751141</identifier>
  <identifier>oai:VTETD:etd-556181169641921</identifier>
  <identifier>oai:VTETD:etd-342482139711101</identifier>
  <identifier>oai:VTETD:etd-1913943975930</identifier>
  <identifier>oai:VTETD:etd-402515359721531</identifier>
  <identifier>oai:VTETD:etd-2025212339731121</identifier>
  <identifier>oai:VTETD:etd-3331171059721601</identifier>
  <identifier>oai:VTETD:etd-18409759651581</identifier>
  <identifier>oai:VTETD:etd-34521672975650</identifier>
  <identifier>oai:VTETD:etd-120142139711101</identifier>
  <identifier>oai:VTETD:etd-4019122049721391</identifier>
  <identifier>oai:VTETD:etd-487142639761151</identifier>
  <resumptionToken>!!!!100</resumptionToken>
</ListIdentifiers>
```

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
- Constructing resumption tokens

5.1. Tools Required

- Code templates if available (currently available for many languages)
- Basic programming environment
- XML generators (for non-trivial encoding)
- Database access libraries/drivers (e.g. DBI, ODBC, JDBC)

5.2. 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
   Error (400, ‘Unknown verb’);
```

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

5.4. Metadata Generation

- Approaches
  - Map from source to each metadata format
  - Use crosswalks to generate additional formats

source		dc		rfc1807
name	=	title	=	title
author	=	creator	=	author

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

- 400
  - Illegal verb value
  - Illegal parameter values, combinations
- 404
  - Archive errors – cannot return a legal response
- In general, everything else returns a legal but empty response !

5.8. 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

5.9. Using 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:
  - vtetd123456789

6. Common Problems

- No unique identifiers !
- No datestamps !
- Incomplete information in database
- New metadata format
- XML responses not validating
- Do I return an HTTP error or not ?

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
- 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 hard-coded 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
- 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
- Register the new format with the OAI **

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

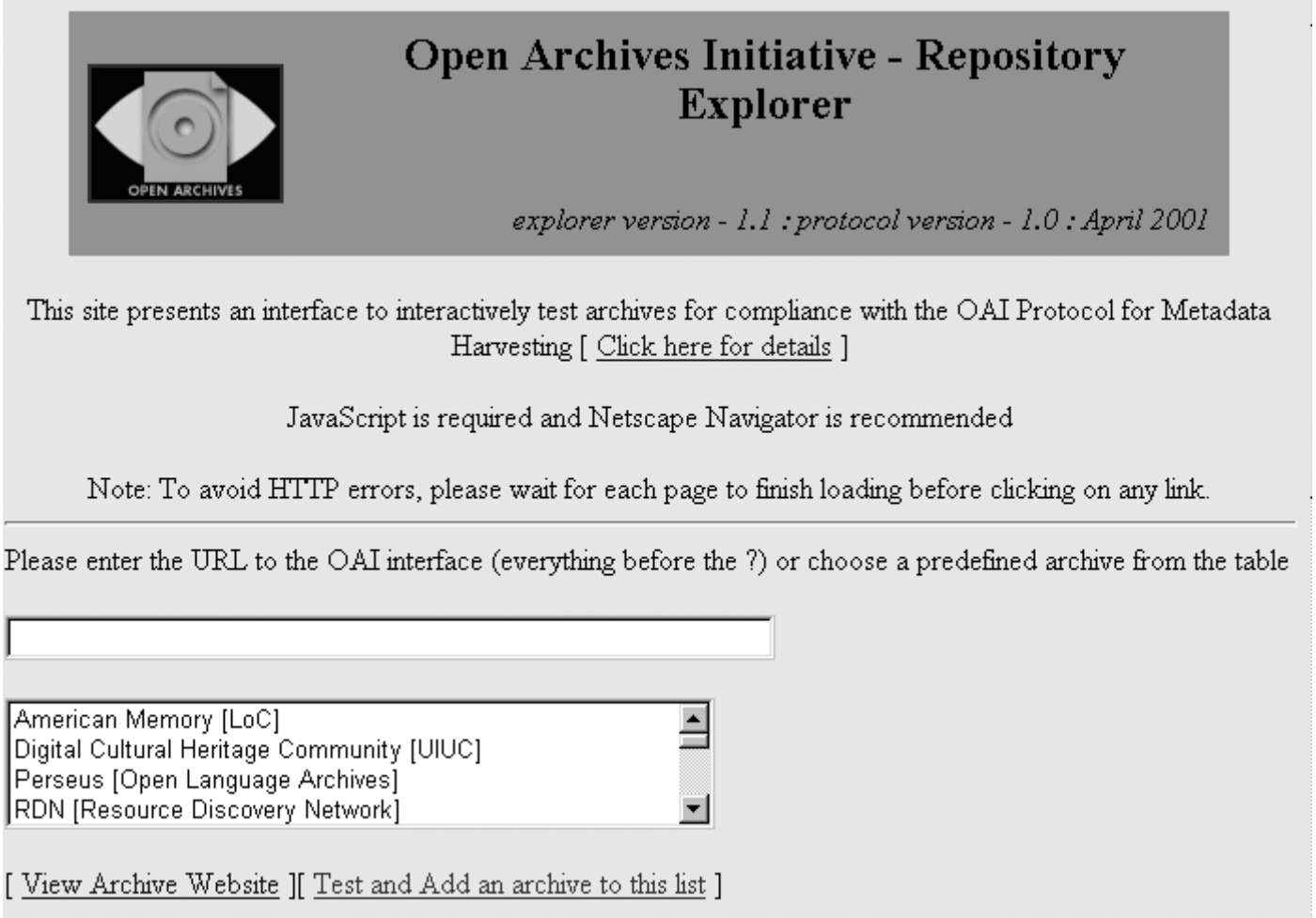
6.6. HTTP Error ?

- Unless the archive is temporarily non-functional or the parameters are intrinsically wrong, do not return an HTTP error
  - If no metadata exists in a particular format, return a record with no metadata field
  - If the set name does not exist, return an empty list
  - If the identifier does not exist, return an empty response

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

7.1. RE Interactive Browsing



7.2. RE Parameter Testing

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

Language	Display	Schema Validation
English	Parsed Raw XML Both	None Local mirror of schemata Online schemata

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

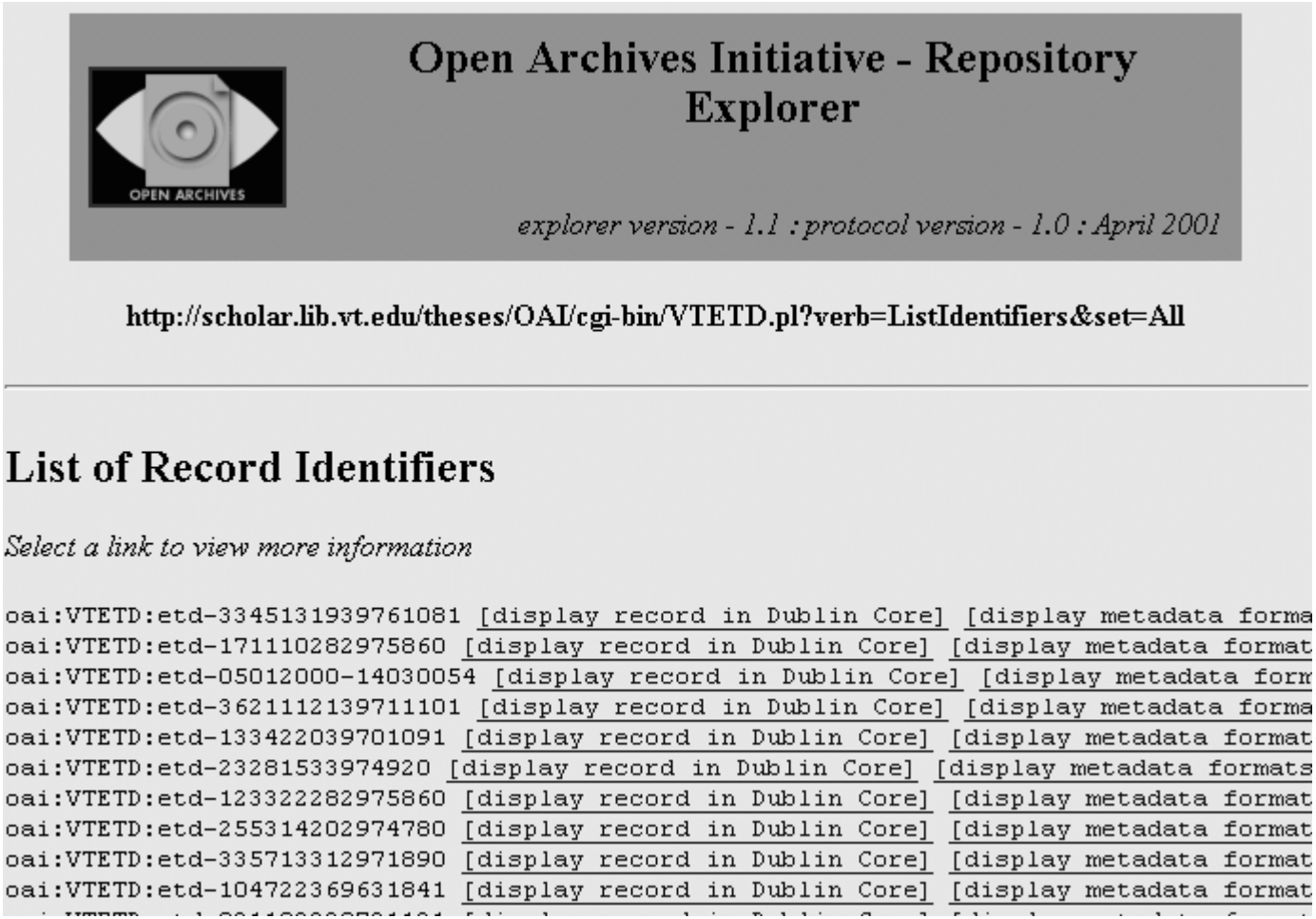
7.3. RE Browsing

Archive Self-Description

Repository Name	Virginia Tech Electronic Thesis and Dissertation Collection
Base URL	http://scholar.lib.vt.edu:80/theses/OAI/cgi-bin/VTETD.pl
Protocol Version	1.0
Admin Email	mailto:webmaster@scholar.lib.vt.edu
Other Information	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 statemen

7.4. RE Browsing

7.5. RE Browsing



Open Archives Initiative - Repository Explorer

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

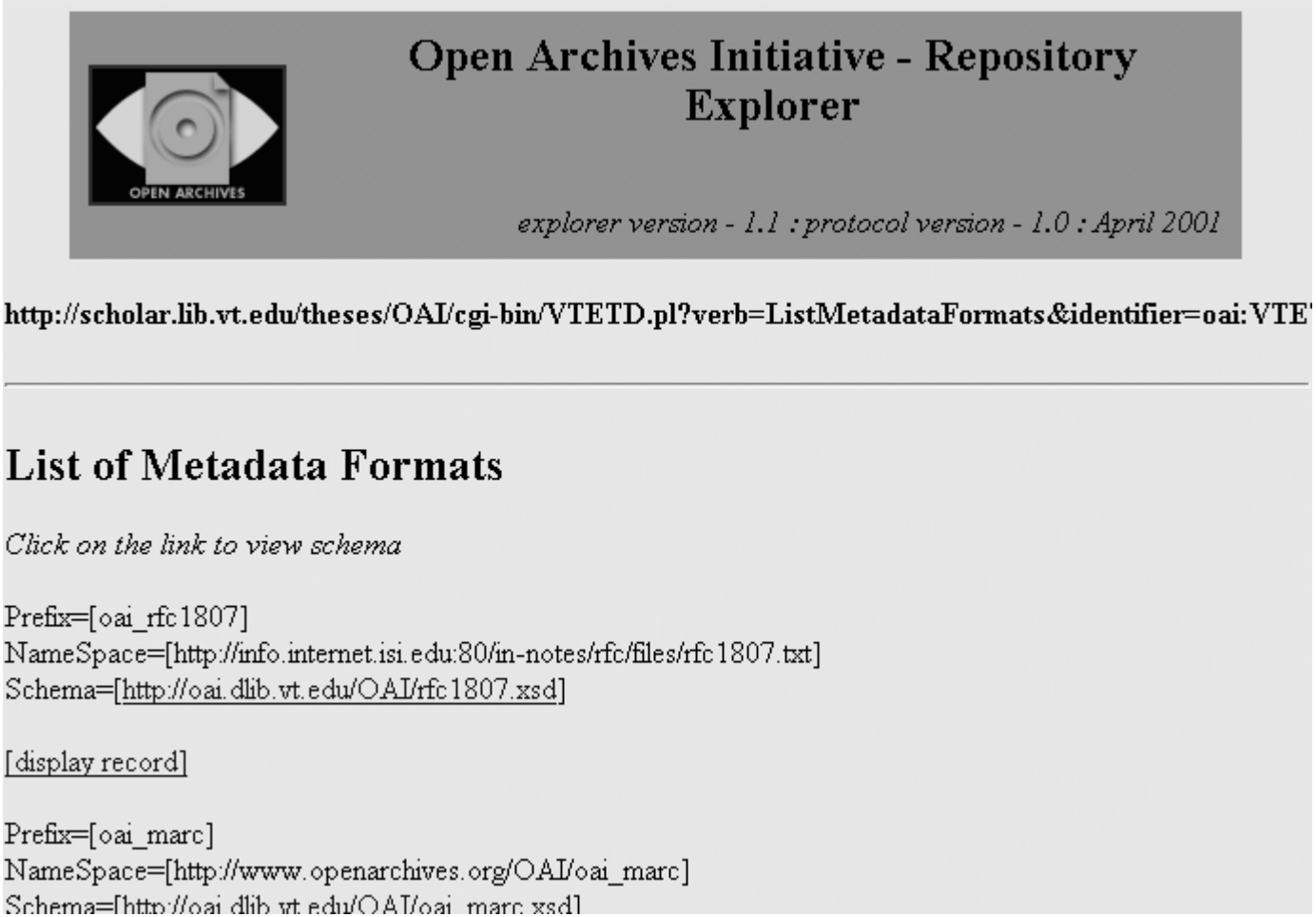
http://scholar.lib.vt.edu/theses/OAI/cgi-bin/VTETD.pl?verb=ListIdentifiers&set=All

List of Record Identifiers

Select a link to view more information

```
oai:VTETD:etd-3345131939761081 [display record in Dublin Core] [display metadata forma
oai:VTETD:etd-171110282975860 [display record in Dublin Core] [display metadata format
oai:VTETD:etd-05012000-14030054 [display record in Dublin Core] [display metadata form
oai:VTETD:etd-3621112139711101 [display record in Dublin Core] [display metadata forma
oai:VTETD:etd-133422039701091 [display record in Dublin Core] [display metadata format
oai:VTETD:etd-23281533974920 [display record in Dublin Core] [display metadata formats
oai:VTETD:etd-123322282975860 [display record in Dublin Core] [display metadata format
oai:VTETD:etd-255314202974780 [display record in Dublin Core] [display metadata format
oai:VTETD:etd-335713312971890 [display record in Dublin Core] [display metadata format
oai:VTETD:etd-104722369631841 [display record in Dublin Core] [display metadata format
```

7.6. RE Browsing



Open Archives Initiative - Repository Explorer

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

http://scholar.lib.vt.edu/theses/OAI/cgi-bin/VTETD.pl?verb=ListMetadataFormats&identifier=oai:VTE

List of Metadata Formats

Click on the link to view schema

Prefix=[oai_rfc1807]
NameSpace=[http://info.internet.isi.edu:80/in-notes/rfc/files/rfc1807.txt]
Schema=[http://oai.dlib.vt.edu/OAI/rfc1807.xsd]

[display record]

Prefix=[oai_marc]
NameSpace=[http://www.openarchives.org/OAI/oai_marc]
Schema=[http://oai.dlib.vt.edu/OAI/oai_marc.xsd]

7.7. RE Browsing

http://scholar.lib.vt.edu/theses/OAI/cgi-bin/VTETD.pl?verb=GetRecord&metadataPrefix=oai_etdms&i

List of Fields

```
header:
  identifier : oai:VTETD:etd-3345131939761081
  datestamp : 1997-03-31

metadata:
   thesis:
      title: Conceptual Development and Empirical Testing of an Outdoor Recreation Exp
      creator: Walker, Gordon James
      publisher: Virginia Polytechnic Institute and State University
      subject: outdoor recreation
      subject: recreation experience preference scales
      subject: recreation experience matrix
      subject: recreation opportunity spectrum
      description: This dissertation examines four issues, including: (a) whether outd
the Recreation Experience Preference (REP) scales exist; (b) whether these experiences
using a framework called the Recreation Experience Matrix (REM); (c) how well the Recr
Spectrum (ROS) variables of activity, setting, and expertise explain the types of expe
recreationists receive; and (d) how well two new variables--primary mode and mode depe
types of experiences outdoor recreationists receive. In order to address these issues,
was distributed at Mount Rogers National Recreation Area in Virginia during October an
A total of 410 people completed this questionnaire. Of these, 336 provided useable add
```

7.8. RE Multiple views of data

l-3345131939761081

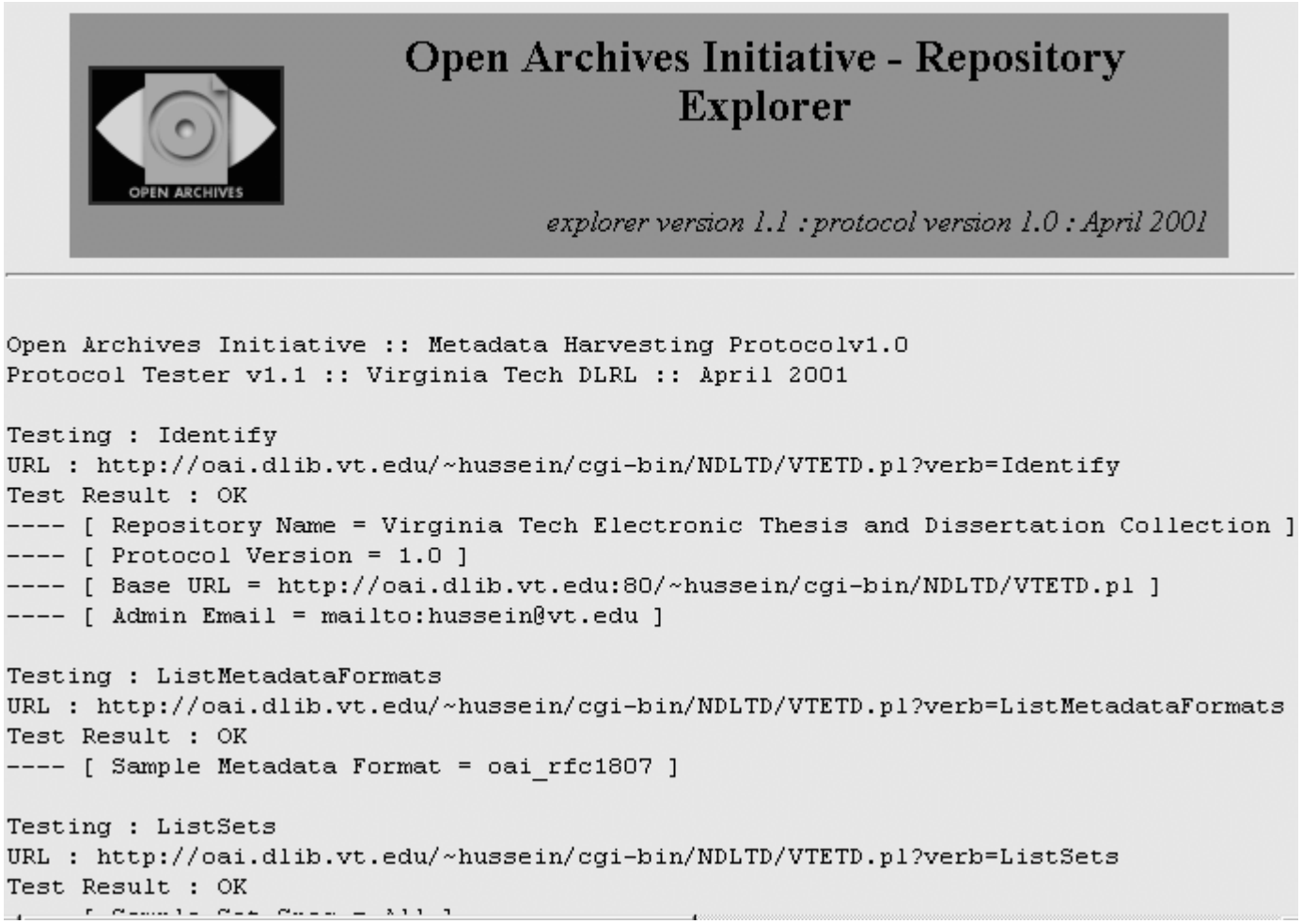
Raw XML Output

31939761081

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

<ListSets xmlns="http://www.openarchives.org/OAI/1.0/OAI_ListSets" xmlns

<responseDate>2001-06-14T20:32:33-05:00</responseDate>
<requestURL>http://scholar.lib.vt.edu:80/theses/OAI/cgi-bin/VTETD.pl?ver

<set>
  <setSpec>All</setSpec>
  <setName>All theses and dissertations</setName>
</set>

</ListSets>
```

7.9. RE Multilingual Support

7.10. RE Automatic Test Suite

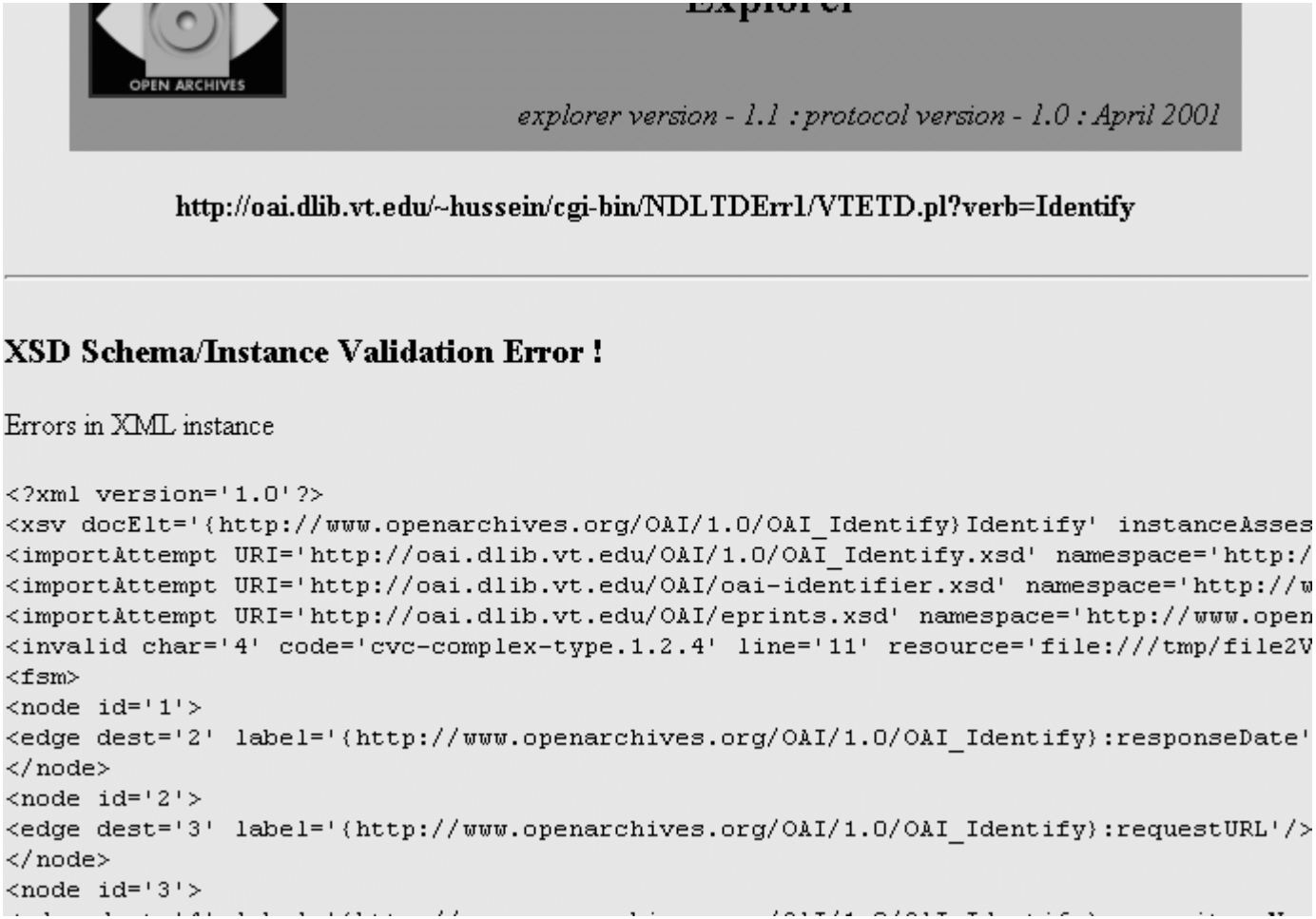


7.11. RE Error in Response

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	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 statemen

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/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'>
```

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 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

7.14. 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
celebration	A Celebration of Women Writers
anlc	Alaska Native Language Center
arXiv	arXiv
CDLCIAS	California International and Area Studies Digital Repository

7.15. XSV Schema Validator

W3C® XML LTG

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 **ALPHA TEST** of a service for a **work-in-progress specification**. This version is for schema documents with the namespace URI http://www.w3.org/2000/10/XMLSchema and is being actively developed: see XSV for XML Schema 200004007 version 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):

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 you work with the timezone of the source)
  - New items may be added for the current day
  - Timezones create up to a day of lag if you ignore them
- Each time a record is encountered, erase previous instances

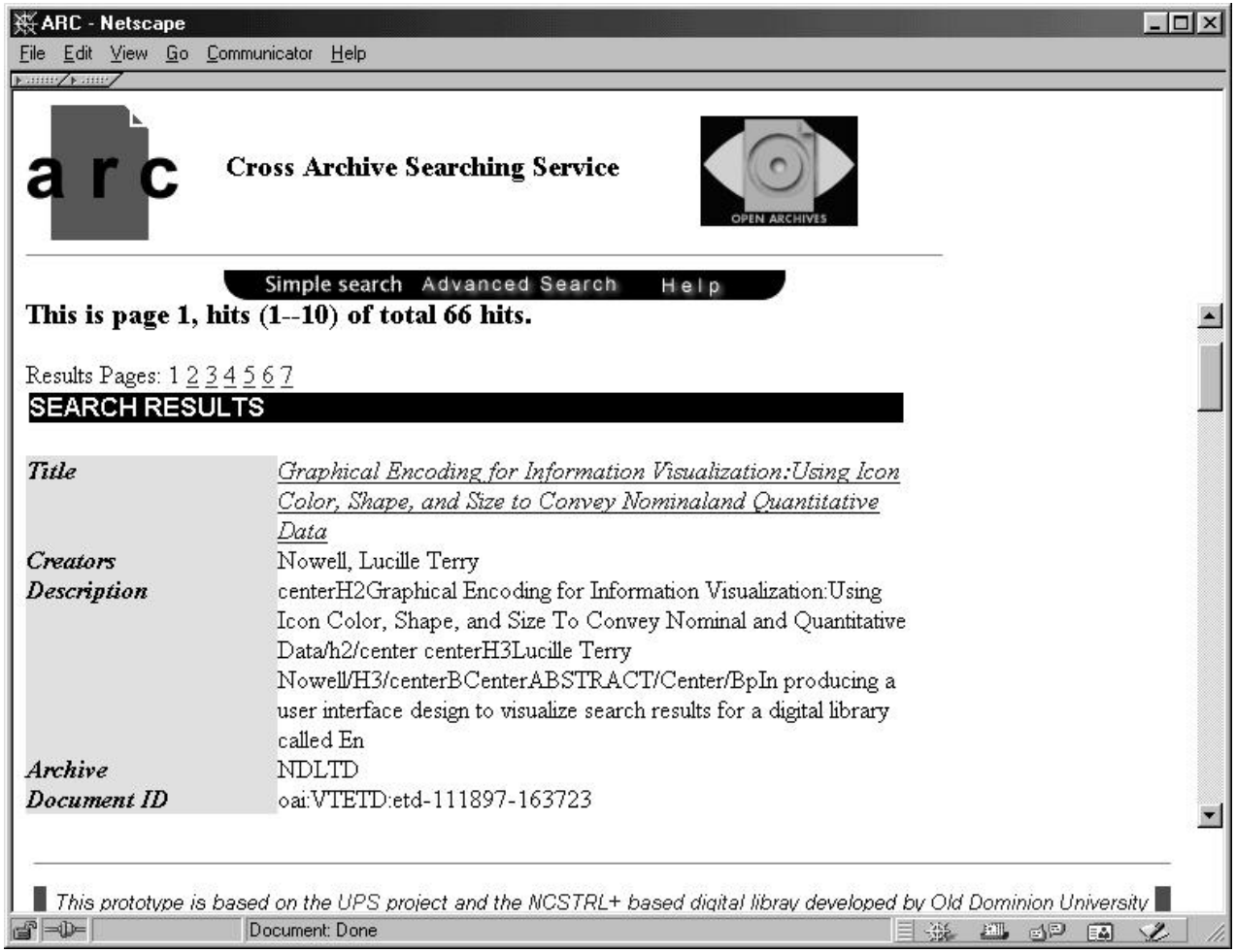
8.3. Intermediate Systems

- 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!

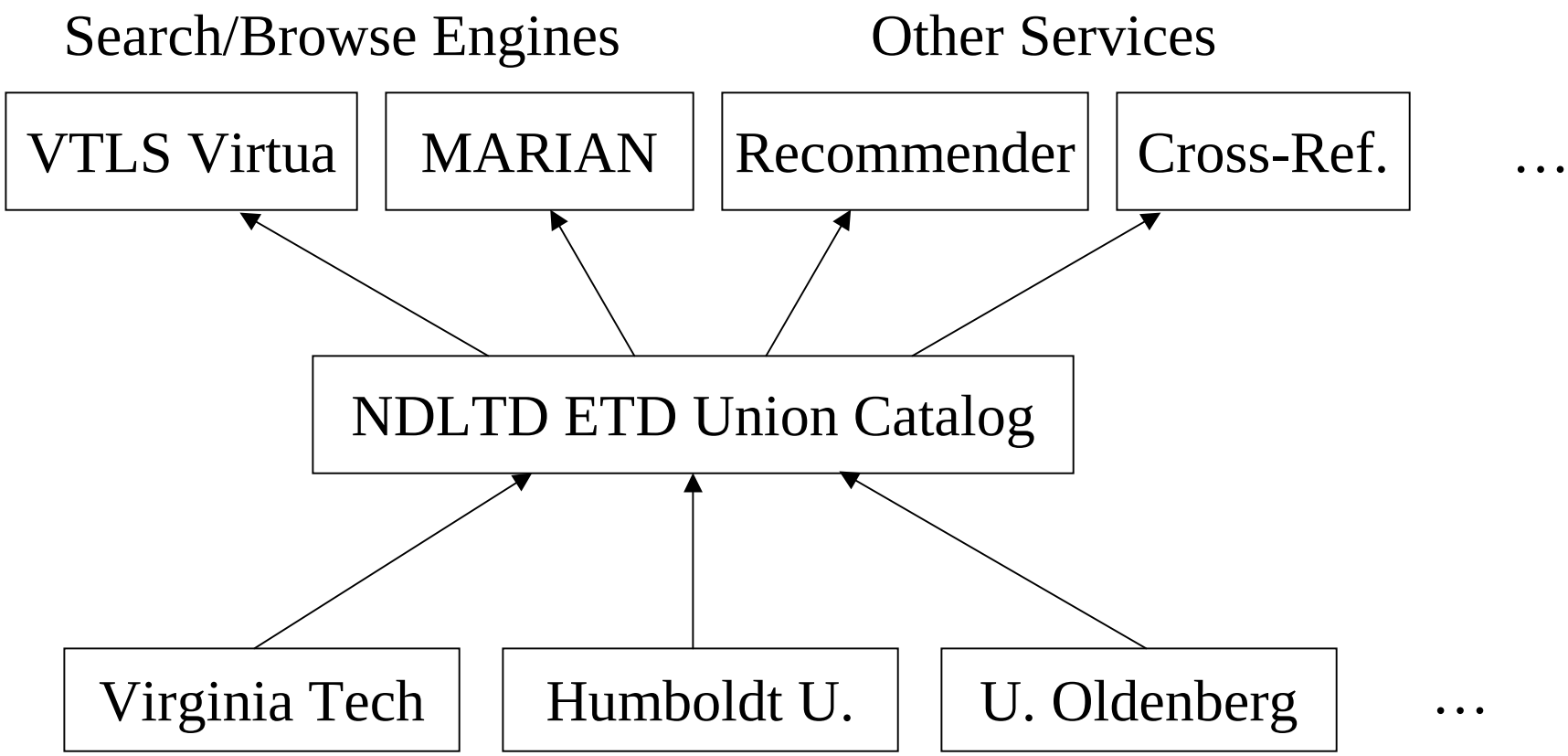
8.4. Tools

- 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
- 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

- 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

- 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

- Use 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

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?

10.2. Future of Protocol

- Version 1.1
  - Soon – minor upgrade to cater for updates to schema language by W3C

- Evaluation
  - Within a year – does this protocol make sense ?

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
  - http://www.openarchives.org
- OAI Metadata Harvesting Protocol
  - 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
  - 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

That’s All Folks !