Toward Universal Accessibility of ETDs: Building the NDLTD Union Archive

Hussein Suleman, hussein@vt.edu

Edward A. Fox, fox@vt.edu

Digital Library Research Laboratory

Virginia Tech

Virginia Tech NDLTD DLRL OPEN ARCHIVES

Overview

Virginia Tech

- Background and History
- OAI Protocol for Metadata Harvesting
- The Union Archive
- Experiences: Encodings
- Experiences: Components
- Experiences: Harvesting
- ETD-MS
- Rights Management
- Service Providers
- How to Participate
- ETD-db
- Future Work

Background and History

Virginia Tech

- Independent collections
- Federated Search
  - Query sent to all sites and results were collected
    - Multiple points of failure
    - Slow
    - Merging is not simple
    - Differences among search technology
- Other federated projects (e.g., NCSTRL) were gradually moving towards other approaches

OAI Protocol for Metadata Harvesting

Virginia Tech

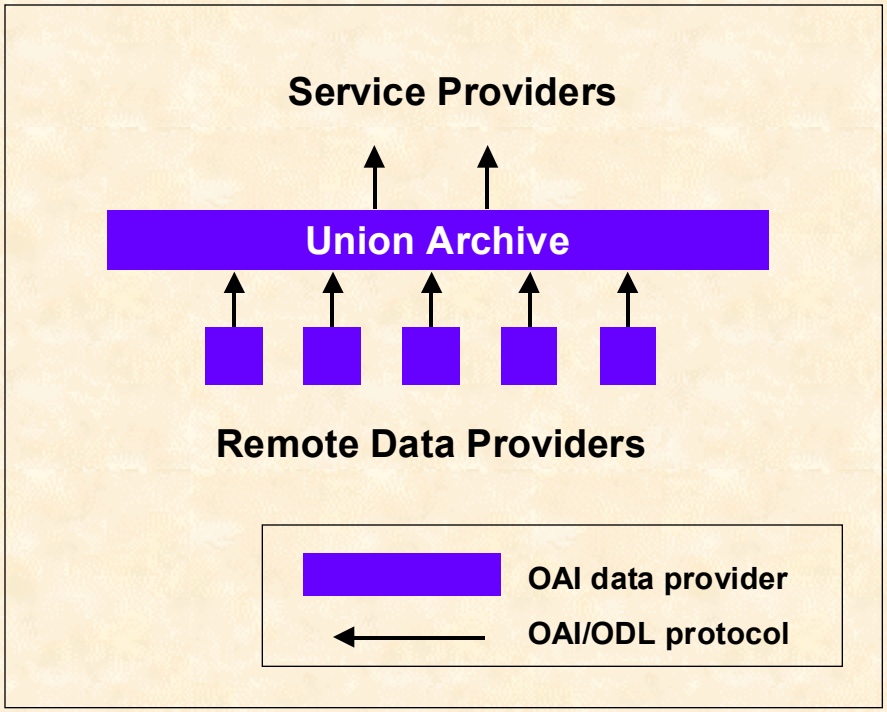
- Open Archives Initiative (OAI)
  - Organization dedicated to solving problems of digital library interoperability by defining simple protocols, most recently for the exchange of metadata
- Protocol for Metadata Harvesting
  - Transfer metadata from one archive to another
    - Incremental updates, any type of metadata

Sample OAI Metadata Record

```
<record>
      <header>
         <identifier>oai:VTETD:etd-232815339749 20</identifier>
         <datestamp>1997-03-31</datestamp>
      </header>
      <metadata>
         <dc>
            <title>Weed Management Programs in Potato, Transplanted
   Tomato and Transplanted Pepper with Rimsulfuron and Other
   Herbicides</title>
            <creator>Ackley, John A.</creator>
            <language>en</language>
            <contributor>Henry P. Wilson</contributor>
            <contributor>Kriton K. Hatzios</contributor>
            <contributor>Ronald D. Morse</contributor>
            <format>application/pdf</format>
         </dc>
      </metadata>
   </record>
```

The Union Archive

- Collects metadata from remote OAI-compliant sources (data providers)



Experiences: Encodings

Virginia Tech

- Some data providers have non-English metadata
- OAI requires using UTF-8 or Unicode numerical entities
- Entities must not be converted in the intermediate stages of processing – (contrary to default XML behavior)
- Some data providers have XML that must be "cleaned"
- All other data is stored as is

Experiences: Components

Virginia Tech

- Open Digital Libraries: Use the OAI protocol as the basis for DL components to communicate
- Examples
  - Search Engines could use dynamic sets to correspond to search results
  - Categorical browsing can be directed by sets
- The Union Archive is one ODL component, serving data to other components and other DLs

Experiences: Harvesting

Virginia Tech

- At least once daily
- Incremental, based on dates
- Delays between requests – to prevent DoS
- Independent for each data provider – aids robustness since data providers fail often!
- Overlapped range of dates to avoid losing records

ETDMS

Virginia Tech

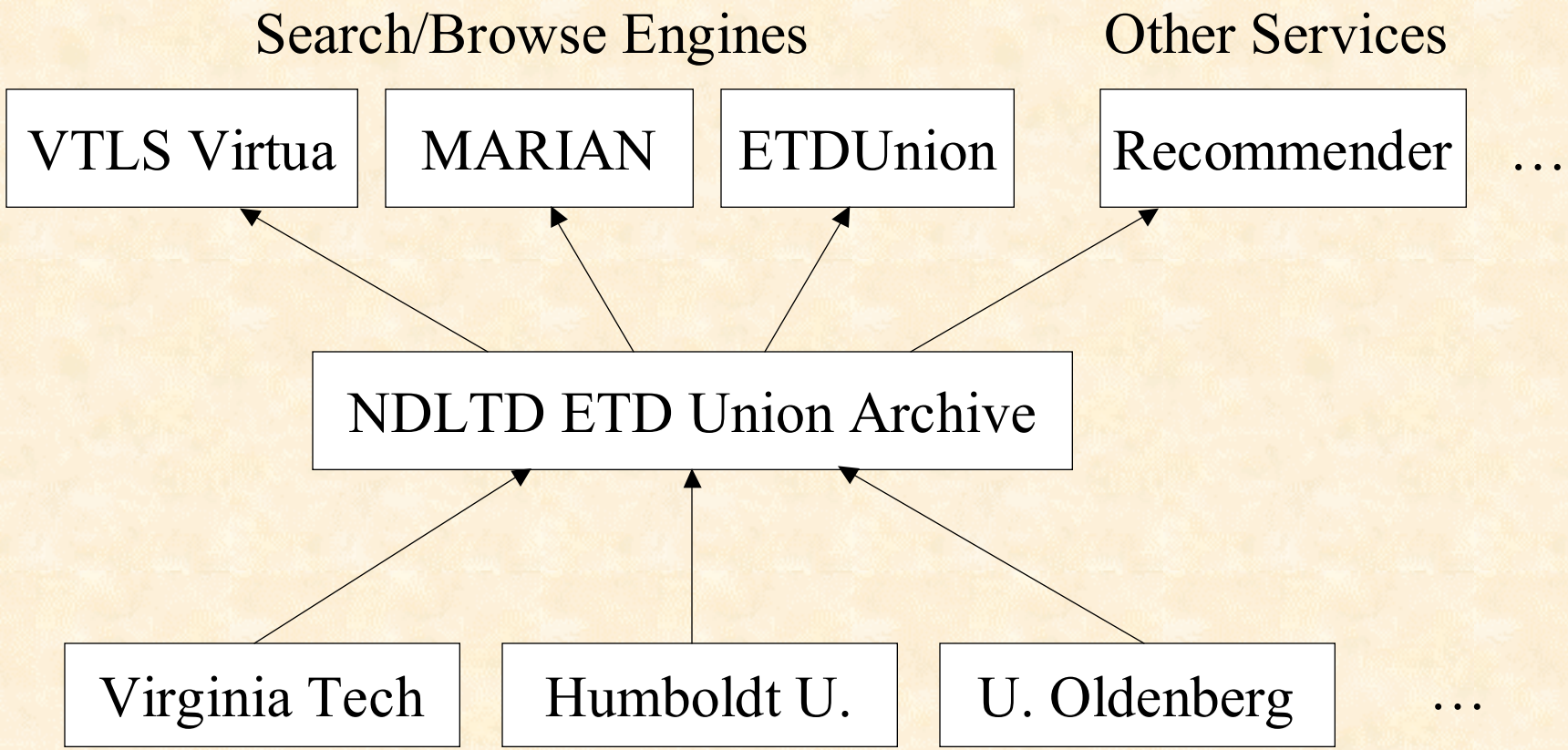
- Electronic Thesis and Dissertation Metadata Set
- Extension of Dublin Core aimed at supporting ETDs
- Supported by ~half of the data providers
- Union Archive harvests DC and ETDMS (if it exists)

Rights Management

Virginia Tech

- Only freely-accessible records are harvested!
- Original identifiers are retained
- Unmodified data is republished
- Metadata can have links back into the source collection
  - branding
  - further rights management

Architecture



VTLS Virtua (production)

Virginia Tech

NETWORKED DIGITAL LIBRARY
OF THESES AND DISSERTATIONS

SEARCHES **Scan** | **Keyword** | **Expert** | English

New Session | Save Session | Cart | View History | Help

Scan Search

Enter your search terms. | Select a search type. | Choose a target database. | Click the **Search** button.

Author | **NDLTD Collection** | Search

Welcome to the Networked Digital Library of Theses and Dissertations Union Catalog. This Union Catalog will serve as a repository of rich graduate educational material contributed by a number of member institutions worldwide. Our hope is that this project will increase the availability of student research for scholars, empower students to convey a richer message through the use of multimedia and hypermedia technologies and advance digital library technology worldwide.

This project is a joint development with NDLTD and VTLS Inc. Read the press release here.

Browse the catalog by theses and dissertation author, subject, title and library call number. Click on "Browse Search" in the toolbar to initiate your search.

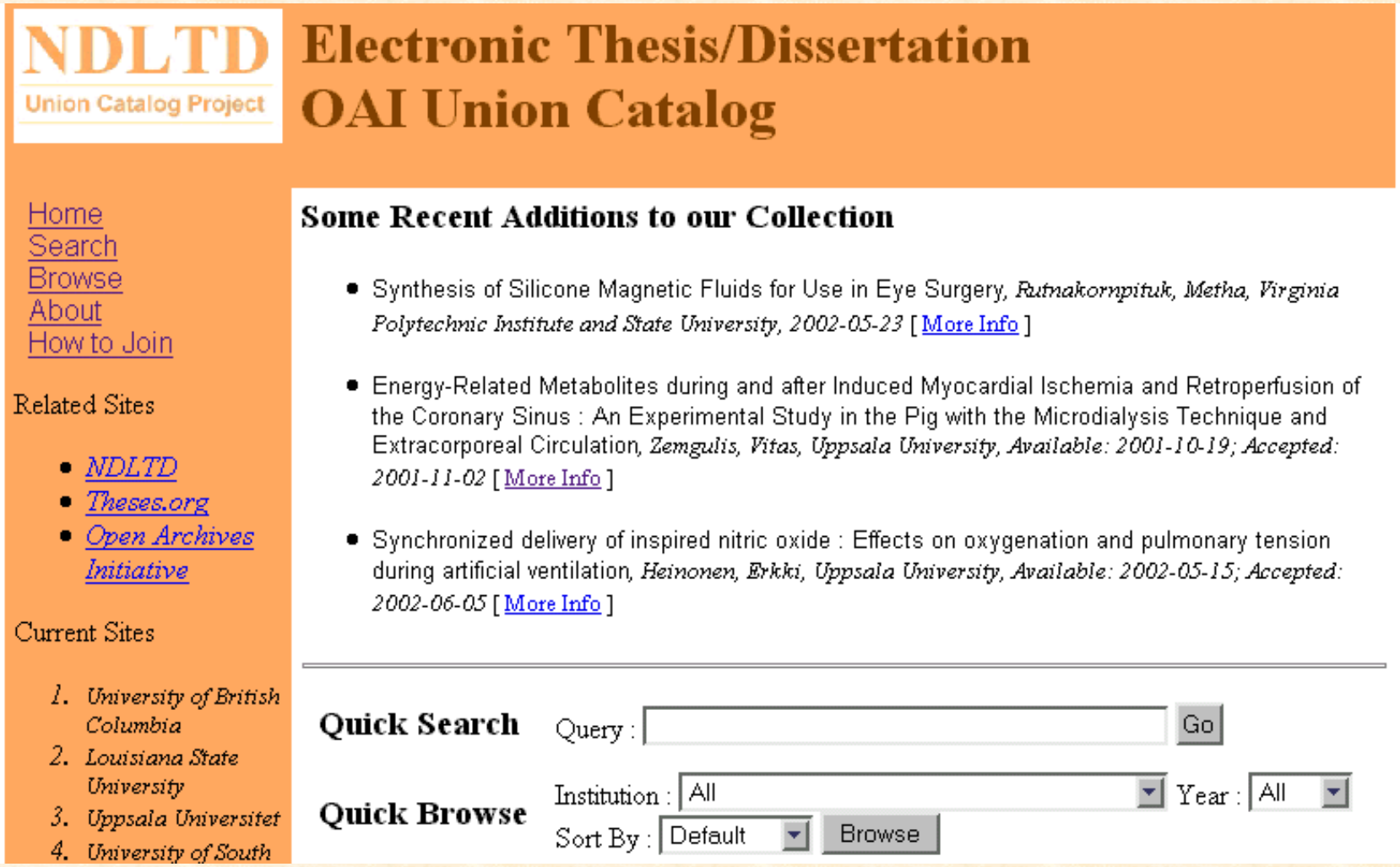
Search for words or exact phrases anywhere in the theses and dissertation catalog record. Specify a keyword field such as author, title or subject to narrow that search. Click on "Keyword Search" in the toolbar to begin.

The Expert search function allows you to use Boolean operators in a command mode. Click on "Expert Search" to get started and follow the helpful search instructions below the data entry box.

Copyright © 2001 VTLS Inc. All rights reserved. 1701 Kraft Drive, Blacksburg, VA, USA 24060 | VTLS.com

ODL ETDUnion (research)

Virginia Tech



Data Provider Participants

Virginia Tech

- Virginia Tech (US)
- Humboldt University of Berlin (Germany)
- University of Duisburg (Germany)
- Technical University of Dresden (Germany)
- PhysDis (Germany)
- MIT (US)
- CalTech (US)
- Uppsala University (Sweden)
- University of South Florida (US)
- Louisiana State University (US)
- University of British Columbia (Canada)
- U. Hong Kong (HK) and U. of the Americas-Puebla (Mexico)
  - coming soon!

How To Participate

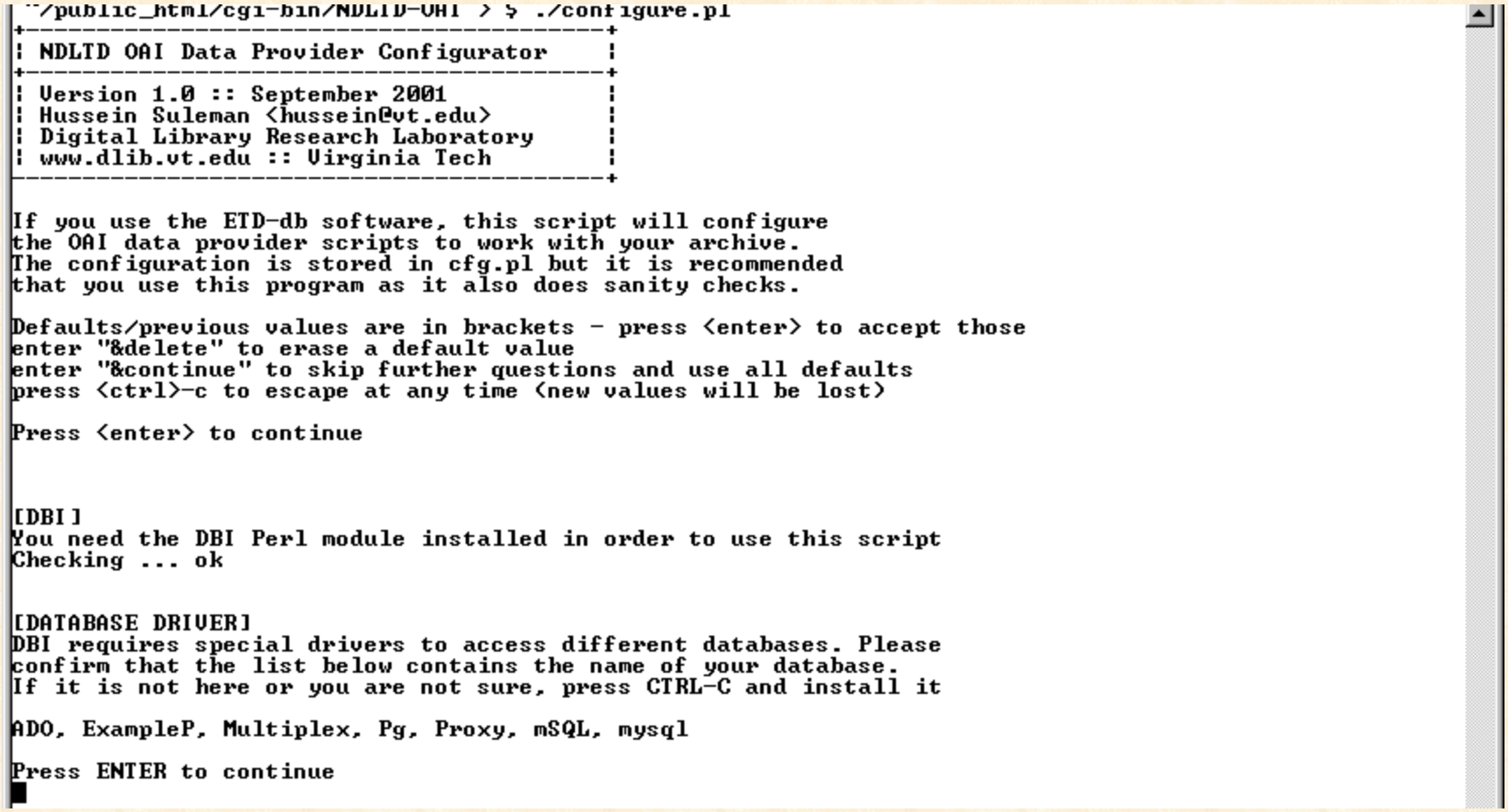
- Make your ETD (or TD) collection into an Open Archive
  - Use software available on OAI website
  - Use ETD-db extensions
- Provide innovative services to the NDLTD community based on the Union Archive

ETD-db Extensions

Virginia Tech

- Server extension to make ETD-db OAI-compliant
- Perl script that works with all versions of ETD-db and is distributed with the latest ones
- Download from
  - http://www.dlib.vt.edu/projects/OAI/software/ndltd/ndltd.html

ETD-db Configuration 1



ETD-db Configuration 2

```
[DATABASE CONNECTION]
The OAI scripts need to connect to your ETD database to obtain
the metadata for your ETDs. For this you need to supply the name of the
database as well as the hostname of the machine that runs it
(you will be asked for username and password next).
If you dont have permission to access the database from this host
you may want to install this OAI script on another machine or grant
permissions before you continue.

Drivers: ADO, ExampleP, Multiplex, Pg, Proxy, mSQL, mysql
Choose database driver from above list [mysql] :
Enter hostname/IP of database server [localhost] :
Enter name of database [etd_available] :
Enter the username to be used to connect [] :
Enter the password to be used to connect [] :
Checking ... ok

[DATABASE TABLES]
Checking etd_main ... ok
Checking keyword_by_urn ... ok
Checking filename_by_urn ... ok
Checking advisor_by_urn ... ok

[ARCHIVE ID]
Each OAI-compliant archive ought to have a unique identifier associated
with it. This is usually the common acronym for the institution.
Examples: VT, MIT, UDW, HUBerlin.
Restrictions: It has to be alphanumeric.

Enter your Archive ID [VTETD] :
```

ETD-db Configuration 3

```
[REPOSITORY NAME]
For identification purposes, you need to specify the full name of
the repository
Example: Virginia Tech Electronic Thesis and Dissertation Collection.

Enter your repository name [VT Electronic Thesis and Dissertation Archive] :

[ADMINISTRATOR EMAIL]
When your repository is registered and used by others, any problems
will be reported to an administator, whose email must be provided
Examples: admin@reposit.org, etds@etdcentral.edu.

Enter your administrator's email [hussein@rocky.dlib.vt.edu] :

[PUBLISHER]
Most metadata formats require some information about the publisher
of the ETD, which is usually the university that granted the degree.

Name of publisher/institution [VT] :
Department responsible for ETDs [University Libraries] :
Acronym for publisher [VT] :
Location of publisher [] :
This field cannot be left blank. Please enter a value.
Location of publisher [] : blacksburg

[LANGUAGE]
Example: en, en-uk, fr, es

Language for metadata [en] :

[SAMPLE IDENTIFIER]
Example: etd-1234567890
Note: Dont prepend an "oai:archiveid:"

Sample identifier [etd-1234567890] :
```

ETD-db Configuration 4

```
[RECORD LIMIT]
This is the number of records disseminated in each batch.

No of records [100] :

[RESUMPTION TOKEN SEPARATOR]
This character can be anything not found in an identifier, typically punctuation.
Example: !, ?

Separator [!] :

[POLICIES]
As additional information, your archive should have policies for use of its data
and metadata by others

What is the content of the archive [Theses and Dissertations produced by students] :
What is the usage policy for the ETDs (not the metadata) [Full texts are individually tagged and the ri
ghts statements must be adhered to] :
What is the usage policy for the metadata [Metadata may be used by commercial and non-commercial users]
 :
What is the policy for making submissions [Submissions are only made internally by students] :



Finis.

Now you can test the archive from the Repository Explorer at
http://purl.org/net/oai_explorer - use the "Test and Add" option
For a URL, use the web server path to the "oai.pl" script in this
directory.
After you confirm that it works properly, register your archive
at http://www.openarchives.org and send email to
etdunion@oai.dlib.vt.edu so that we can add your collection to the
NDLTD Union Catalog.
```

Data Provider Registration

Virginia Tech

- Test your OA with the Repository Explorer
  - http://purl.org/net/oai_explorer
- Send the URL to your OAI interface to NDLTD
- Register your OA with the OAI
- And that's it !

Future Work

Virginia Tech

- Build more services
- Encourage more data providers
- TD and ETD collections
- Scalability and reliability for union collection

Links

- NDLTD
  - http://www.ndltd.org
- 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
- Open Digital Libraries
  - http://oai.dlib.vt.edu/odl

Virginia Tech

That's All Folks !