Creating a National Electronic Thesis and Dissertation Portal in South Africa

Lawrence Webley
University of Cape Town
lwebley@cs.uct.ac.za

Tatenda Chipeperekwa
University of Cape Town
chippytdm@gmail.com

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

ABSTRACT

A national Electronic Theses and Dissertations (ETD) portal has been developed in South Africa to provide access to a country-specific collection of ETDs and, more importantly, to coordinate, manage, monitor and support the development of ETD programmes at the various universities. This portal required the development of a custom software solution, using a multi-tiered simple architecture of complex components. It is argued in this paper that this tiered architecture, tightly integrated into a commonly-used application/operating system framework, is a good approach to develop such central repository architectures to interconnect into the larger repository ecosystems of NDLTD and similar organisations.

Keywords (Required)

ETD, OAI-PMH, repository, harvester, portal, national archive, metadata.

INTRODUCTION

Since the inception of the Networked Digital Library of Theses and Dissertations (NDLTD), many institutions around South Africa and the world have developed operational Electronic Theses and Dissertations (ETDs) repositories, collectively amassing more than 1.5 million ETDs as of early 2010 [7].

Unfortunately, although ETD digital library (DL) collections are now widely available online, there are few cross-archive services specific to ETDs. NDLTD and its collaborators manage a suite of international services but these are not easily replicable for more localised communities, such as the universities within a single country or discipline.

ETD services in South Africa have reached the stage where almost every institution has a suitable repository and associated workflow processes in place. It has been recognised that completeness, sustainability and expansion are best served by a national archive, one aspect of which brings together the metadata from distributed source archives using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) into a central portal. However, at the onset of the project, there was no readily-available open source ETD portal management software to meet this objective.

This paper reports on an attempt to create such a system, composed of 3 separable layers to support expansion and assist in future scalability. Initially metadata records for the ETDs are harvested and verified from a selection of remote repositories. These records then are stored in a local database that is accessible via the OAI-PMH. Finally an instance of the Lucene [6] open source search engine is used to index the repository and provide a human-usable Web interface that can be used to search through the metadata. This three-part system allows a user to navigate through multiple different institutions' ETDs via keywords and browsing operations within a single portal. The metadata that the user discovers contains a link to the original repository and, indirectly, the original documents. On a global level, the system serves as a central repository for other, potentially larger, repositories (e.g., that of NDLTD) to harvest from, allowing them to simplify the harvesting of a region/country.

PREVIOUS WORK

Initial efforts at providing a unified search system for NDLTD involved the creation of a federated search system [8]. This system stored the search language and syntax of all the participating sites. Upon receiving a user query, the system reformulated it such that it could be used to search on each participant site. The federated search system would then return

the results from each site's search, but did not merge them. There were many problems with this approach, however. Primarily, it relied heavily on each participating member site. If a site was unreachable or changed its search interface, their results could not be obtained. If a site took a long time to reply, the search results would be delayed. Since each site provided its own form of HTML results page, there was no simple way to merge the search results. Furthermore, adding a new remote site to the system was a non-trivial process, often involving the specification of a new search syntax and/or language.

These problems were partly addressed by the emergence of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [2]. The OAI-PMH is a HTTP based client-server protocol that provides an application-independent interoperability framework based on metadata harvesting. The protocol is designed to facilitate incremental transfers of metadata in a simple and general way. Primarily, the protocol is designed for the transfer of metadata as opposed to the remote searching of metadata. This results in less interaction between remote sites in a distributed discovery environment. In practice this means that if a DL provides an OAI-PMH interface to their ETD repository, a service provider can access their ETD metadata in a generalised and uniform manner, using less bandwidth and with greater stability of service provision.

If a service provider wishes to use the ETDs contained in a DL, it must first harvest those records using the OAI-PMH and then store them locally before using them. At a more general level, the OAI-PMH allows for a system of interconnected components, each of which is essentially a digital library (DL). This mode of operation forms the basis of the NDLTD Union Catalog [1] and is data transfer mechanism used in the South African National Electronic Thesis and Dissertation (NETD) Portal [3].

Similar efforts have been undertaken in some other countries e.g., Brazil [9], United Kingdom [10]. However, a common problem is the lack of reusable systems. While there are popular repository tools such as Dspace and EPrints, there are few popular portal creation tools. Thus, a further aim of the South African NETD project was to create a reusable set of tools.

SYSTEM ARCHITECTURE

All components were written in Java, using Java Servlet Web technology hosted on Tomcat servers within an Ubuntu Linux platform.

The system architecture is based on a complex component model. Each step in the process of providing a centralised ETD discovery portal forms a single and separable component that can operate independently. There are multiple reasons for this approach. Firstly, the scalability of the service as a whole is far greater with a componentised solution. For example, if the data collection is small, all components can run on a single server. However, as the data collection grows, each component can be moved off onto its own server. Additionally, with a componentised solution, if the solution needs to be updated or changed to suit a particular environment, the user need only replace the offending component of the solution, as opposed to rewriting the entire system. A final important feature of a componentised solution is that if any part of the system fails, the system as a whole does not.

The system has been split into three distinct components, as described in Figure 1: the Harvester, the Repository and the Web Portal. The Harvester retrieves metadata from a set of ETD repositories. The Repository provides a set of machine access points to the metadata harvested by the Harvester. The Web Portal then uses the metadata and other services from the Repository in order to provide a unified discovery interface over all the repositories harvested. Prior attempts at designing such an architecture (e.g., ODL, OpenDLib) have included service-based components with finer granularity but this invariably leads to increased complexity at the component management level. This system attempts to strike a better balance between component size and functionality on the one hand and inter-component management on the other hand.

A further consequence of the complex component approach is that each component potentially consumes multiple machines interfaces and provides multiple machine and end-user interfaces. The Harvester consumes OAI-PMH interfaces from multiple source repositories and provides a database with a machine interface to the Repository in addition to an end-user management interface. The Repository consumes the database interface and provides multiple machine interfaces (OAI-PMH, RSS, Summary) to the Web Portal and other high level services. Finally, the Web Portal consumes all machine interfaces provided by the Repository and provides an end-user interface for resource discovery. This many-to-many component interaction is arguably appropriate for well-defined and repeatable systems such as national metadata portals that can be used in different countries and for different tasks.

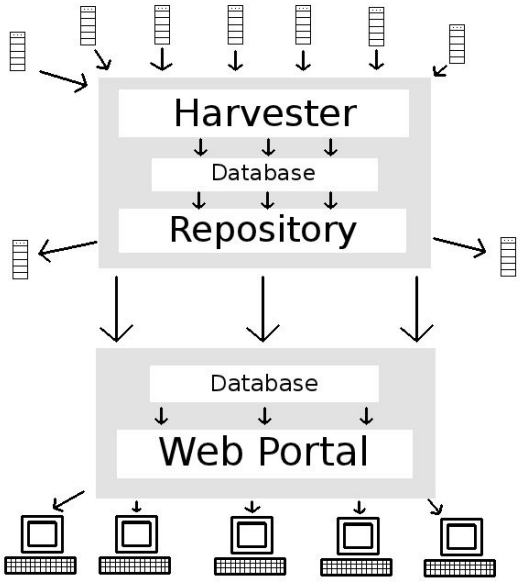


Figure 1. Diagram of the System Architecture

Use of the OAI-PMH

Much of this solution relies on the use of a standardised metadata transfer protocol, namely the OAI-PMH. In order for the Harvester to easily retrieve metadata from various institutional ETD repositories, it must be able to use a single predetermined protocol that all of the repositories support. The widespread adoption of the OAI-PMH by repository tools makes this possible. Without this standardised protocol, the harvester would need to support an array of different protocols in order to effectively harvest multiple repositories. This would complicate its construction greatly. The Repository provides a public OAI-PMH interface, so that its records can be accessed in a standardised fashion by others. The Web Portal uses the OAI-PMH interface offered by the Repository to access the Repository metadata.

Harvester

The Harvester component of the system is responsible for collecting metadata records from a set of remote repositories. The Harvester initially performs a full harvest of each remote repository, meaning that it will retrieve all metadata records stored in the repository. Thereafter, the Harvester will only retrieve records that have been added since the last harvest date, thus effecting incremental harvesting. This is done in order to minimize the transfer of data between the repositories and the Harvester. All metadata records retrieved by the Harvester will be stored directly into the shared Harvester/Repository data store. This is implementation-specific, but the system defaults to using MySQL as the database layer.

The Harvester performs simple validation checks on records as they are inserted into the database. This initial checking is required in order to ensure that no malformed records are passed to the Repository via the shared data store. The Harvester is typically run at set time intervals, which could be as often as hourly, but is more likely to be daily. ETDs are not a rapidly changing resource, and a propagation delay of a day from an institution's repository to the national Repository is generally considered to be acceptable.

The Harvester also includes an end-user interface to manage the list of remote repositories and the harvesting process. Using this interface, an administrator is able to add, modify or delete repositories and monitor the status of the harvesting of individual repositories.

Repository

The Repository data store defaults to using MySQL for its SQL-compliant database. The Repository retrieves records from the data store and makes them accessible via the OAI-PMH protocol. The actual data store form and structure is transparent to the user of the OAI-PMH machine interface. This OAI-PMH interface is important for scalability because the national Repository might be harvested by another Repository higher up the hierarchy, such as a continental or international Repository. Such a hierarchical system of repositories and harvesters at an international level will result in minimum

duplication of records and minimal network traffic.

The Repository also provides an RSS feed of the latest harvested records using the RSS 2.0 standard.

Finally, a custom Web Service is provided to list a summary of repositories and the number of records in each. This service is meant to be consumed by a portal to provide an overview of the collection.

Web Portal

The purpose of the Web Portal is to provide a human usable interface to the metadata archive. The Web Portal queries the OAI-PMH interface of the Repository in order to access all the records it maintains. It harvests from the Repository into its own local database. Once the metadata records are stored in its local database it can process the records using Lucene [6] for search indices and MySQL for browse indices. Currently Lucene performs a full re-index of the repository periodically as opposed to an incremental update. In future versions of the portal, this will be converted to an incremental re-index. Lucene is integrated into the Web front-end, which provides the user with the ability to search by keyword and by various categories, such as title, author, subject and institution.

The Web Portal offers a Web-based interface through which users can browse and search through the indexed metadata records, and find links to the original documents on the source repositories. Although the primary focus of the Web Portal is searching and browsing, it also encapsulates and makes all standard Web Services (OAI-PMH, RSS) and end-user services available to end users in a single location.

Operating System Integration

All components store configuration information in the standard Ubuntu Linux configuration directory (/etc); log files in the standard location (/var/log); and Web applications in the standard root Web server. This high level of conformance with one of the most popular Linux distributions will make it possible to package the application for software repository-based installation in future. The system's complex components also map well to the granularity of software components already defined in Ubuntu Linux, where an advanced dependency management regime will ensure that installation of the Web Portal will automatically trigger installation of all pre-requisites if necessary (Repository/Harvester, MySQL, Java, Tomcat, Apache HTTPD).

CASE STUDY: IMPLEMENTING THE NETD SYSTEM

Development of the open source South African NETD software began in 2009 and a fully functional implementation was completed in 2011. The software was instantiated at the National Research Foundation to create a single national portal for ETDs (www.netd.ac.za).

Eleven institutional sub-collections were included in the portal at the time of writing, each of which had implemented an OAI-PMH interface for their metadata repositories. Prior to this development, there existed no central point of access for South African ETDs (only non-electronic documents via NEXUS).

The harvester was initially configured with only support for oai_dc metadata, for which it performs XML validity checking per record harvested. Additional metadata formats are planned to be supported in the future and the harvester allows for this multiplicity of metadata already. A MySQL database was created to store the harvested metadata records for the Repository/Harvester.

The Web Portal also uses MySQL for its data store. The front page of the Web Portal was developed as a simple servlet specific to NETD to bring together the various services and impose an appropriate CSS theme on all pages.

In this initial implementation, all three components are run from a single server, but as the scope of the project grows, the components might be separated off onto different servers.

Figure 2 shows the Web interface of the system while Figure 3 shows the Web interface of the administrative part of the harvester.

AWStats was used to analyse the server log files of the portal after approximately 6 months of operation to monitor usage and the following trends were observed:

- There were usually 500-600 unique visitors to the site each month, with a total of approximately 43000 hits in the period from 1 January to 10 June 2011.

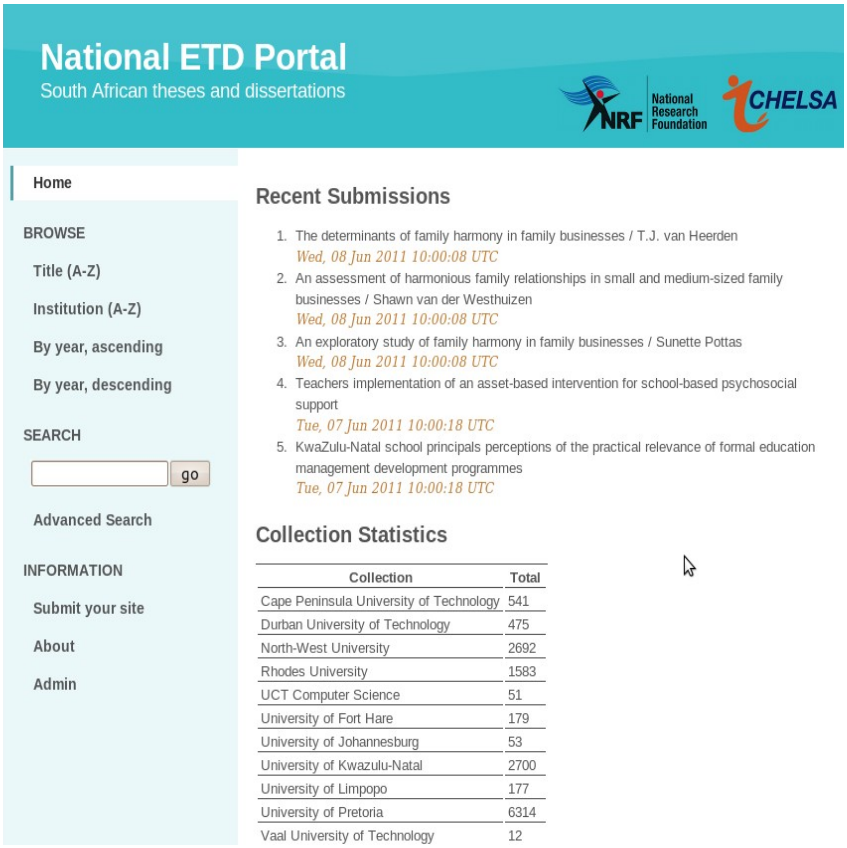
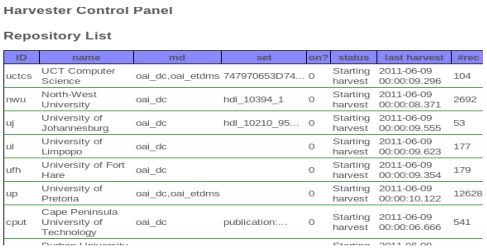


Figure 2. Web discovery interface



1Figure 3. Harvester administrative interface

- The site was accessed more during the hours of 9am-4pm (local time). This implies that most hits were from South Africa and countries in similar time zones, such as most of Europe.
- A large percentage (approximately 30%) of the hits came from South Africa.
- The site was indexed by 37 different Web crawlers.
- More than 80% of hits were to the main portal interface. The remaining hits were to the Web service interfaces (OAI-PMH, RSS, etc.)
- Most visitors to the site coming from another site (rather than a search engine) were in fact arriving from the NDLTD site (1118 hits of 1441).

-
-

CONCLUSION

The framework proposed in this paper represents a reusable, componentised system for the creation of a central repository and Web Portal. It can be easily generalized for use in other DL systems, and offers a balance between manageability and scalability. Due to the component nature of the system, it is customisable, and can be made to be extended and incorporated into other systems. Ongoing work is investigating the use of a Content Management System to replace the simple servlet front-end, for example.

The extensive use of the OAI-PMH in this system is in line with the concept of an interconnected, hierarchical global ETD network. The South African part of this framework can easily be harvested by an Africa-wide segment of the framework, which could subsequently be harvested by a global service. The advantage of this is that management is decentralised and occurs closer to the responsible entities, with automatic metadata propagation to higher levels.

The implementation of this system in the form of the South African NETD project represents a proof of concept of the proposed framework. All code produced is being released under open source licences for others to reuse, adapt and extend.

ACKNOWLEDGMENTS

We would like to thank Alexander Van Olst for his work on the harvester portion of the software. Thanks also are due to: the South African National Research Foundation (NRF), who funded the project; the NETD Steering Committee and the Committee of Higher Education Librarians of South Africa (CHELSA), who provided feedback during the development. The software development has as its home the Digital Libraries Laboratory at the University of Cape Town.

REFERENCES

1. Suleman, H. and Fox, E.A. (2003), “Leveraging OAI harvesting to disseminate theses”, *Library HiTech*, Vol. 21 No. 2, pp. 219-27.
2. Lagoze, C., Van de Sompel, H., Nelson, M. and Warner, S. 2002. The Open Archives Initiative Protocol for Metadata Harvesting, Open Archives Initiative. Available http://www.openarchives.org/OAI/openarchivesprotocol.html
3. National Electronic Thesis and Dissertation Portal. Available http://www.netd.ac.za/
4. South African National Research Foundation. Available http://www.nrf.ac.za
5. Digital Libraries Laboratory at University of Cape Town. Available http://www.cs.uct.ac.za/research/dll
6. Apache Lucene. Available http://lucene.apache.org/
7. Fox, E. A., Hickey, T., Chachra, V., and Erb, C. (2010) NDLTD Union Catalog Surpasses One Million Electronic Theses and Dissertations, Press Release, NDLTD. Available http://www.ndltd.org/find/ndltd-union-catalog-surpasses-one-million-electronic-theses-and-dissertations/
8. Powell, J., and Fox, E. A. (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
9. Marcones, C. H., Sayão, L. F., Triska, R., and Pavani, A. M. B. (2002) Brazilian electronic theses and dissertations consortium, in Proceedings of Fifth International Symposium on Electronic Theses and Dissertations, 30 May – 2 June 2002, Provo, Utah, USA. Available http://docs.ndltd.org:8081/dspace/handle/2340/206
10. Copeland, S., and Bevan, S. (2005) 'EThOS': an 'Electronic Theses Online Service' for the UK, in Proceedings of Eigth International Symposium on Electronic Theses and Dissertations, 29 September 2005, Sydney, Australia. Available http://docs.ndltd.org:8081/dspace/handle/2340/255