ALA 2002
LITA Open Source Software

Open Archives Initiative

Hussein Suleman <hussein@vt.edu>
AmericanSouth.org

14 June 2002

Outline

1. Introduction to OAI
2. Definitions and Concepts
3. Protocol for Metadata Harvesting
4. OAI and ODL Open Source Software
5. Installation of XML-File software
6. Testing of XML-File
7. Installation of harvester
8. Installation of IRDB
9. User interface for IRDB
10. Wrap-up and discussion

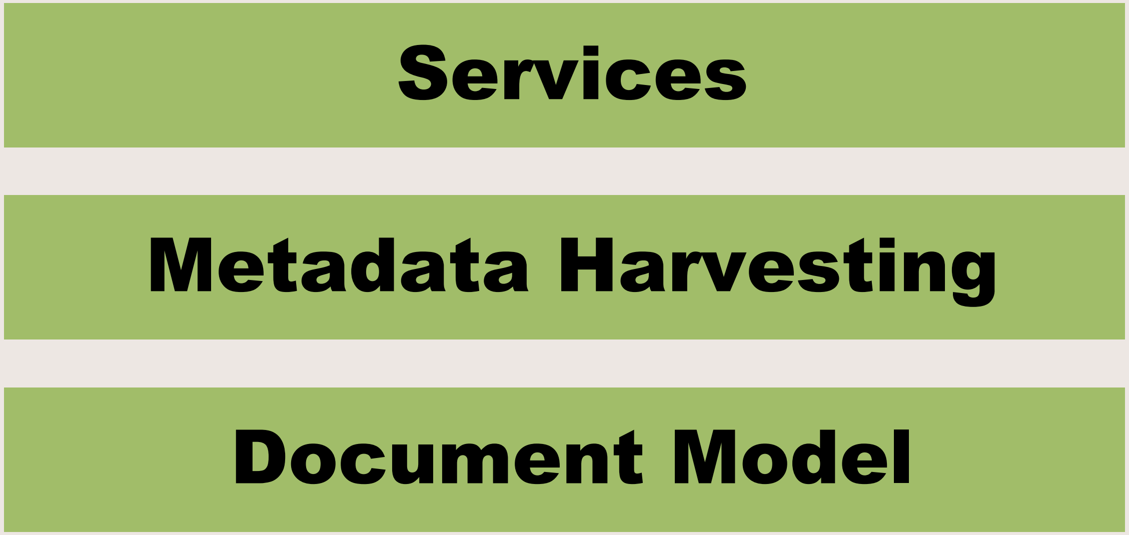
1. Introduction to OAI

- What is the Open Archives Initiative ?
  - Group of people and organizations dedicated to solving problems of digital library interoperability by developing simple protocols.
- Major Accomplishment:
  - Protocol for Metadata Harvesting (OAI-PMH)

1.1. What is the OAI-PMH ?

- What is the Protocol for Metadata Harvesting?
  - Protocol to transfer metadata from one archive to another
    - Any metadata
    - In a continuous stream
    - As simply as possible

1.2. General System Strategy



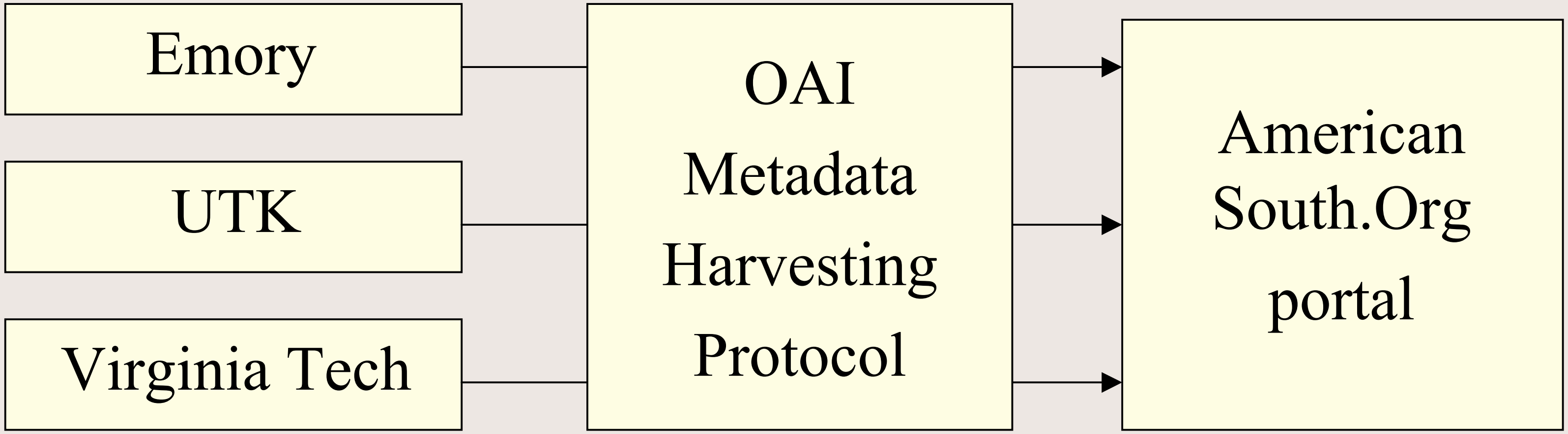
Services

Metadata Harvesting

Document Model

1.3. Case Study: AmericanSouth

- Digital library of resources related to Southern history and culture
- Multiple independent university-based collections of electronic documents



Emory

UTK

Virginia Tech

OAI Metadata Harvesting Protocol

American South.Org portal

1.4. Versions of OAI-PMH

- v1.0 January 2001
- v1.1 July 2001
  - Minor revision from v1.0
  - These notes are based on version 1.1 !
- v2.0 June 2002 (expected)
  - Mostly syntactical changes

2. Definitions / Concepts

- Basic Principles
  - What is an Open Archive?
  - Harvesting vs. Federation
  - Data and Service Providers
- Underlying Technology
  - HTTP and XML
- Protocol Policies
  - 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. 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)

2.4. 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

2.5. 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

2.6. Sample OAI Record

```
<record>
   <header>
      <identifier>oai:sigir:ws3</identifier>
      <datestamp>2001-08-13</datestamp>
   </header>
   <metadata>
      <dc>
         <title>OAI Workshop at SIGIR</title>
         <creator>Hussein Suleman</creator>
         <language>English</language>
      </dc>
   </metadata>
   <about>
      <metadataID>oai:sigir:ws3md</metadataID>
   </about>
</record>
```

2.7. 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)

2.8. 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

2.9. 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

2.10. 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

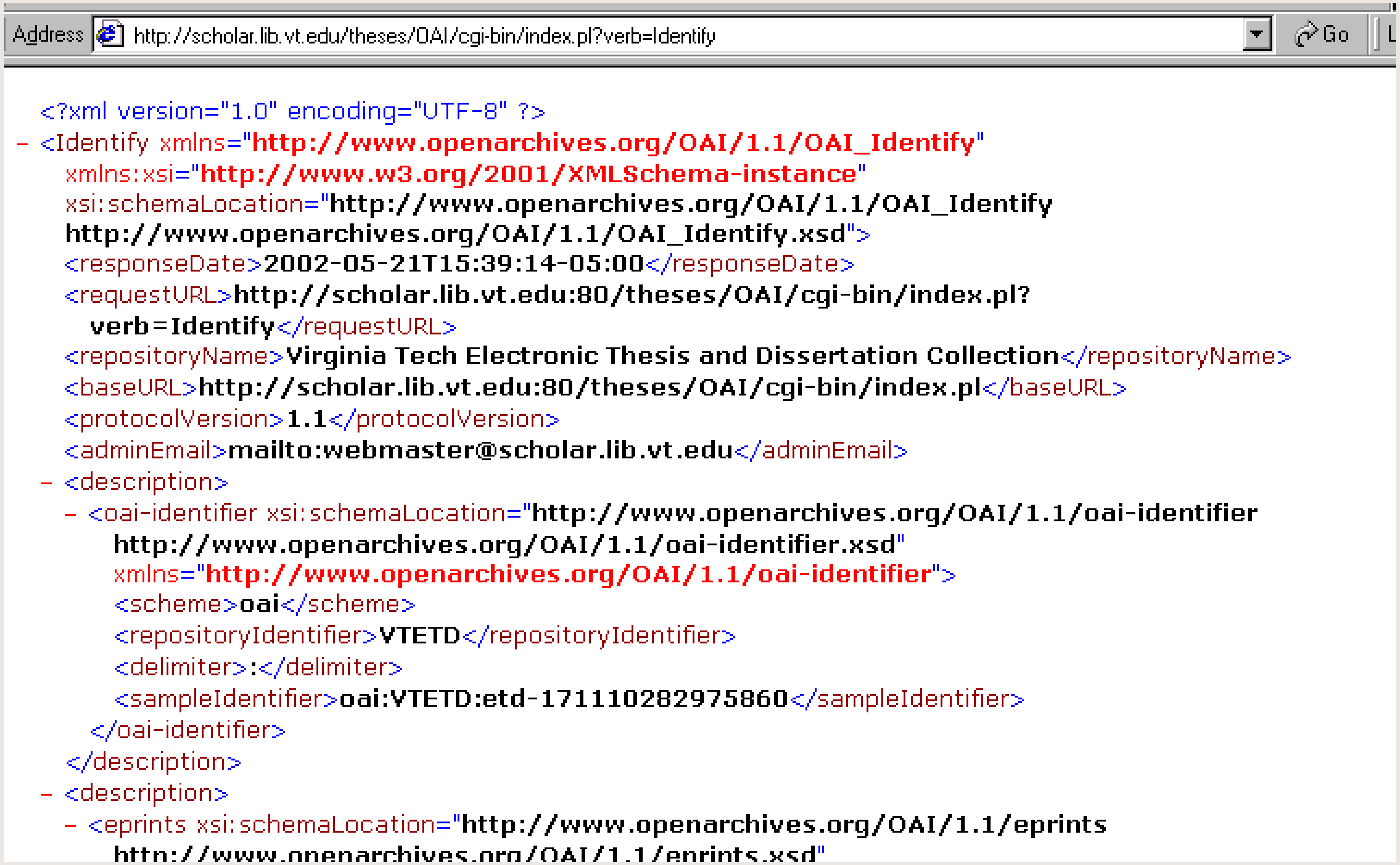
3. Protocol for Metadata Harvesting

- Service Requests
  - Identify
  - ListMetadataFormats
  - ListSets
  - GetRecord
  - ListIdentifiers
  - ListRecords
- Metadata Multiplicity
- Date Ranges
- Resumption Tokens

3.1. Identify

- Purpose
  - Return general information about the archive and its policies
- Parameters
  - None
- Sample URL
  - http://www.anarchive.org/cgi-bin/OAI?verb=Identify

3.2. Identify - Response



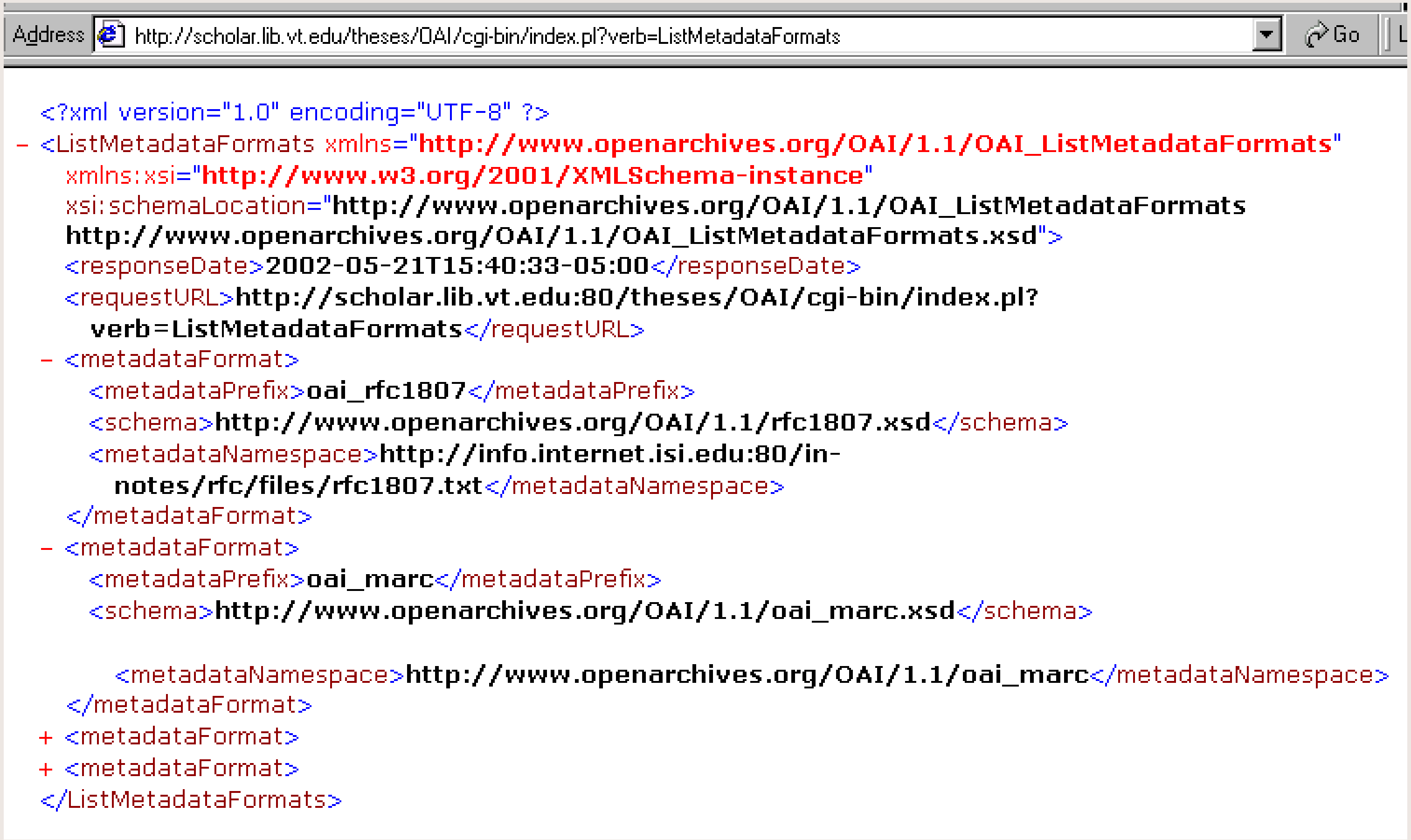
http://scholar.lib.vt.edu/theses/OAI/cgi-bin/index.pl?verb=Identify

```
  <?xml version="1.0" encoding="UTF-8" ?>
- <Identify xmlns="http://www.openarchives.org/OAI/1.1/OAI_Identify"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.openarchives.org/OAI/1.1/OAI_Identify
    http://www.openarchives.org/OAI/1.1/OAI_Identify.xsd">
    <responseDate>2002-05-21T15:39:14-05:00</responseDate>
    <requestURL>http://scholar.lib.vt.edu:80/theses/OAI/cgi-bin/index.pl?
      verb=Identify</requestURL>
    <repositoryName>Virginia Tech Electronic Thesis and Dissertation Collection</repositoryName>
    <baseURL>http://scholar.lib.vt.edu:80/theses/OAI/cgi-bin/index.pl</baseURL>
    <protocolVersion>1.1</protocolVersion>
    <adminEmail>mailto:webmaster@scholar.lib.vt.edu</adminEmail>
  - <description>
    - <oai-identifier xsi:schemaLocation="http://www.openarchives.org/OAI/1.1/oai-identifier
        http://www.openarchives.org/OAI/1.1/oai-identifier.xsd"
        xmlns="http://www.openarchives.org/OAI/1.1/oai-identifier">
        <scheme>oai</scheme>
        <repositoryIdentifier>VTETD</repositoryIdentifier>
        <delimiter>:</delimiter>
        <sampleIdentifier>oai:VTETD:etd-171110282975860</sampleIdentifier>
      </oai-identifier>
    </description>
  - <description>
    - <eprints xsi:schemaLocation="http://www.openarchives.org/OAI/1.1/eprints
        http://www.openarchives.org/OAI/1.1/eprints.xsd"
```

3.3. 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

3.4. ListMetadataFormats - Response



http://scholar.lib.vt.edu/theses/OAI/cgi-bin/index.pl?verb=ListMetadataFormats

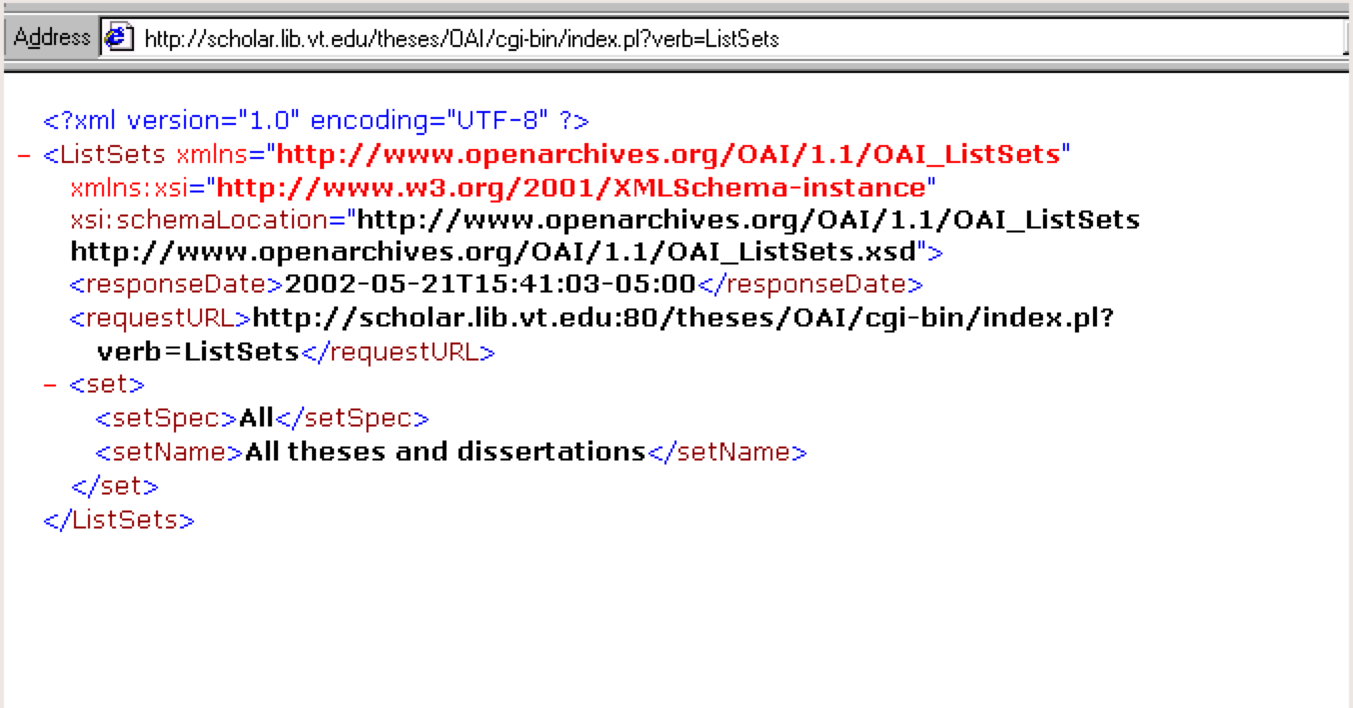
```
 <?xml version="1.0" encoding="UTF-8" ?>
- <ListMetadataFormats xmlns="http://www.openarchives.org/OAI/1.1/OAI_ListMetadataFormats"
   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
   xsi:schemaLocation="http://www.openarchives.org/OAI/1.1/OAI_ListMetadataFormats
   http://www.openarchives.org/OAI/1.1/OAI_ListMetadataFormats.xsd">
   <responseDate>2002-05-21T15:40:33-05:00</responseDate>
   <requestURL>http://scholar.lib.vt.edu:80/theses/OAI/cgi-bin/index.pl?
     verb=ListMetadataFormats</requestURL>
 - <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>
 - <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>
 + <metadataFormat>
 </ListMetadataFormats>
```

3.5. 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

3.6. ListSets – Response



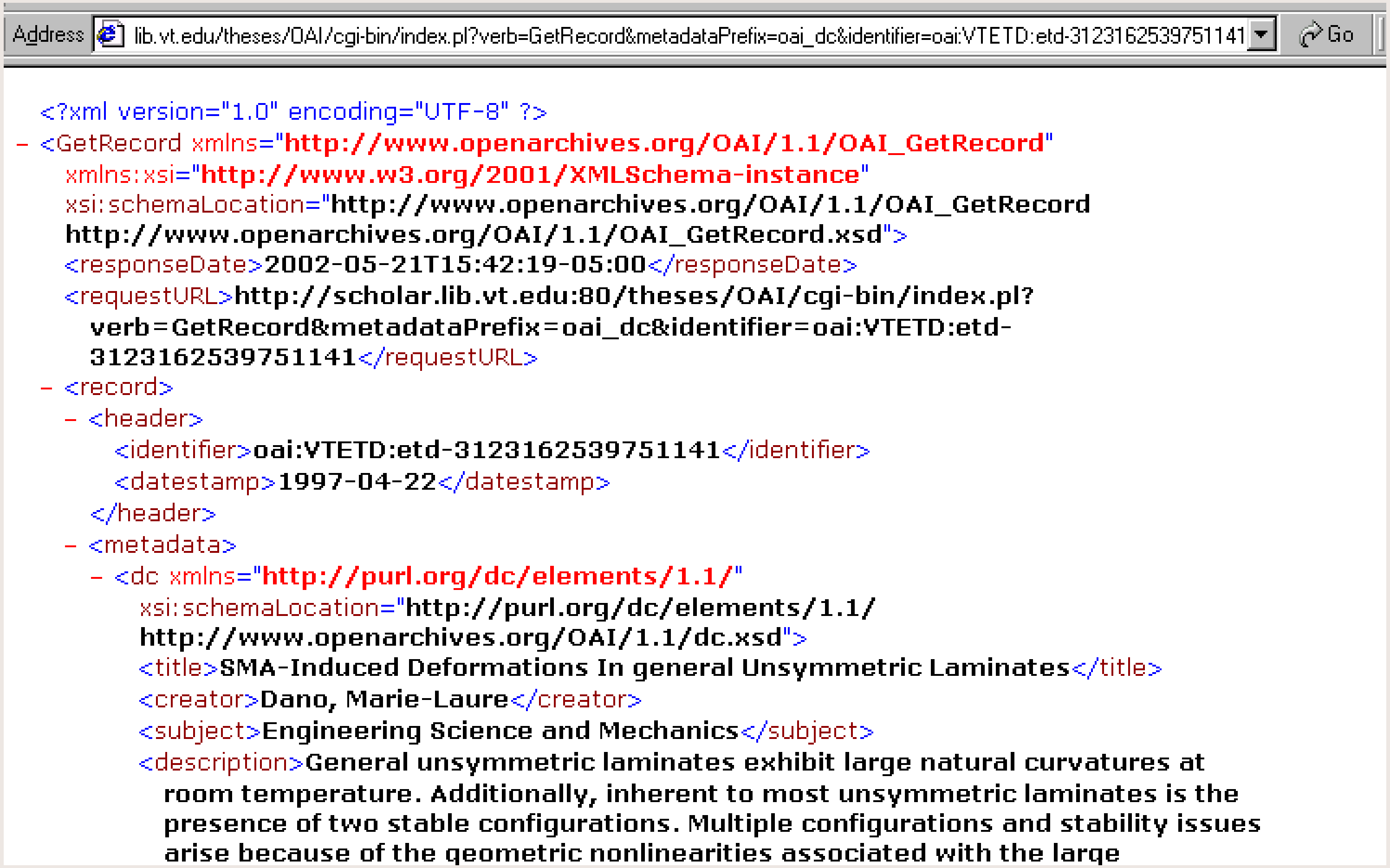
Address http://scholar.lib.vt.edu/theses/OAI/cgi-bin/index.pl?verb=ListSets

```
  <?xml version="1.0" encoding="UTF-8" ?>
- <ListSets xmlns="http://www.openarchives.org/OAI/1.1/OAI_ListSets"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.openarchives.org/OAI/1.1/OAI_ListSets
    http://www.openarchives.org/OAI/1.1/OAI_ListSets.xsd">
    <responseDate>2002-05-21T15:41:03-05:00</responseDate>
    <requestURL>http://scholar.lib.vt.edu:80/theses/OAI/cgi-bin/index.pl?
      verb=ListSets</requestURL>
  - <set>
      <setSpec>All</setSpec>
      <setName>All theses and dissertations</setName>
    </set>
  </ListSets>
```

3.7. 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

3.8. GetRecord - Response



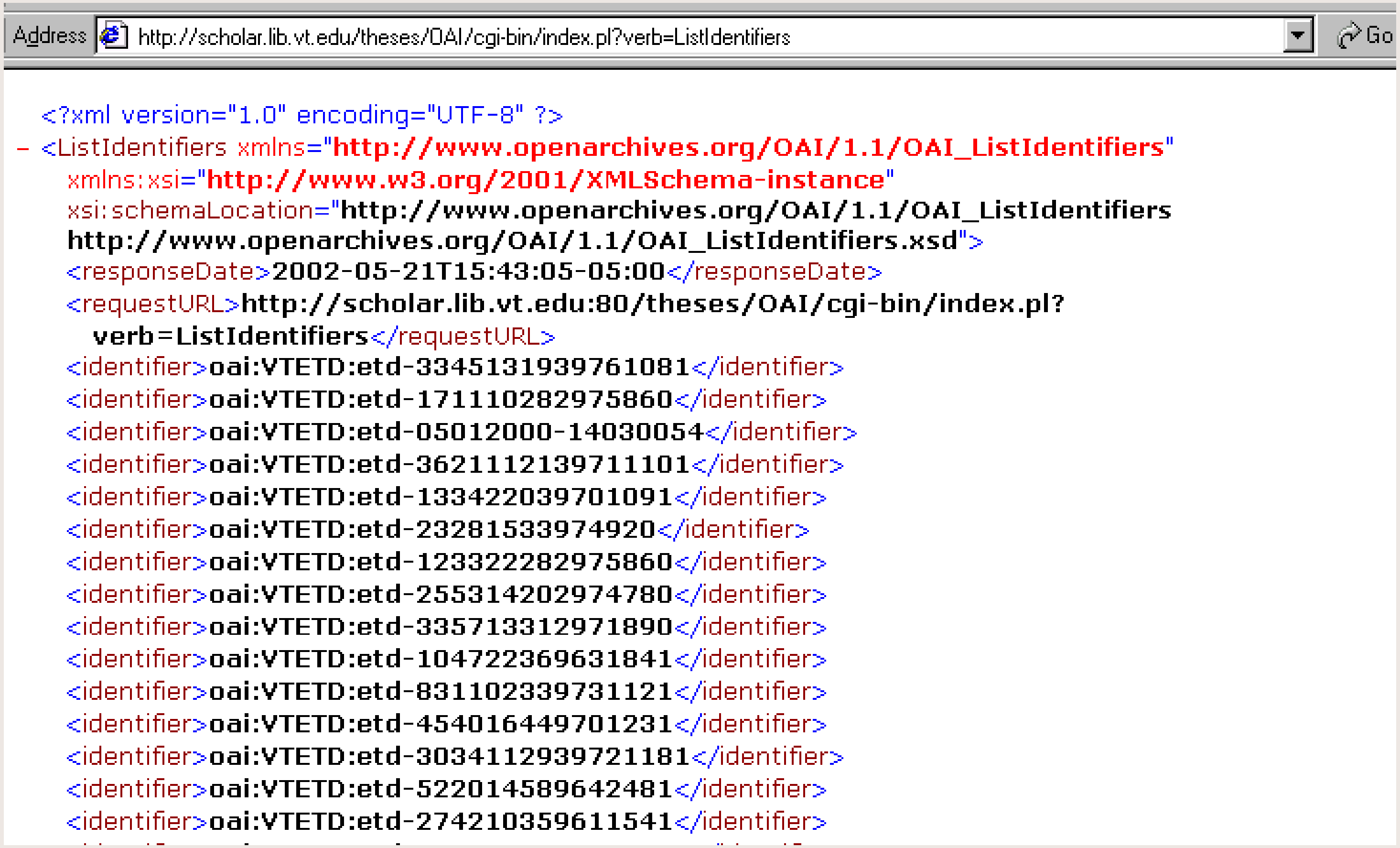
lib.vt.edu/theses/OAI/cgi-bin/index.pl?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:VTETD:etd-3123162539751141

```
  <?xml version="1.0" encoding="UTF-8" ?>
- <GetRecord xmlns="http://www.openarchives.org/OAI/1.1/OAI_GetRecord"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.openarchives.org/OAI/1.1/OAI_GetRecord
    http://www.openarchives.org/OAI/1.1/OAI_GetRecord.xsd">
    <responseDate>2002-05-21T15:42:19-05:00</responseDate>
    <requestURL>http://scholar.lib.vt.edu:80/theses/OAI/cgi-bin/index.pl?
      verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:VTETD:etd-
      3123162539751141</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/1.1/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
```

3.9. 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

3.10. ListIdentifiers - Response



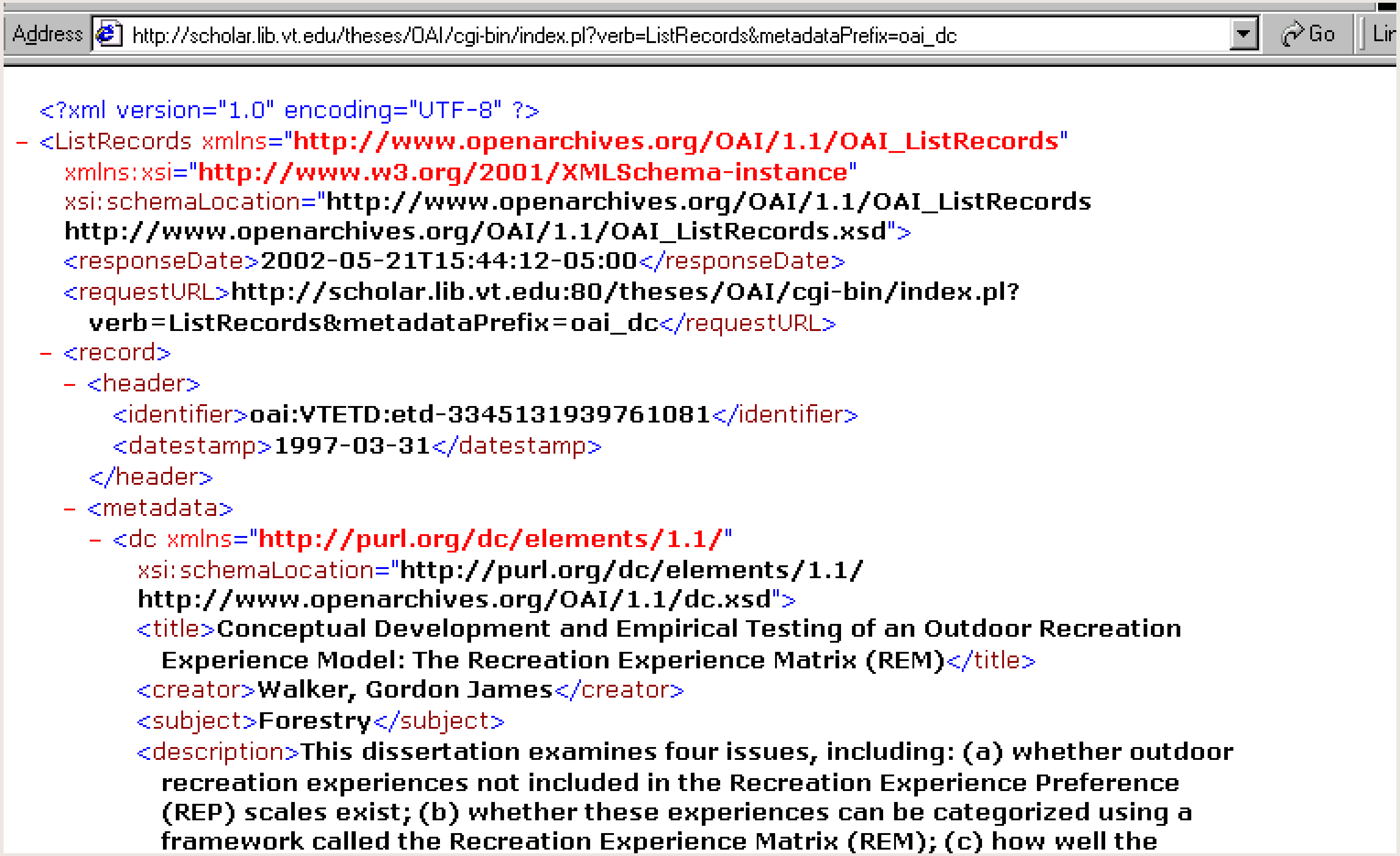
http://scholar.lib.vt.edu/theses/OAI/cgi-bin/index.pl?verb=ListIdentifiers

```
  <?xml version="1.0" encoding="UTF-8" ?>
- <ListIdentifiers xmlns="http://www.openarchives.org/OAI/1.1/OAI_ListIdentifiers"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.openarchives.org/OAI/1.1/OAI_ListIdentifiers
    http://www.openarchives.org/OAI/1.1/OAI_ListIdentifiers.xsd">
    <responseDate>2002-05-21T15:43:05-05:00</responseDate>
    <requestURL>http://scholar.lib.vt.edu:80/theses/OAI/cgi-bin/index.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>
```

3.11. 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

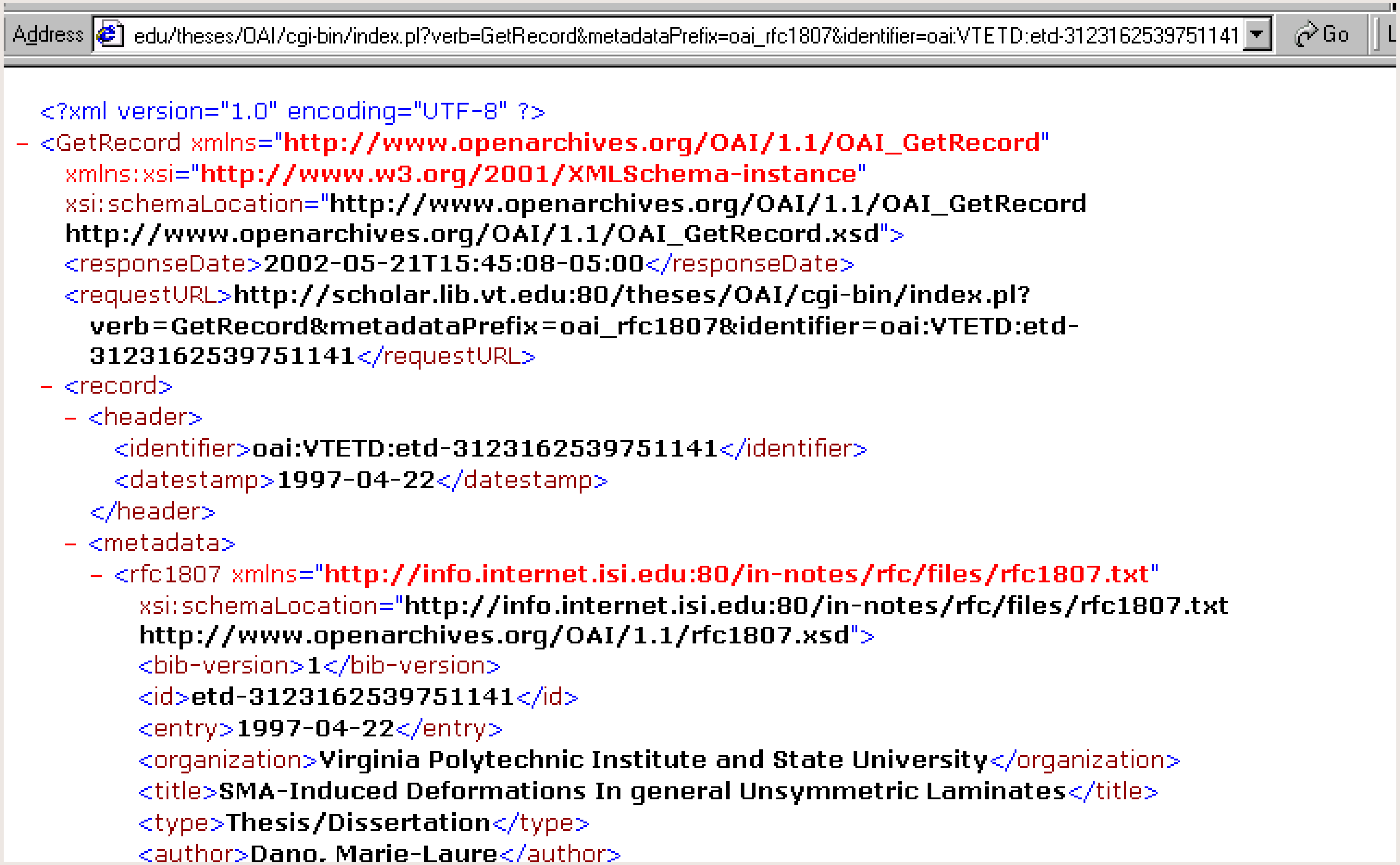
3.12. ListRecords - Response



http://scholar.lib.vt.edu/theses/OAI/cgi-bin/index.pl?verb=ListRecords&metadataPrefix=oai_dc

```
 <?xml version="1.0" encoding="UTF-8" ?>
- <ListRecords xmlns="http://www.openarchives.org/OAI/1.1/OAI_ListRecords"
   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
   xsi:schemaLocation="http://www.openarchives.org/OAI/1.1/OAI_ListRecords
   http://www.openarchives.org/OAI/1.1/OAI_ListRecords.xsd">
   <responseDate>2002-05-21T15:44:12-05:00</responseDate>
   <requestURL>http://scholar.lib.vt.edu:80/theses/OAI/cgi-bin/index.pl?
     verb=ListRecords&metadataPrefix=oai_dc</requestURL>
 - <record>
   - <header>
       <identifier>oai:VTETD:etd-3345131939761081</identifier>
       <datestamp>1997-03-31</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/1.1/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 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
```

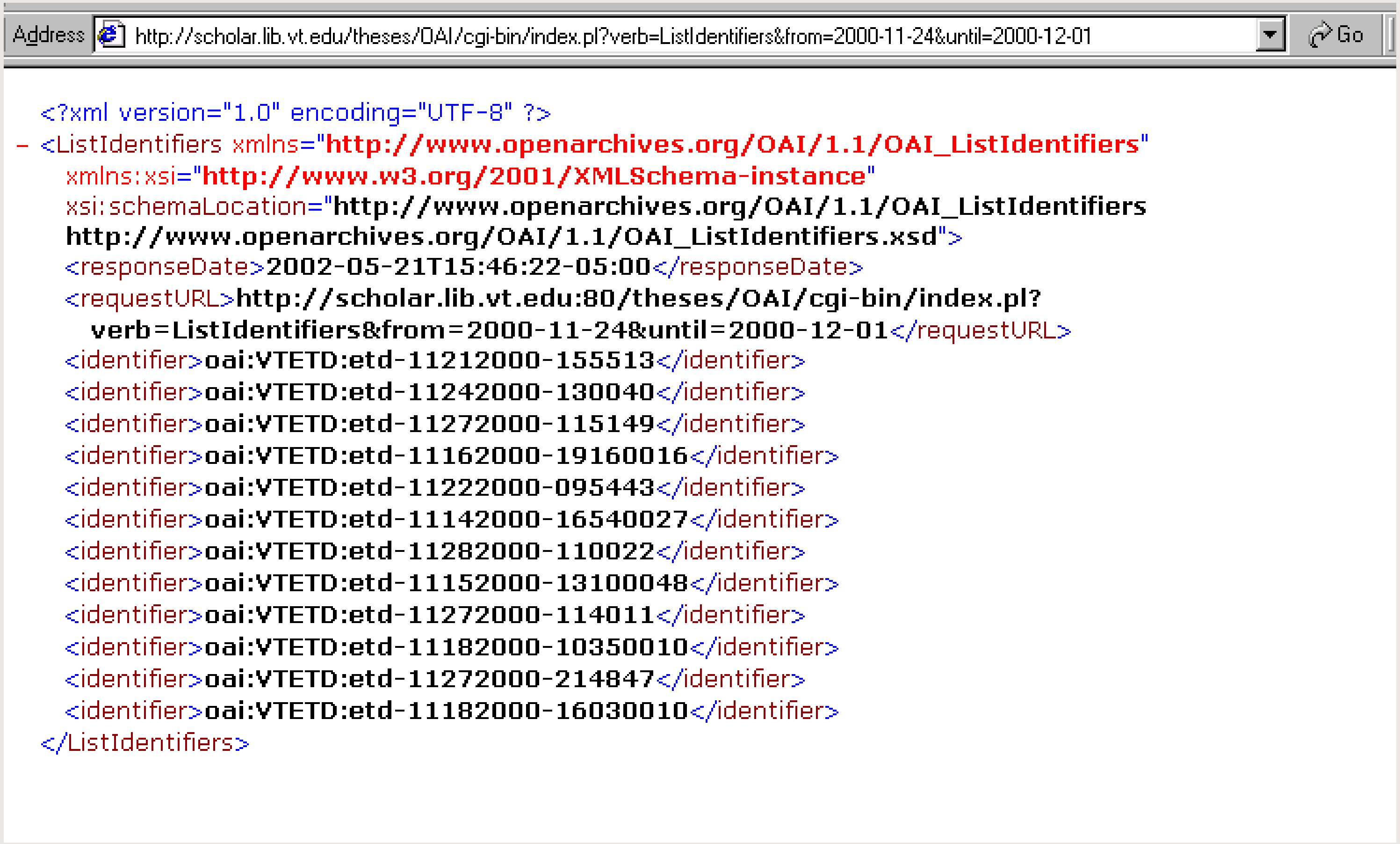
3.13. Metadata Multiplicity



edu/theses/OAI/cgi-bin/index.pl?verb=GetRecord&metadataPrefix=oai_rfc1807&identifier=oai:VTETD:etd-3123162539751141

```
  <?xml version="1.0" encoding="UTF-8" ?>
- <GetRecord xmlns="http://www.openarchives.org/OAI/1.1/OAI_GetRecord"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.openarchives.org/OAI/1.1/OAI_GetRecord
    http://www.openarchives.org/OAI/1.1/OAI_GetRecord.xsd">
    <responseDate>2002-05-21T15:45:08-05:00</responseDate>
    <requestURL>http://scholar.lib.vt.edu:80/theses/OAI/cgi-bin/index.pl?
      verb=GetRecord&metadataPrefix=oai_rfc1807&identifier=oai:VTETD:etd-
      3123162539751141</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/1.1/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>
          <author>Dano, Marie-Laure</author>
```

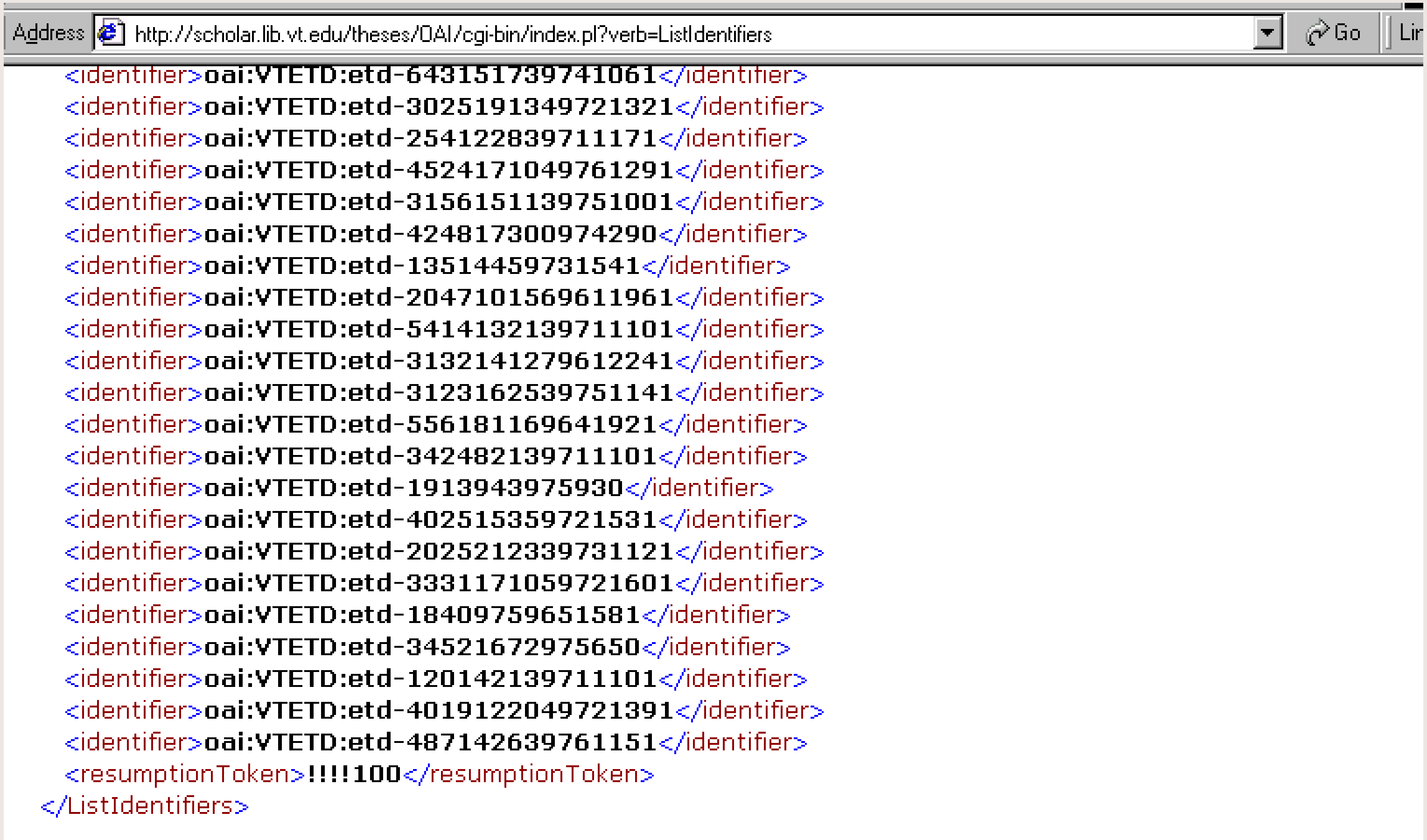
3.14. Date Ranges



Address http://scholar.lib.vt.edu/theses/OAI/cgi-bin/index.pl?verb=ListIdentifiers&from=2000-11-24&until=2000-12-01

```
  <?xml version="1.0" encoding="UTF-8" ?>
- <ListIdentifiers xmlns="http://www.openarchives.org/OAI/1.1/OAI_ListIdentifiers"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.openarchives.org/OAI/1.1/OAI_ListIdentifiers
    http://www.openarchives.org/OAI/1.1/OAI_ListIdentifiers.xsd">
    <responseDate>2002-05-21T15:46:22-05:00</responseDate>
    <requestURL>http://scholar.lib.vt.edu:80/theses/OAI/cgi-bin/index.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>
```

3.15. Resumption Token



http://scholar.lib.vt.edu/theses/OAI/cgi-bin/index.pl?verb=ListIdentifiers

```
    <identifier>oai:VTETD:etd-643151739741061</identifier>
    <identifier>oai:VTETD:etd-3025191349721321</identifier>
    <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>
```

4. OAI and ODL software

- No one needs to start from scratch !
- OAI supports the creation and distribution of toolkits and templates to implement the OAI-PMH.
- ODL (Open Digital Libraries) is a component framework for simple services that work with OAI-PMH-compliant archives.

4.1. Software to be installed

- To create an Open Archive using XML files: XML-File
- To test that it works: Repository Explorer
- To try harvesting data: Harvester
- To create a search engine: IRDB

4.2. Web Server Setup

- CGI capability needed for web server
  - Example for Apache

```
<Directory /home/*/public_html/cgi-bin>
   Options ExecCGI
   SetHandler cgi-script
</Directory>
```

- Note: May need minor tweaking for modperl

5. Creating an Open Archive: XML-File

- Data provider module that operates over a set of XML files which contain the metadata
- Requires minimal effort while retaining all the flexibility of the OAI protocol.

5.1. Features of XML-File

- OAI v1.1 protocol support
- Clean separation between engine, configuration and data
- FastCGI support (www.fastcgi.com)
- Hierarchical sets mapped from directory structure
- Multiple metadata formats generated on the fly
- Harvesting by date based on the file modification dates

5.2. Installation 1/4

- Extract all files into a directory from which the scripts can be executed using CGI.
  - Change to ~/public_html/cgi-bin/<station> where <station> is your machine number e.g., user01

    ```
    cd ~/public_html/cgi-bin/<station>
    ```

  - Download the file from the OAI-VT website if you don’t already have it

    ```
    wget http://www.dlib.vt.edu/
      projects/OAI/software/oai-file/
      oai-file.tar.gz
    ```

  - Decompress the file

    ```
    gzip –cd oai-file.tar.gz | tar –xf -
    ```

5.3. Installation 2/4

- Change to oai-file directory

  ```
  cd oai-file
  ```

- There will be three sub-directories: “config”, “scripts” and “data”
- Edit all the configuration files in the "config" directory
  - Define the archive name in “archiveid”

    ```
    joe config/archiveid
    ```

    - (or use your favorite *nix text editor)
    - change the word “oai-file” to your station name eg. user01

5.3. Installation 3/4

– Define/edit the metadata mappings in “metadata.pl”

```
joe config/metadata.pl
```

- (or use your favorite *nix text editor)
- change the phrase “/usr/local/bin/xsltproc” to “/usr/bin/xsltproc” since that is the location of the XSL transformation program on this server
- Do not change anything else!

5.5. Installation 4/4

- Define the response to Identify in "identity.pl“

  ```
  joe config/identity.pl
  ```

  - Replace “oai-file” in repositoryIdentifier and sampleIdentifier with your station name
- Look at some of the files in the “data” directory but don’t edit any.
- We will use the defaults for everything else !

6. Testing XML-File

- The script that implements an OAI data provider is

  `scripts/oaicgi.pl`

- The full baseURL is

  http://oss1.library.emory.edu/~hussein/cgi-bin/<station>/oai-file/scripts/oaicgi.pl

6.1. Direct execution

- First we can test by directly invoking the script to see if the script executes without any errors. Change to the “scripts” directory and run the following command:

```
QUERY_STRING=‘verb=Identify’
  ./oaicgi.pl
```

- You should see the XML response to Identify

6.2. Internet Explorer

- Run Internet Explorer and type in the following URL:

  http://oss1.library.emory.edu/~hussein/cgi-bin/<station>/oai-file/scripts/oaicgi.pl?verb=Identify

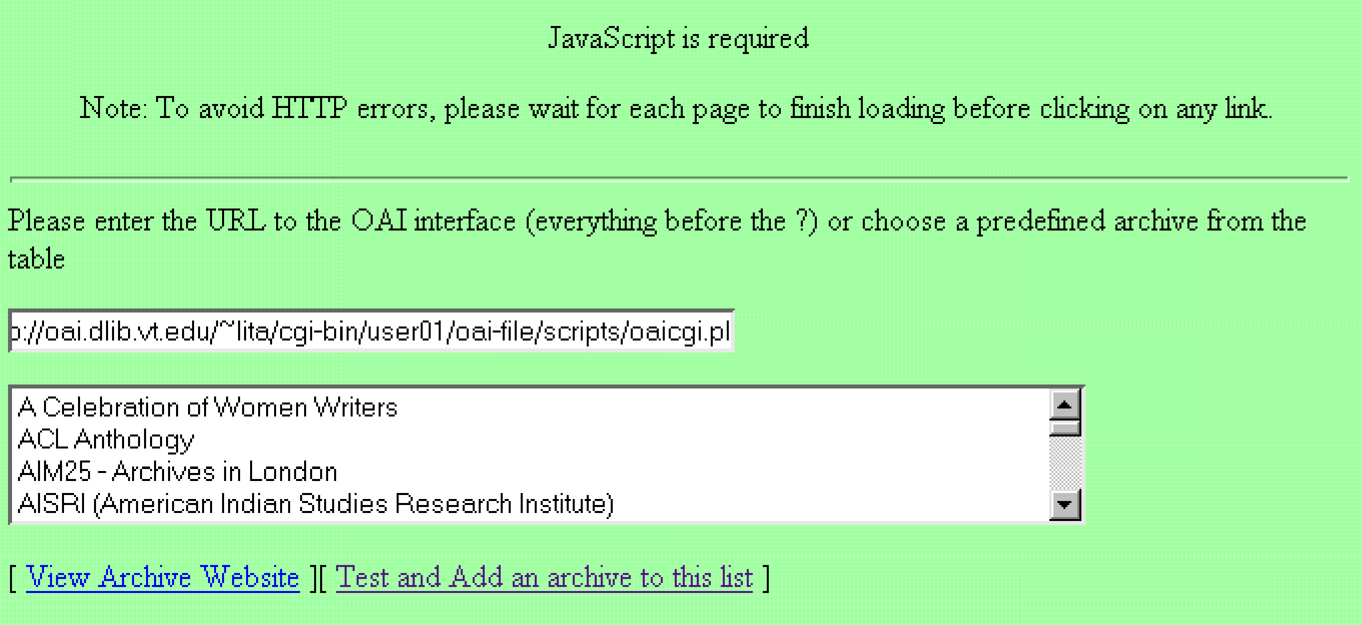
- You should get the response as before
- This also works in Netscape 6 but you have to “View Source” to see the output nicely formatted

6.3. Repository Explorer

- The Repository Explorer is a tool for testing Open Archives.
- You can issue individual commands and validate the results (using XML Schema)
- You can also perform a sequence of automatic tests
- http://purl.org/net/oai_explorer

6.4. Identify in RE

- Enter your baseURL in the RE and click on Identify



JavaScript is required

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

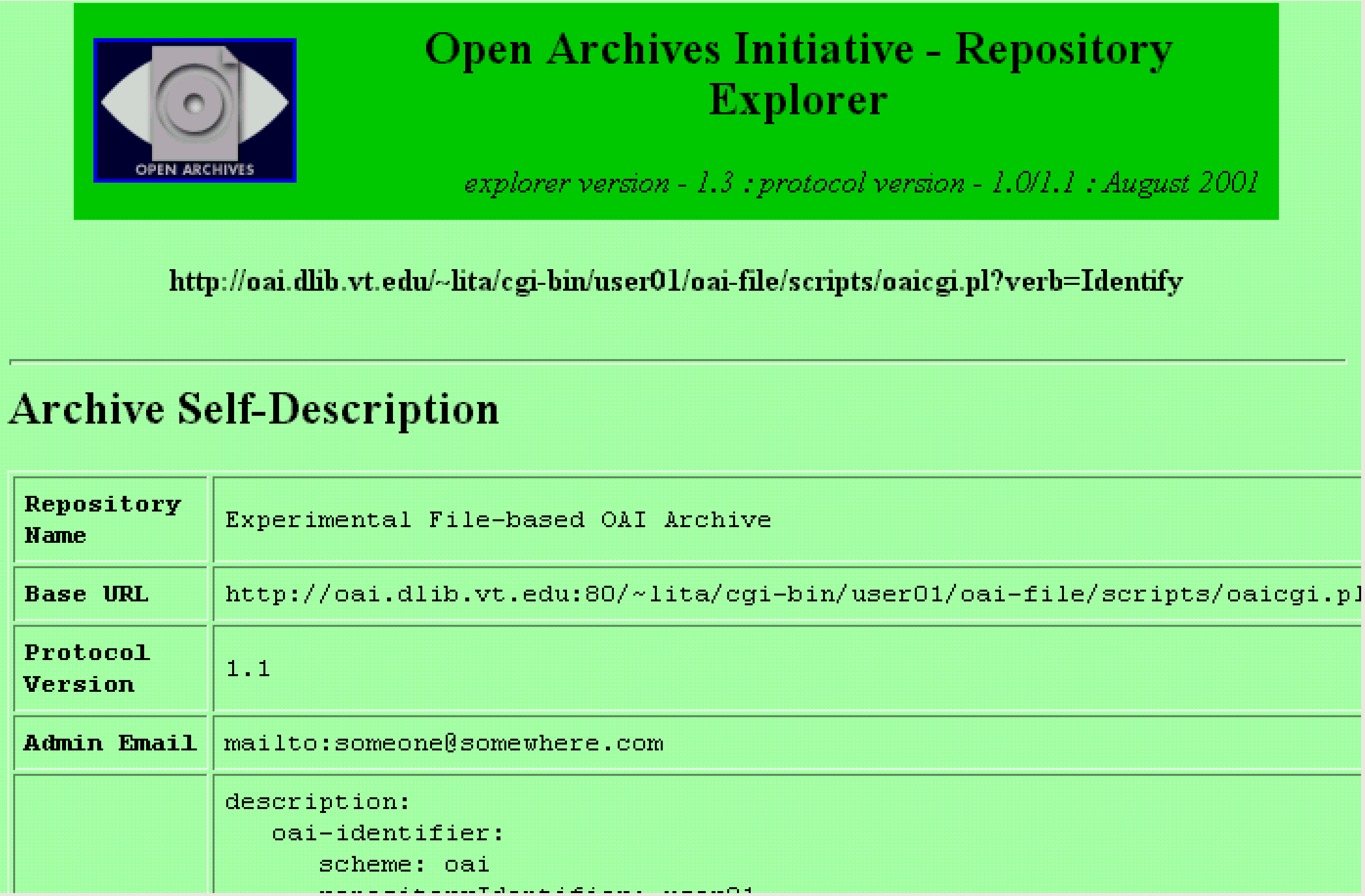
Please enter the URL to the OAI interface (everything before the ?) or choose a predefined archive from the table

p://oai.dlib.vt.edu/~lita/cgi-bin/user01/oai-file/scripts/oaicgi.pl

A Celebration of Women Writers
ACL Anthology
AIM25 - Archives in London
AISRI (American Indian Studies Research Institute)

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

6.5. Identify



Open Archives Initiative - Repository Explorer

explorer version - 1.3 : protocol version - 1.0/1.1 : August 2001

http://oai.dlib.vt.edu/~lita/cgi-bin/user01/oai-file/scripts/oaicgi.pl?verb=Identify

Archive Self-Description

Repository Name	Experimental File-based OAI Archive
Base URL	http://oai.dlib.vt.edu:80/~lita/cgi-bin/user01/oai-file/scripts/oaicgi.pl
Protocol Version	1.1
Admin Email	mailto:someone@somewhere.com
	description: oai-identifier: scheme: oai repositoryIdentifier: user01

6.6. Other functions

- Try clicking on the other verbs to see what the effect is
- Parameters are necessary for some verbs (like GetRecord) and optional for others
- “Display” can change whether you see the original XML, a parsed version (default), or both

6.7. Automatic Tests

- Click on “home” at the bottom of the page and select “Test and Add an archive”
- Enter the baseURL on the next page and click “Test the archive”
- This will perform a set of tests to verify that the OAI interface works and is somewhat robust – (do not register your archive)

6.8. Add more data

- Switch to your telnet session
- Change to the “data” directory
- Make a duplicate of one of the files there (e.g., compend1.xml) – choose any name with a “.xml” extension
- Edit some or all of the fields in the file
- Go back to the browser, click home, enter the baseURL, and try ListIdentifiers again. You should have one more entry.

7. Installing a Harvester

- Harvester is a service provider module that implements an algorithm to get periodic updates from an Open Archive
- Object-Oriented Perl allows subclassing to integrate this into other tools.
- The supplied sample code outputs records to the screen.

7.1. Installation

- Extract all files
  - Change to ~/public_html/cgi-bin/<station> where <station> is your machine number e.g., user01

    ```
    cd ~/public_html/cgi-bin/<station>
    ```

  - Download the file from the ODL website if you don't already have it

    ```
    wget
      http://oai.dlib.vt.edu/odl/software/harve
      st/Harvest-1.11.tar.gz
    ```

  - Decompress the file

    ```
    gzip -cd H* | tar -xf -
    ```

7.2. Configuration

- Change to “ODL-Harvest/Harvest”
- Run

```
./configure.pl <station>
```

- Add one archive - the one we just created
- Answer all questions as indicated on next slide

7.3. Harvester Parameters

- Archive identifier: <station>
- baseURL of the archive:
  - from previous exercise
- How often to harvest: 86400 (default)
- Overlap: 172800 (default)
- Granularity: day (default)
- metadataPrefix: oai_dc
- set (leave empty):   (default)

7.4. Harvesting

- Run

```
<station>/harvest.pl
```

- This will do an initial harvest of the archive – records will be displayed on screen
- Run it again – since the time interval has not elapsed, nothing will be displayed
- Force an immediate (now) harvest of all records (start) from all defined archives (all) by issuing:

```
<station>/harvest.pl now all start
```

8. Installing a Search Engine: IRDB

- Harvesting is useful to either import data into a system or to create services such as search engines
- IRDB is a small-scale search engine that gets its data from an Open Archive and has a simple machine interface to issue queries

8.1. Features

- Works with any OAI source
- Indexes any metadata format
- No pre-requisite software except a database that can be accessed by Perl's DBI
  - *We will use mySQL, where the administrator has already created a database and assigned "all privileges" to the user account.*

8.2. Installation

- Extract all files
  - Change to ~/public_html/cgi-bin/<station> where <station> is your machine number e.g., user01

    ```
    cd ~/public_html/cgi-bin/<station>
    ```
  - Download the file from the ODL website if you don’t already have it

    ```
    wget
      http://oai.dlib.vt.edu/odl/software/irdb/
      IRDB-1.02.tar.gz
    ```
  - Decompress the file

    ```
    gzip -cd I* | tar -xf -
    ```

8.3. Configuration

- Change to “ODL-IRDB/IRDB”
- Run

```
./configure.pl <station>
```

- Answer questions as in the following slide

8.4. IRDB Parameters

- Database connection
  - Driver : mysql
  - Database : lita
  - Username : hussein
  - Password (leave blank):
- Database Table, Repository Name, Admin Email, Archive Identifier: leave at defaults
- Archive URL: enter the baseURL for the XML-File archive
- Use defaults for everything else

8.5. Test IRDB

- To populate with data from the Open Archive:

```
<station>/harvest.pl
```

- To run a test query from the command-line:

```
<station>/testsearch.pl “test”
```

- To issue a query to the machine (ODL) interface try:

```
QUERY_STRING='verb=ListRecords&metadat
  aPrefix=oai_dc&set=odlsearch1/test/1
  /10' <station>/search.pl
```

8.6. Web Server Permissions

- The apache web server will not run a script if the directory is group-writable. IRDB uses default permissions so you may need to disable group-writing with:

```
chmod 755
  /home/hussein/public_html
  /cgi-bin/<station>
  /ODL-IRDB/IRDB/<station>
```

9. A quick user interface

- A search engine is not very useful without a user interface
- We can either parse the XML and generate HTML or use some kind of transformation or stylesheet
- IRDB has a sample interface that can be installed

9.1. Installation

- Extract all files
  - Change to ~/public_html/cgi-bin/<station>
    where <station> is your machine number e.g., user01

    ```
    cd ~/public_html/cgi-bin/<station>
    ```

  - Download the file from the ODL website if you don't already have it

    ```
    wget
      http://oai.dlib.vt.edu/odl/software/compu
      te_ui/compute_ui.tar.gz
    ```

  - Decompress the file

    ```
    gzip -cd c* | tar -xf -
    ```

9.2. Configuration

- Edit the “search.pl” file in the UI directory – change the baseURL to:

```
http://oss1.library.emory.edu
  /~hussein/cgi-bin/<station>
  /ODL-IRDB/IRDB/<station>
  /search.pl
```

- The rest of the file can be changed to change the interface appearance, but we will ignore it for now!

9.3. Testing the interface

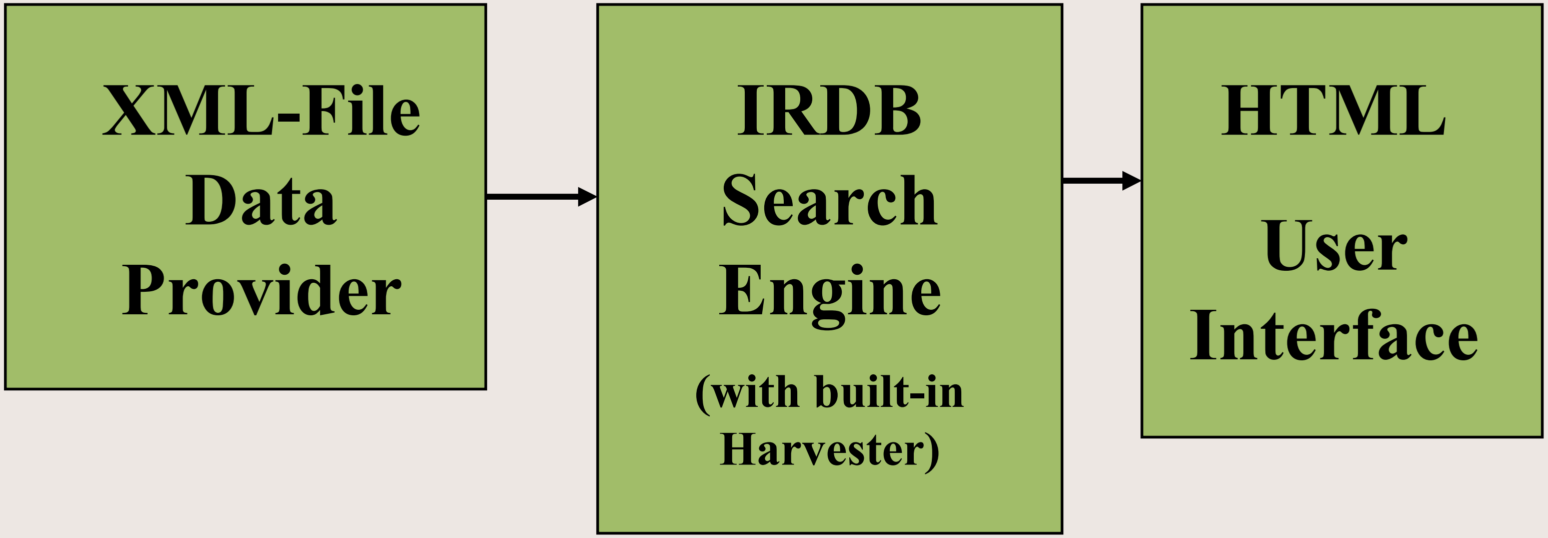
- Enter the URL into your web browser as:

  http://oss1.library.emory.edu/~hussein/cgi-bin/<station>/UI/search.pl

  - Try a query such as “test”, “art”, or “war” (or any other word that appeared in the metadata)
- Note: The links will not work since we did not edit that part of the search.pl script

10. Wrap up and discussion

- You have just built a digital library out of components !



XML-File Data Provider

IRDB Search Engine (with built-in Harvester)

HTML User Interface

10.1 Final Thoughts

- OAI-PMH is a simple protocol for exporting and importing metadata
- Components based on OAI can be used to build modular systems
- Lots of tools available now !
- Lots of interest from other people already, even publishers!

11.1. Links

- Open Archives Initiative
  - http://www.openarchives.org
- OAI Metadata Harvesting Protocol
  - http://www.openarchives.org/OAI/openarchivesprotocol.htm
- Virginia Tech DLRL OAI Projects (XML-File)
  - http://www.dlib.vt.edu/projects/OAI/
- Repository Explorer
  - http://purl.org/net/oai_explorer
- Open Digital Libraries (Harvester, IRDB)
  - http://oai.dlib.vt.edu/odl

11.2. 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