ETD 2012

The NDLTD Union Catalog: Issues at a Global Scale

Hussein Suleman
University of Cape Town
hussein@cs.uct.ac.za

Format: paper

Length of presentation: 30 minutes

Abstract

The NDLTD Union Catalog is an international collection of ETD metadata that is harvested from various institutional, regional and cross-institutional collections. The Union Catalog has grown substantially in the 10 years since its launch and now contains almost 2 million records. However, various issues have surfaced during the maintenance of the Union Catalog and its downstream service providers. For example, at this scale, the well-known best practice of the OAI-PMH to restrict the size of a response to 100 records or 1MB has a severe impact on harvesting time. This paper describes this problem and other issues that are relevant to the Union Catalog and similar projects. For each such issue, solutions are discussed. Together these present a set of guidelines not only for large union catalogues but also for the design of large digital library collections in general.

Introduction

The NDLTD Union Catalog Project began in 2001, with the primary objective being to gather metadata from ETD repositories around the world into a central collection where users could use discovery tools to find ETDs across collections. While it is possible to use general-purpose Web search engines, an ETD-specific suite of services could provide higher quality results of a very specific nature; and this could provide users with specialised services to navigate through a homogenous collection. The Union Catalog aimed also at popularising ETDs by creating a showcase where users could see what was being produced in other institutions, particularly those who are members of NDLTD. Membership of NDLTD was and is not, however, a prerequisite to be part of the Union Catalog.

The first prototype of the Union Catalog service was produced as a demonstrator for the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) (Lagoze, et al, 2002). There were only 14 sites, with a total of less than 50000 records, using a harvester and OAI-PMH data provider from the ODL suite of digital library system components (Suleman and Fox, 2001). This service was developed at and run by the Virginia Tech Digital Libraries Research Laboratory until the experimental phase ended (Suleman and Fox, 2002).

Eventually, NDLTD migrated the service to OCLC, where a production service was run for approximately 10 years. The collection grew to well beyond a million records, with the addition of ETD metadata extracted from WorldCat, OCLC's international collaborative metadata generation system. WorldCat contains records in the MARC format and ETD records were extracted if the record specified that it was a thesis or dissertation and included a URL to a full text electronic representation. A custom-built harvester was used in conjunction with the XTCat data provider software produced and maintained by OCLC. Initially, the number of records rose

rapidly, approximately doubling each year for the first 5 years because of new sites joining the project – in the last 5 years, this growth has slowed down.

When the service moved to OCLC, NDLTD also decided to divide the Union Catalog into a metadata-only Union Archive and downstream service providers who would be responsible for search and browse services. VTLS and Scirus offered to provide the latter services, while OCLC provided the former archive service.

In 2011, OCLC transferred the Union Archive to the University of Cape Town's Digital Libraries Laboratory, where it currently resides. The current system uses the ETDPortal software tool that was developed for the South African National ETD Project (Webley, et al, 2011). Data was migrated from OCLC and, after some months of tweaking, the system transitioned into limited production mode in 2011. As of early 2012, the system has reached a stable state where automated harvesting of remote sites takes place daily.

In the process of making the transition and during the early operation of the site, a number of issues surfaced that were unique to the problem of running a large repository. Some of these have been resolved or addressed adequately while others are still unresolved. The rest of this paper presents details of the Union Archive system and the various issues that were encountered in its development and continued operation.

Architecture of the Union Catalog

The Union Archive collects metadata on a regular basis from each remote site using the OAI Protocol for Metadata Harvesting (Lagoze, et al, 2002). The OAI-PMH is a Web-based protocol to transfer metadata from one machine to another in incremental mode such that only changes are transmitted each time a harvesting operation is performed.

Metadata records are harvested from every site in the Dublin Core format that is mandatory for the OAI protocol. Records also are harvested in the ETDMS metadata format (Hickey, et al, 2010), which was designed specifically for ETDs, if remote sites can provide this metadata.

The architecture of the system is depicted in Figure 1.

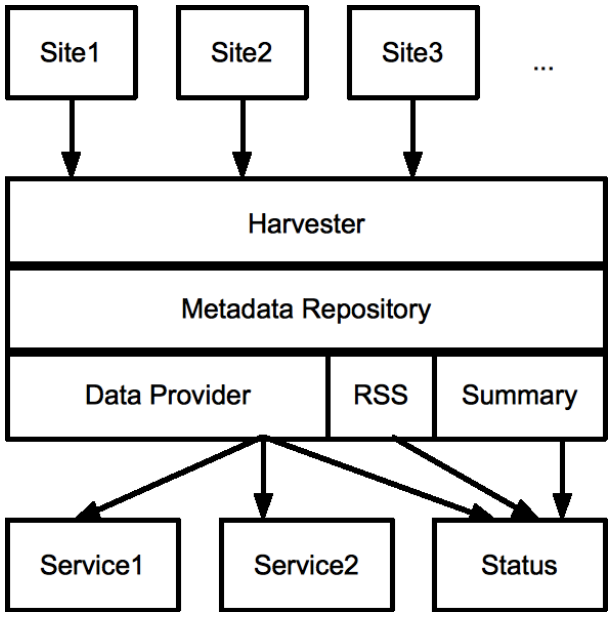

Figure 1: Architecture of Union Catalog / Archive

The central metadata repository, in a MySQL database, is shared by 4 components: the metadata harvester, OAI-PMH data provider, RSS feed generator and summary generator. The metadata harvester collects metadata on a regular basis from remote sites. The data provider serves this metadata to any external service providers who wish to obtain a copy of the metadata to enable provision of services. The RSS feed is an awareness service that provides a snapshot listing of 5 of the most recent records ingested into the system. The summary generator is a custom service that generates a listing of all sites and their numbers of records. There also is a simple front-end status interface that lists recent activity and statistics for the service. This is shown in Figure 2.

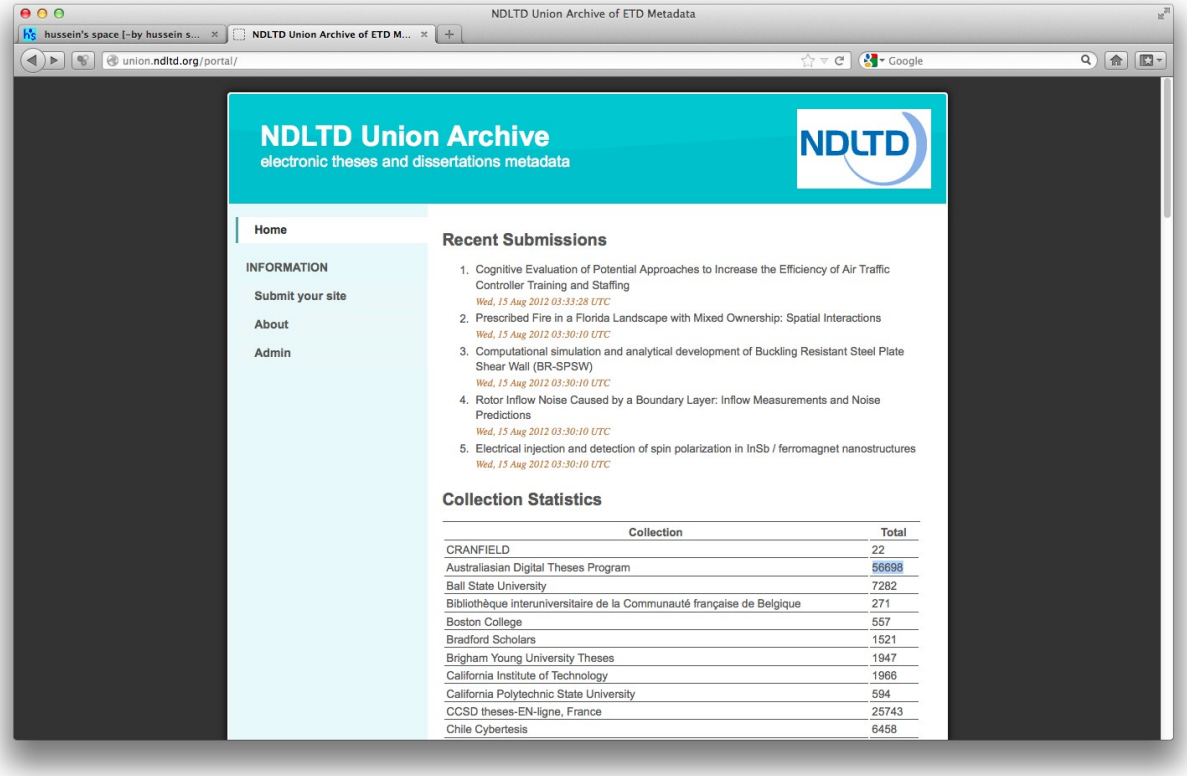

Figure 2: Union Archive status page

The Union Archive website provides only a thin user interface layer to integrate the RSS and summary feeds into a status page. An authenticated administrator is also able to access an administrative interface, where repositories can be monitored and modified and harvesting schedules changed. This is depicted in Figure 3.

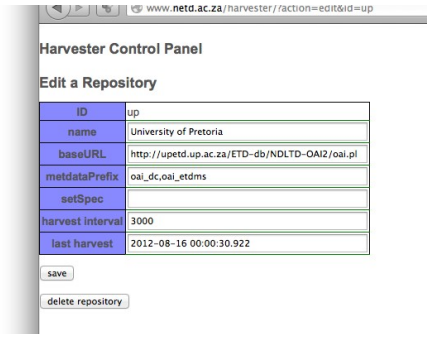

Figure 3: Repository management interface

While the central metadata repository does not provide any discovery services, these are provided by external service providers. VTLS provides a search and faceted browsing service while Scirus provides a full-text search service. Both of these service are linked directly from the NDLTD website – it is not intended that any end users should ever visit the Union Archive status page. Any additional service providers may obtain the full metadata feed from the Union Archive via its OAI-PMH data provider. The details are provider on the Union Archive status page as well as the NDLTD website.

Status: August 2012

As of 15 August 2012, there are 1982309 records in the Union Archive, from a total of 132 different sites and sub-collections.

The top contributors include OCLC (1228554), the IBICT Brazilian national collection (142252), Libraries and Archives Canada (121677) and the Australasian Digital Theses Program (56698). All of these are themselves collections from multiple sites. The smaller contributors are individual institutions and, in a few rare cases, divisions within institutions.

Harvesting is triggered once a day at 3am UTC for all sites.

Analysis: Problems and Fixes

The following is a list of the problems encountered during the transition and how they were resolved.

- When records were migrated from OCLC, identifiers were not unique to each record. This is a problem for the data provider as a request for a record will not map to a distinct entry in its database. The OAI-PMH recommends the use of domain names within a URI syntax to ensure global uniqueness, with the domain name registry serving as the global registry, thus avoiding duplication of effort. A typical identifier would then be: *oai:www.someuniversity.edu:recordid*. Many ETD sites do not follow this convention, so all incoming records had their identifiers modified to conform to a global scheme, using the site ID as a prefix for the recordid and placing all records within the union.ndltd.org domain.
- Each site was assigned a unique ID based on the abbreviation of the institution but this identifier space did not work for long as there are different institutions internationally with the same abbreviated name (e.g., UBC). Newer sites are now identified by a fully qualified domain name.

- Some sites are no longer active or have moved without notification but there are records within the Union Archive. These records have been maintained. In future, a detailed investigation of all sites should help to resolve problems with server relocation.
- The scale of some sites, especially national sites, has caused unforeseen problems. The nature of OAI-PMH requires that sites are able to generate sets of records at arbitrary positions within the full list of records held by a repository. A repository with a large collection of metadata needs a fast lookup index to extract these records – a database row scan does not work as the time taken to extract each subsequent batch of records is greater than that for the previous batch. In practice, this means that some sites timed out after a large number of records had been tranferred. Two solutions were applied for this problem. Firstly, the source repository administrator was contacted to request an improvement in the indexing of their records. Secondly, the Union Archive harvester was modified to include different per-repository timeout values.
- The scale of the Union Archive, with almost 2 million records, is itself a problem for external service providers who try to harvest the metadata. OAI-PMH data providers have historically used batches of approximately 100 records to transfer metadata. This proved too slow for large-scale harvesting internationally because of latencies in TCP/IP connections. The data provider component was modified to provide 1000 records in each batch as a workable fix. In future, the OAI-PMH may need to be updated to better meet the needs of repositories attempting to perform large data transfers.
- The ETDPortal sofware used for the Union Archive also contained a simple Lucene-based search service that was initially turned off because of memory requirements for indexing of large collections. The software has since been modified to index OAI-PMH-based batches of records, thus placing an upper bound on memory requirements. The search service will be reactivated in future because it is now a viable option.
- Encoding of data has always been a problem, in the harvested records as well as the names of repositories and other meta-information. The software was carefully checked and updated to ensure that Unicode and UTF-8 were used throughout all workflows, thus ensuring that no data corruption would take place provided that the incoming data is itself valid.
- Errors in incoming records are a recurring problem. Many sites do not validate or clean the XML in data provider responses. Thus, characters such as the quotation marks that are cut-and-pasted from Microsoft Word create invalid XML. To minimise the impact of such problems, OAI-PMH responses are split into text records before being parsed as XML. Each record is parsed individually and erroneous records are logged and not included in the Union Archive. In future, an automated message may be sent to the source archive administrator to inform them of the error.

The following is a list of unresolved issues in the Union Archive – problems that either have no solution or where solutions have not yet been implemented:

- Some collections are harvested directly but also indirectly via national collections and the OCLC extracted collection. In those cases, the metadata may also have been generated from different sources (e.g., the library catalogue and the graduate school repository) so deduplication is non-trivial. Currently, the collection does not remove duplicates but this must be addressed in future.

- There are rudimentary site maintenance tools for harvesting, but a more sophisticated curation tool is necessary to track down individual records, view subsets, etc. in order to answer queries and make changes manually when needed. Any changes should also result in updates to the datestamp so that the changes are propagated to downstream service providers.
- There is a need for greater automation in various aspects of the system in order to make it more scalable. For example, a repository administrator should be able to test and submit their site's baseURL or make changes to site information without direct human intervention. Errors should be reported to source repository administrators automatically.
- While ETDMS has been widely advertised, most sites still do not provide metadata in this format. The few sites that do provide ETDMS metadata use the older pre-2011 version of the standard instead of v1.1. The Union Archive could automatically translate ETDMS records and generate partial ETDMS records from DC records on-demand.
- While the design philosophy of the Union Catalog project has always been to support external service providers, there is an inevitable delay before they update their indices. Some end-users have complained that service provider indices are substantially out-of-date with the Union Archive. A basic search/browse service that is collocated with the harvester could help to address the needs of such users.
- Metadata push may need to be considered as an alternative to harvesting. Whenever new records arrive at the Union Archive, the service could contact key service providers with updates or a notification that updates are available. This could be similar to the RSS trackback feature.
- There are many problems with records and it is not clear who should take responsibility for correction of these errors. Where a repository manager responds quickly, it is usually preferable that the errors are fixed at the source. However, where the source repository or source repository administrator is no longer responsive, it may be necessary to make corrections in the Union Archive directly. Thus, a hybrid solution may be most effective for record-level quality control and management.

Conclusions

The NDLTD Union Catalog is a well-established production repository and service system. However, as the scale of the repository changes over time, new problems have emerged. Some of these problems have been addressed by architectural changes while others are as yet unresolved. Most importantly, the scale of operations suggest that digital library systems need to be designed differently to deal with large data sets, where efficiency and automation are more important than user interaction. Even established standards like OAI-PMH are desperately in need of an update to enable efficient large data transfers. While the NDLTD Union Catalog is fully operational at the current scale, future design efforts need to consider further orders of magnitude for this and related systems.

References

Hickey, Thom, Ana Pavani and Hussein Suleman (2010) ETD-MS v1.1: an Interoperability Metadata Standard for Electronic Theses and Dissertations, NDLTD. Available http://www.ndltd.org/standards/metadata/etd-ms-v1.1.html

Lagoze, Carl, Herbert Van de Sompel, Michael Nelson, and Simeon Warner (2002), The Open Archives Initiative Protocol for Metadata Harvesting Version 2.0, Open Archives Initiative, June 2002. Available http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

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). Towards Universal Accessibility of ETDs: Building the NDLTD Union Archive, in Proceedings of Fifth International Symposium on Electronic Theses and Dissertations (ETD2002), Provo, Utah, USA, 30 May-1 June 2002. Available http://www.husseinsspace.com/research/publications/etd_2002_paper_union.pdf

Webley, L., T. Chipeperekwa and H. Suleman (2011), Creating a National Electronic Thesis and Dissertation Portal in South Africa, in Proceedings of 14th International Symposium on Electronic Theses and Dissertations (ETD) 2011, Cape Town, South Africa, 13-15 September 2011. Available http://pubs.cs.uct.ac.za/archive/00000748/