Towards Universal Accessibility of ETDs: Building the NDLTD Union Archive

Hussein Suleman
Department of Computer Science
Virginia Tech, Blacksburg, VA, USA
+1 540 231 3615
hussein@vt.edu

Edward A. Fox
Department of Computer Science
Virginia Tech, Blacksburg, VA, USA
+1 540 231 5113
fox@vt.edu

ABSTRACT

By virtue of its name, NDLTD (Networked Digital Library of Theses and Dissertations) suggests to newcomers the existence of a collection of accessible ETDs. While many current members do indeed have such collections at the level of institutions or consortia, there is no single mechanism to aggregate all ETDs to provide NDLTD-wide services (e.g., searching). Past attempts have focussed on federated searching and have achieved a small measure of success. However, current trends in digital library architecture suggest that metadata harvesting is a simpler and more robust solution. This approach is actively advocated by the Open Archives Initiative (OAI), which recently devised a protocol to support such a mode of interoperability. The Union Archive project is an attempt to implement and build upon the work of the OAI by setting up and maintaining a central metadata collection, fed by OAI-compliant remote sites. NDLTD members are encouraged to contribute from their host sites to this archive by supporting the OAI protocol, with particular conventions that are specific to ETDs. A preliminary analysis of the operation of the Union Archive both supports the interoperability approach of the OAI and paves the way for a larger-scale project within the community of ETD archives.

Keywords

Interoperability, standards, protocols, union archive

BACKGROUND

NDLTD is a loose federation of member institutions and organizations that publish ETDs. While not a primary goal, it has long been a desire of NDLTD to unite the various member sites into a single collection in the perception of the average researcher or student seeking ETDs [Fox, et al., 1996; Fox, et al., 1997].

The first attempt to provide services across multiple NDLTD members was the federated search system [Powell and Fox, 1993]. This system allowed users to go to a central site and perform a meta-search across multiple ETD collections. The search language and syntax at each site was known and stated internally in a precise specification language. This then was used to reformulate the users' queries appropriately for each site. The search results were cached as they arrived but the results were not merged. This federated search system was popular as a proof-of-concept of the capabilities of federated searching. There were, however, many shortcomings with this initial system, including:

- The reliance on every participating member site resulted in more points of failure – every time a member site changed any aspect of its search engine interface, the federated search system would need reconfiguration.
- Searches were as slow as the slowest member site. And if a remote site were down, the federated search results would be incomplete.
- There was no simple way to merge result sets since this would require parsing of the result set pages (which were in HTML) as well as retrieving each individual document or metadata record listed prior to ranking.
- It was not always simple to add new remote sites to the system since each new site could require the specification of a new search language and/or syntax. Furthermore, when remote search engines were not based on single-request stateless operations, additional complexity could result from the client-server interaction.

Similar observations were made by other projects that relied on federated searching. A notable example is the original NCSTRL project [Leiner, 1998], which used the Dienst protocol and software [Lagoze and Davis, 1995] to create a distributed digital library of technical reports in the field of computer science. Reliability and speed problems were addressed in NCSTRL by replication and caching. The complexity of mapping search queries and results was avoided by establishing a standard protocol and an associated software toolkit. Dienst's major shortcomings were that adoption was considered a technical hurdle by many implementers and support of remote sites became an increasing burden as the library grew larger.

In order to address all of these problems, representatives of NCSTRL, NDLTD, and various other organizations that had an interest in digital library interoperability, met in Santa Fe in October 1999 to develop a simpler solution for large-scale production-quality interoperability. The result of this meeting was the Santa Fe Agreement [Van de Sompel and Lagoze, 2000], a set of guiding principles for interoperability and a protocol for transferring metadata among sites. The organization that grew out of that meeting, the Open Archives Initiative [OAI, 2002], has further developed the original agreement and actively advocates and supports it in its current form – the Open Archives Initiative Protocol for Metadata Harvesting [Van de Sompel and Lagoze, 2001].

THE OAI PROTOCOL FOR METADATA HARVESTING (OAI-PMH)

The OAI-PMH is a client-server protocol based on HTTP, which facilitates the incremental transfer of metadata among networked systems. The protocol is designed to be simple and general, thus suitable for use in various different contexts and communities. NDLTD, as one such community, has adopted this protocol to develop an alternative to the federated search system used in the past.

The OAI-PMH differs from previous approaches to interoperability in that there is very little interaction among remote sites and the central site. The primary purpose of the protocol is incremental bulk transfer of metadata (harvesting) – there is no remote search facility. Instead, a provider of services acquires data from a data provider, stores and processes it locally, and then supplies services to users based on that data. A brief summary of the requests that make up the core of the protocol is provided in Table 1.

Service Request	Expected Response
Identify	Description of archive - standards and protocols implemented
ListMetadataFormats	List of supported metadata formats
ListSets	List of archive sets and subsets
ListIdentifiers	List of record identifiers, optionally corresponding to a specified set and/or date range
GetRecord	Single metadata record corresponding to a specified identifier and in a specified metadata format
ListRecords	List of metadata records corresponding to a specified metadata format and, optionally, a set and/or date range

Table 1. OAI-PMH service requests and expected responses

The OAI-PMH uses current standards wherever applicable. All data that is transferred in response to a request is encoded in an XML format defined using XML Schema [Fallside, 2001], a more precise structural specification language than DTDs. Based on these specifications, tools such as the Repository Explorer [Suleman, 2001] can perform machine validation of the responses. Requests are encoded using the CGI mechanism because of its stability as a Web technology – when Web Services standards such as SOAP [Box, et al., 2000] are eventually finalized, the OAI-PMH can easily be retargeted to use those mechanisms.

NDLTD has adopted use of the OAI protocol for metadata transfer because of the ease with which the protocol can be customized to handle the specific needs of NDLTD. Included among these are:

- Simplicity: various toolkits can be (and have been) developed to ease adoption.
- Support for arbitrary metadata formats: NDLTD has developed a new standard for metadata specific to ETDs (ETD Metadata Set) and this is trivially supported by the OAI protocol.
- Sets within the archive: in the context of NDLTD these can correspond to subject areas, geography, and organizational structures.

THE UNION ARCHIVE

As NDLTD members implement the OAI protocol on their individual archives, work has begun on developing mechanisms and policies for collecting metadata into a central site for the provision of user-level services across all participating NDLTD members.

In order to support multiple service providers, the Union Archive was created with the purpose of collecting metadata from remote sites and republishing it as a single collection. Sites that provide user-level services such as searching and browsing then can harvest from this single collection. Some of the advantages of a central merged collection of metadata are:

- Multiple services can be developed at a central site, without having to harvest metadata multiple times – this is analogous to caching or replication.
- The development of experimental services does not negatively impact the production servers of individual members by repeatedly requesting transfers of the same metadata.
- The system is robust – the number of points of failure is reduced to exactly one. This is traditionally not considered wise but, in the networked information community, it is simpler to replicate or replace a single point of failure than deal with arbitrary and distributed points of failure.

The Union Archive is designed to function as both a provider of services (harvester) and a provider of data. It harvests metadata from the remote sites, stores it in an internal database and then republishes this metadata through its own OAI data provider interface. This is illustrated in Figure 1.

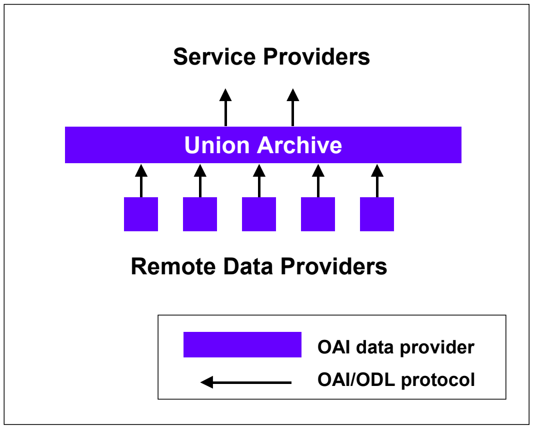

Figure 1. Union Archive Architecture

There are currently 9 ETD collections that participate in this project, and more are in the process of making their collections OAI-accessible. The nine collections and their respective document sizes are listed in Table 2.

Archive	# Records
Virginia Tech	3096
Humboldt University of Berlin	299
University of Duisburg	145
Technical University of Dresden	18
PhysNet	185
MIT	72
CalTech	27
Uppsala University	1482
University of South Florida	22

Table 2. Contents of Union Archive collection (as of 14 February 2002)

Some additional sites contributed metadata directly to the Union Catalog maintained by VTLS. These records are in the process of being incorporated into the Union Archive.

EXPERIENCE AND LESSONS LEARNT

Encodings

With many NDLTD members being from non-English-speaking countries, languages and character sets are an important problem that needs to be solved in any distributed digital library solution. Of the 9 current participants in the project, 5 use metadata that cannot be encoded using standard ASCII. While extended ASCII may work for now, it is anticipated that other members from Asian countries will provide data that cannot be encoded in any variation of ASCII.

As a general solution to this problem, the OAI requires that all data transmitted by the OAI protocol be encoded in UTF-8 with numerical Unicode entities wherever necessary. This ensures that there is no loss of data and that validation of XML is always possible without the need for external entity files. However, it raises the bar on the development of tools to process this data. All tools must support UTF-8 – this is usually not much of a problem as long as data is converted to a human-readable format when needed. But since the XML standard requires that parsers translate all entities, any intermediate XML parser or XSLT transformation converts entities to an appropriate human-readable form. To work around these conversions of the intermediate data, Unicode entities are double-escaped in the Union Archive before any XML fragments are passed to XML parsers.

Components

The Union Archive was developed in conjunction with and as a test case for the Open Digital Library (ODL) methodology [Suleman and Fox, 2001]. ODL is a set of guiding principles and protocols for building digital libraries as networks of communicating components, where each component is an extended Open Archive.

Componentized software development is the latest trend in software development for Web-based systems, and ODL is an attempt to leverage existing OAI standards to make the transition simpler and faster. By building the Union Archive as an ODL component, it can easily be integrated to interoperate with other OAI and ODL components. This makes it simpler to gather data from remote sources, as well as serve as an OAI/ODL data source for other components that provide innovative services.

Harvesting Schedule

The current version of the OAI protocol (v1.1) supports incremental harvesting whereby service providers may request records by specifying only a date range. Harvesting algorithms designed to deal with this absence of timestamps have to contend with either marginally outdated data or duplicate records [Suleman and Fox, 2002]. In practice, the Union Archive harvests data from remote locations once daily since ETD metadata is relatively static and it is not critical to incorporate new entries immediately.

ETD Metadata Set (ETDMS)

ETDMS is a standard developed for expressing metadata related to theses and dissertations [Atkins, et al., 2001]. The Union Archive harvests records in ETDMS from all archives that support it, in addition to the default Dublin Core format. Currently, only 4 of the 9 archives support ETDMS, but all new sites are encouraged to support this richer standard.

Tools and ETD-db extensions

To support participation in the Union Archive, an extension to the ETD-db ETD management software [Atkins, 2001] was created. This extension is a drop-in module to retrospectively add OAI support to an archive that uses any version of the ETD-db software. The current version of ETD-db has OAI support built in.

Similar tools are available for the E-Prints software [OpCit, 2002] and various toolkits are made available at the OAI website to ease the development of OAI interfaces to custom-built digital libraries or digital libraries with no OAI support.

Rights Management

Currently, only freely accessible metadata is harvested from ETD archives. Thus, when the unmodified metadata is republished, no intellectual property rights are violated. Also, the Union Archive retains all original record identifiers so that they may be traced to their respective sources, and to enable duplicate detection.

In terms of digital objects, the OAI can allow linking from metadata to HTML pages within a digital library, rather than directly to the data. The net effect of this is that the source archives can enforce rights management for their resources. The Union Archive uses this mechanism to avoid dealing with the myriad of rights policies in use among NDLTD members.

Configuration

The configurable information for the Union Archive is stored in an XML file, defining the archives to harvest metadata from as well as parameters to control the harvesting algorithm and the Union Archive in general. Figure 2 shows an extract from the configuration file and Table 3 lists the semantics of its fields.

```
<unionconfig>

  <database>DBI:mysql:etdunion</database>
  <dbusername>etdunion</dbusername>
  <dbpassword></dbpassword>
  <table>unioncat</table>

  <archive>
    <identifier>VTETD</identifier>
    <url>http://oai.dlib.vt.edu/cgi-bin/VTETD/VTETD.pl</url>
    <metadataPrefix>oai_dc</metadataPrefix>
    <interval>0.25</interval>
    <interrequestgap>15</interrequestgap>
  </archive>

  <archive>
    <identifier>HUBerlin</identifier>
    <url>http://dochost.rz.hu-berlin.de/OAI-script</url>
    <metadataPrefix>oai_dc</metadataPrefix>
    <interval>1</interval>
    <set>HUBerlin:dissertationen</set>
  </archive>

</unionconfig>
```

Figure 2. Extract from Union Archive configuration

Field	**Semantics**
database	Name of the SQL database
dbusername	User name to supply when connecting to the database
dbpassword	Password to supply when connecting to the database
table	Prefix for database table names associated with this instance of the component
archive/	Container for definition of a single archive
identifier	Unique label for the archive
url	*baseURL* of the archive
metadataPrefix	*metadataPrefix* for metadata format to harvest
interval	Periodicity of harvesting (in days)
interrequestgap	Number of seconds to delay between consecutive requests to the archive
set	*setSpec* of set to harvest (if omitted, harvest entire archive)

Table 3. Semantics of fields in configuration file

SERVICE PROVIDERS

Currently, the primary service providers are the NDLTD Union Catalog [VTLS, 2002] and the experimental ODL-based Union Catalog [Suleman and Fox, 2001]. Both of these systems harvest metadata from the Union Archive and provide searching and browsing services [Suleman, et al., 2001]

FUTURE WORK

As the user-level services become more popular, the Union Archive is becoming an important component in ETD discovery solutions.

Research is currently underway to build more reliable archives with better performance by replicating metadata over high-speed network connections – this is related to the Internet2 Distributed Storage Initiative. Within the ODL project, higher-level digital library services are being developed to facilitate alternative forms of discovery (e.g., by recommendation).

Ultimately, the Union Archive is one part of a larger initiative to provide high quality and high availability digital library services to students and researchers with minimum effort from contributing sites.

To get closer to this goal, all NDLTD members are encouraged to contribute to the Union Archive. More information can be found on the NDLTD website (http://www.ndltd.org/union.html).

ACKNOWLEDGEMENTS

Thanks are given for the support of NSF through its grants: IIS-9986089, IIS-0002935, IIS-0080748, IIS-0086227, DUE-0121679, DUE-0121741, and DUE-0136690.

REFERENCES

Atkins, Anthony. (2001) Resources for Developers of ETD databases. Website http://scholar.lib.vt.edu/ETD-db/developer/

Atkins, Anthony, Edward A. Fox, Robert France and Hussein Suleman (editors). (2001) ETD-ms: an Interoperability Metadata Standard for Electronic Theses and Dissertations -- version 1.00. Available http://www.ndltd.org/standards/metadata/ETD-ms-v1.00.html.

Box, Don, David Ehnebuske, Gopal Kakivaya, Andrew Layman, Noah Mendelsohn, Henrik Frystyk Nielsen, Satish Thatte, and Dave Winer. (2000) Simple Object Access Protocol (SOAP) 1.1, W3C Note, 08 May 2000. Available http://www.w3.org/TR/SOAP/.

Fallside, David C. (editor). (2001) XML Schema, Part 0, Part 1, and Part 2. W3C Recommendation, 2 May 2001. Available http://www.w3.org/TR/xmlschema-0/, http://www.w3.org/TR/xmlschema-1/, and http://www.w3.org/TR/xmlschema-2/.

Fox, Edward A., John L. Eaton, Gail McMillan, Neill A. Kipp, Laura Weiss, Emilio Arce, and Scott Guyer. (1996) National Digital Library of Theses and Dissertations: A Scalable and Sustainable Approach to Unlock University Resources, *D-Lib Magazine*, September 1996. Available http://www.dlib.org/dlib/september96/theses/09fox.html.

Fox, Edward A., Brian DeVane, John L. Eaton, Neill A. Kipp, Paul Mather, Tim McGonigle, Gail McMillan, and William Schweiker. (1997) Networked Digital Library of Theses and Dissertations: An International Effort Unlocking University Resources, *D-Lib Magazine*, September 1997. Available http://www.dlib.org/dlib/september97/theses/09fox.html.

Lagoze, C., and Herbert Van de Sompel. (2001) The Open Archives Initiative: Building a low-barrier interoperability framework, in *Proceedings of JCDL 2001*, Roanoke VA, June 2001, ACM Press, pp. 54-62.

Lagoze., C., and J. R. Davis. (1995) Dienst – An Architecture for Distributed Document Libraries, in *Communications of the ACM*, 38(4), April 1995, p. 47.

Leiner, B. M. (1998) The NCSTRL Approach to Open Architecture, in *D-Lib Magazine,* December 1998. Available http://www.dlib.org/dlib/december98/leiner/12leiner.html.

OAI. (2002) Open Archive Initiative. Website http://www.openarchive.org/.

OpCit. (2002) E-Prints. Website http://www.eprints.org/.

Powell, J., and E. A. Fox. (1998) Multilingual Federated Searching Across Heterogeneous Collections, in *D-Lib Magazine*, 4(8), September 1998. Available http://www.dlib.org/dlib/september98/powell/09powell.html.

Suleman, H. (2001) Enforcing Interoperability with the Open Archives Initiative Repository Explorer, in *Proceedings of JCDL 2001*, Roanoke, VA, June 2001, ACM Press, pp. 63-64.

Suleman, H., and E. A. Fox. (2001) A Framework for Building Open Digital Libraries, in *D-Lib Magazine* 7(12), December 2001. Available http://www.dlib.org/dlib/december01/suleman/12suleman.html.

Suleman, H. and E. A. Fox. (2002) Beyond Harvesting: Digital Library Components as OAI Extensions, Technical Report, Virginia Tech NCSTRL Collection, January 2002.

Suleman, H., Atkins, A., Gonçalves, M. A., France, R. K., Fox, E. A., Chachra, V., Crowder, M., and J. Young. (2001) Networked Digital Library of Theses and Dissertations: Bridging the Gaps for Global Access - Part 1: Mission and Progress, and Part 2: Services and Research, in *D-Lib Magazine,* 7(9), September 2001. Available http://www.dlib.org/dlib/september01/suleman/09suleman-pt1.html and http://www.dlib.org/dlib/september01/suleman/09suleman-pt2.html.

Van de Sompel, Herbert, and Carl Lagoze. (2000) The Santa Fe Convention of the Open Archives Initiative, in *D-Lib Magazine*, 6(2), February 2000. Available http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html.

Van de Sompel, Herbert, and Carl Lagoze. (2001) The Open Archives Initiative Protocol for Metadata Harvesting. Open Archives Initiative, 2001. Available http://www.openarchives.org/OAI_protocol/openarchivesprotocol.html.

VTLS. (2002) NTLTD Union Catalog. Website http://www.vtls.com/ndltd/.