Implementing a Low-Cost South African National Heritage Portal built from Metadata Aggregation

A literature Review

Alex Priscu†
Department of Computer Science
University of Cape Town
Cape Town Western Province South Africa
prsale003@myuct.ac.za

ABSTRACT

South African memory institutions have been struggling to build digital heritage archives. Most digital libraries are separate and unlinked, creating the challenge of searching through numerous archives for resources. This discoverability challenge was resolved by Organisational Architecture, where a centralised repository collects the relevant cultural heritage resource metadata from remote archives, known as metadata aggregation. Web portals use the aggregation of metadata as input using a metadata harvesting protocol such as the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) to allow for interoperability among repositories. Efforts have been made to create a national heritage portal, where everyone can search through a unified collection based specifically on heritage documents. This paper reviews local and international metadata portals that resulted in either a high-cost design or being built for a different domain. Technologies such as the OAI-PMH and AJAX are discussed and compared in various metadata aggregation systems providing cross-archive search and browse services to end-users.

CCS CONCEPTS

• Information Systems • Information systems applications • Digital libraries and archives

KEYWORDS

Metadata, Aggregation, Portal, Heritage, Archive, OAI-PMH.

1. INTRODUCTION

For years, memory institutions in South Africa have been trying to create digital archives of heritage content with mixed success levels. Online digital repository systems, digital libraries or digital archives are a popular framework for the storage of various forms of digital content [11]. Academic documents and heritage collections are commonly archived in such repositories for the purpose of discovery and preservation [12]. These online systems have the advantage of general accessibility during the current period of increasing Web browser use [11]. These systems store and manage digital items and provide users with access to them by providing search and browse services through a Web-based interface [12].

Early digital archives, such as the Bleek and Lloyd Collection [16] and early publications scanned by Digital Imaging South Africa (DISA) [18] have been created with the help of external funding. These digital archives, however, are all independent and not linked. Cross-archive discovery is a problem that affects digital libraries, as there are a large number of existing digital libraries hosted by different institutions [4].

In the case of cultural heritage data, the most prevalent challenge for data management is its variety, due to the lack of homogeneity in the application of data models, languages and other data-related practices [4]. Ideally, historians, researchers and students should be able to search through multiple archives [3]. This is currently only possible with a generic search engine like Google [4]. Discoverability is typically addressed by Organisational Architecture, where the central organisation operates services to improve the discovery of cultural heritage resources through the collection of their associated metadata.

Generally, Web portals use the aggregation of metadata provided as input [4]. Specific data aggregation technologies are used in the domain of cultural heritage. These technologies are different from technologies used in other domains, such as Internet search engines or in the Web of data. The dominant technology is the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [3] since it specialises in the aggregation of metadata, covering some specific requirements for the aggregation of cultural heritage data [4]. This technology has been applied since its early stages in the field of cultural heritage.

Over the last ten years, efforts have been made to create a national heritage portal, where everyone can search through a unified collection specifically based on heritage documents [4]. The Europeana project [3], which links many European archives, is an exemplary but costly project. The South African National Electronic Theses and Dissertations (NETD) portal [13] is an example of a local system, designed and built at a low cost but for a different document domain.

The aim of this project is to create a small-scale prototype of a South African National Heritage Portal, using best practices in low-resource heritage archive systems to collect metadata spanning multiple local heritage archives and provide end-users with services such as searching and browsing across the archives using a central Web portal.

This paper is composed of four main sections: Web portals; metadata aggregation; metadata aggregator systems; metadata archive systems.

2. WEB PORTALS

This section discusses why Web portals are vital component of metadata aggregator systems and how they enable cross-archive discovery.

The Web portal is a human-usable interface that provides access to a metadata archive [13]. Users can browse and search the Web portal as they access indexed metadata records and link to the original documents in the source repositories. Web portals also encapsulate and make all standard Web services such as the OAI-PMH, Really Simple Syndication (RSS) and end-user services available to end-users in a single location [13]. Web portals are usually set up to deploy cultural heritage-adapted search engines tailored to the retrieval of metadata records [4]. The approach generally used to feed these portals is the aggregation of metadata, reviewed in Section 3.

A local project, the NETD [13], is implementing a Web portal that uses the OAI-PMH protocol to query a repository interface. The aggregated metadata records in the respective repository can then be accessed through a query via the Web portal. Mostly by harvesting from a repository, the Web portal creates a local database. Metadata records stored in the local database are processed for search indices and browser indices. The NETD system used Lucene to process search indices and MySQL to process browsing indices [13]. Following this framework, the Bonolo archive system provides a Web-based user interface that functions as a portal to end-users that exposes digital content stored in the repository for discovery through searching and browsing [10].

Finally, end-users of metadata aggregator Web portals can the search through a single archive or through a collection of archives using keywords and different filters, such as title, author and subject [10]. In the case of the NETD metadata aggregation system [13], a search service has been enabled by the integration of Lucene into the Web front end.

2.1. AJAX

This section reviews the use of AJAX technology to improve digital library Web portals and implement core end-user services that could be used for a low-cost national heritage Web portal.

Considering web portals, Asynchronous JavaScript and XML (AJAX) represents a fundamental transition from the classic Web application model and to what is possible on the Web by improving the richness and responsiveness of Web applications [5][11]. AJAX uses Extensible HyperText Markup Language (XHTML) and Cascading Style Sheets (CSS) for standard presentation; Document Object Model for dynamic display and interaction; XML and XML Stylesheet Language Transformations (XSLT) for data interchange and manipulation; XMLHttpRequest for asynchronous data retrieval and JavaScript for linking everything together [11].

In the classic Web application model, the Web interface action of the user triggers an HTTP request back to the Web server [5]. The Web server processes the request by retrieving the data and returns the HyperText Markup Language (HTML) page to the client. This classic Web interaction model allows one-way communication that limits the interaction and feedback of the Website [11].

Between the user and the server, AJAX applications introduce the AJAX engine [11]. Adding a layer to the application makes it more responsive and improves the flow of Web interaction [5]. AJAX allows two-way communication between the browser and the Web server without the need to install additional software or reload the page each time by the introduction of an AJAX engine, which acts as an intermediary [11][16].

At the start of a user's Web session, the browser loads in a hidden frame an AJAX engine which is written in JavaScript [5]. The engine renders the interface and communicates with the server. This allows the user to interact with the application independently of communication with the server. The user will not have to wait on the server to return the result to the user's request. This is achieved by engine handling actions that do not require a return trip to the server. JavaScript calls to the AJAX engine replace the HTTP request that would be generated by the action of the user in the classic Web application model. In the case of the engine requiring the server to respond, the request is sent asynchronously, usually using XML, without stopping the user's interaction with the Web application as compared in Figure 1 and Figure 2 [5][16].

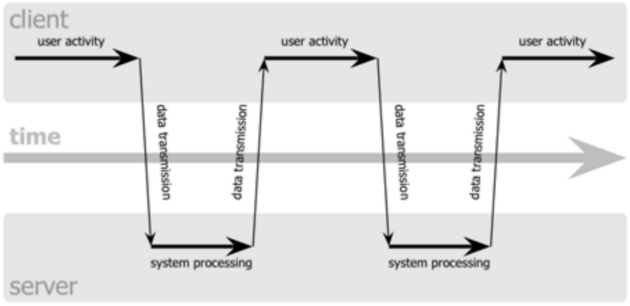

Figure 1: The synchronous traditional Web application model [5]

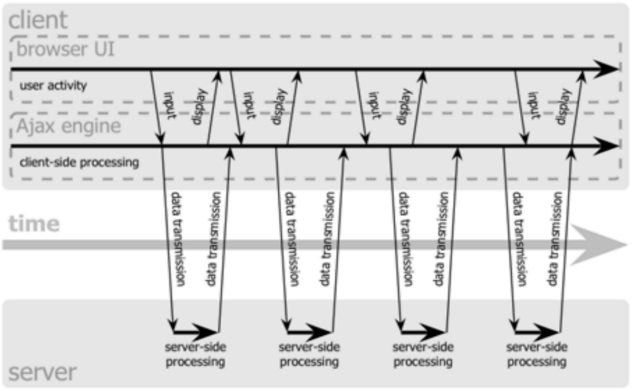

Figure 2: The asynchronous AJAX application model [5]

Traditionally, the server would perform most of the computations, while the AJAX in-browser service model differs, as most of the computation occurs on the client machine, resulting in the user interface being partially generated by server data and in-browser applications [16].

2.2 CLIENT-SIDE SERVICES

This section discusses services required by a Web portal, such as search and browse, and reviews how the client or server-side services have been implemented.

In most digital repository toolkits, core end-user services are search and browse [17]. These services incorporate elements of information retrieval and database operations, often by implementing a server-side indexer such as Apache SOLR. For example, the Bonolo archive system provides search and browse services by integrating an Apache SOLR search engine [10]. This allows an end-user to find content effectively by performing organised search requests filtered by specified search fields. In OAI-PMH implementations, search services are performed on harvested metadata by a search service provider, rather than by each metadata provider from whom the metadata was fully or selectively harvested [19].

If a small collection needs to be static, and portable, or stored client-side, browser-based search and browsing is possible using standard browser facilities [17]. Digital library service models tend to mainly use server-side applications [16]. Service models can be flexible by decomposing services into modules or be performed client-side to minimise bandwidth [18]. These services can be created to run within the Web browser of end-users using advancements in Web technologies such as AJAX. AJAX is a technology that allows browsers to perform more computing tasks than those traditionally included in server-side digital library services and enhances performance by moving computation to client-side [16].

In-browser services, performance and scalability are more powerful and flexible than the traditional server-side service model [16]. These client-side services can perform various useful tasks, with only minimal server interaction instead of loading new pages for each request as shown in Figure 3 [18]. Services could be invoked from a local device, operating without any Internet connection depending on the data collection size and offline storage capabilities.

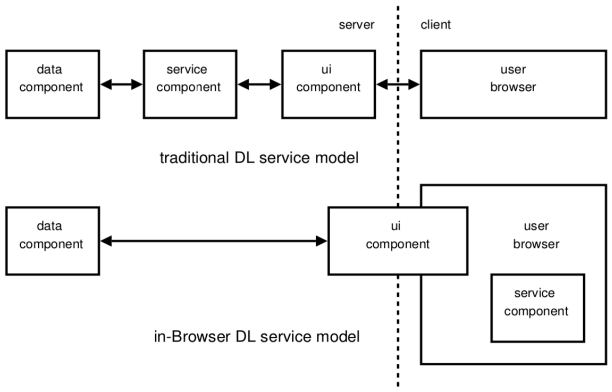

Figure 3: Traditional and in-browser data/service provider models [16]

Following this approach to services, the Bleek and Llyod Collection system uses AJAX to query a fully in-browser service [16]. Single word queries perform efficiently due to optimised document indexing and inverted files storing part of the document index relevant to its documents. Multi-word queries incur a penalty if document lists overlap. This problem is minimal, as users prefer shorter queries [16]. This altered information retrieval system requires more storage space for indices but decreases access time to documents for client-side applications.

CALJAX is a hybrid, digital library system that integrates repository updates into a large offline collection, giving full access to information to users in developing countries with limited Internet bandwidth [11]. When CALJAX is online, an AJAX engine retrieves updates for the offline client-side CD-ROM copy of a central repository. Combining the system-independent components provides search and browse services and online repository management. Similarly, to the Bleek and Llyod Collection system [16], the central repository stores indices which are generated by pre-processors for searching and browsing metadata content and a metadata collection that can be copied to removable media [11]. Implementing AJAX alone cannot list the directory's contents; thus, pre-processors are used to allow users to browse sorted object lists.

Indices are XML files that can be parsed in JavaScript and accessed via Web browsers [11]. Pre-processing is used to generate indices of files in a repository and to increase access speed. For offline use in systems like CALJAX, access speed may be slower than server-based systems. The pre-processors used for searching, generate input in the form of inverted files for an extended Boolean search implementation in the access Web interface. Common Web-based repository interfaces offer search and browse features [11]. The system responds to a user invoking a service by generating and displaying results from the relevant indices. A user can view interesting and relevant digital objects and navigate the result pages.

As a proof of the functionality of indices for search and browse systems based completely within a browser, a general-purpose offline client-side metadata search and browse system, also called a faceted search system, developed for browser use, was implemented and evaluated for varying behaviours and collection sizes [17]. Two sets of indices were stored as XML documents in order to be processed using integrated browser facilities. The offline tool was first evaluated with small amounts of FHYA [12] test data (about 100 items) and performed correctly [17].

However, it was unknown how well the system would work for larger realistic collections. The system performed quickly for typical small-to-medium-sized collections queries. The experiment concluded that such a system performance with no network connection, no software installation, and using JavaScript is possible with data pre-indexed and stored in XML files.

3. METADATA AGGREGATION

This section introduces the concept of metadata aggregation, its benefits for data discovery and use, and how it used to build a low-cost Web portal with the OAI-PMH as an example for aggregation implementation.

On the Internet, digital libraries expose a large number of resources [3]. Potential users face the challenges of discoverability and usage of these resources due to the fact that individual digital libraries are

maintained by different organisations. A standard solution to these challenges is the aggregation of metadata. Metadata aggregation is where a central organisation [4], called an aggregator, helps the discovery and usage of digital library resources that each digital library cannot offer in isolation [3]. This is achieved by an aggregator collecting or "harvesting" metadata from the storage of digital libraries or "repositories".

Recently, research data is considered to be a valuable research product in the data-centric period [20]. The generation, collection, organisation, analysis, mining and visualisation of data adds value to research data. Research data is vulnerable to issues such as storage, processing, transmission, curation, discoverability, sharing and re-use of research datasets. Suitable sharing of data for re-use to increase research efficiency and speed up the research development cycle is therefore crucial. Ensuring the availability of research data increases the verifiability of research and the citation count of related academic articles and papers.

Accessing digital objects from a range of archives requires the processing of different metadata encoding schemes and standards [7]. The data is only accessible through a Web portal after the aggregator has completed the processing of the metadata.

In the research area of cultural heritage, the OAI-PMH has become the most widely accepted solution for the aggregation of metadata [3]. Internet search engines are unable to retrieve user-requested resources based on the cultural heritage metadata domain combined with the World Wide Web hypertext documents. Subsequently, the retrieval of cultural heritage resources through search engines has shown to be ineffective [3]. The solution to this data synchronisation problem is the low-cost interoperability protocol - the OAI-PMH [3], and therefore, its frequent implementation [6].

Effectively, the OAI-PMH allows the sharing of metadata by the adopted data model of each aggregation case, as it is not limited to the data model to be used [3]. The only restriction imposed by the OAI-PMH is that the metadata must be in Extensible Markup Language (XML) form [8].

3.1 OAI-PMH

This section reviews the OAI-PMH, a low-cost, low-complexity interoperability protocol commonly used by international and local metadata aggregators, such as those discussed in Section 4.

Subsection 3.1.1 to Subsection 3.1.4 discuss key concepts that make up the OAI-PMH, such as repositories, records, requests, and harvesting.

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a simple and general HyperText Transfer Protocol (HTTP) - based client-server protocol that facilitates the incremental transfer of metadata between networked systems [14]. This protocol is an approach to interoperability with minimal interaction between remote sites and the central site, which primarily transfers metadata incrementally in bulk. Services are provided to users on internally stored and processed data acquired by an OAI-PMH service provider from an OAI-PMH data provider. By avoiding remote metadata searching and allowing the transfer of metadata, there is less interaction between remote sites in a scattered discovery environment. Typically, there are two types of entities in the OAI-PMH interactions: a data provider that exposes metadata to interested clients and a service provider that provides value-added services in addition to metadata collected from data providers [6].

Recently, more metadata aggregators have been able to obtain metadata-only information from individual data sources due to an increase in the number of repositories worldwide [6]. If a digital library offers an OAI-PMH interface to its repository, a service provider may access the repository of the digital library [14]. The digital content of the repository is systematically accessed using less bandwidth and with improved stability of service provision.

3.1.1 OAI REPOSITORIES

Repositories are accessible network servers of data providers that store digital data [8]. Digital data includes items that relate to metadata records harvested from a data provider. If a repository supports a collection of OAI-PMH requests, it is called OAI-compliant [8]. The association between records and items is many-to-one, as the metadata can be expressed in multiple formats. The scope of the OAI-PMH does not include the properties of an item, such as the type of metadata that is stored in the item, the type of metadata that is derived and whether the item includes the "full content" described in the metadata.

Specifically intended, the nature of the item is indistinct, as the OAI-PMH is defined as unaware of the data provider's standards [8]. The OAI-PMH supports data providers that have fixed metadata content, those that computationally derive metadata in different formats from some intermediate form or content itself, or those that are stores or intermediaries for external content providers of metadata.

Systematically, repository records are retrieved from a data store and made accessible via the OAI-PMH [8]. The repository data store of the NETD system uses MySQL for its SQL-compliant database [13]. The user of the OAI-PMH machine interface is unaware of the form and structure of the actual data storage [8].

Furthermore, the OAI-PMH interface is essential for scalability as another repository higher up the hierarchy, such as a continental or international repository, is capable of harvesting a national repository such as the NETD [13], resulting in minimal duplication of records and minimal network traffic.

3.1.2 OAI RECORDS

The OAI-PMH defines an XML encoded byte stream as a record. A record is made of three parts and serves as a packaging mechanism for harvested metadata [8]:

3.1.2.1 Header

This part contains the information required for harvesting and is shared among all records. Information is independent of the format of the metadata in the record. The information specified in the header is a unique record identifier and a date stamp indicating the date of creation, deletion or change in the record metadata.

3.1.2.2 Metadata
This part contains metadata in one format. All OAI-PMH data providers must be able to issue records containing unqualified Dublin Core (DC) metadata. Other metadata formats are acceptable.

3.1.2.3 About
This part is an optional container to hold data about the metadata record. Typically, the container could be used to hold information on metadata, terms and conditions, etc. The protocol does not define the inner structure of the container. Individual communities are allowed to decide on their syntax and semantics by defining a schema.

3.1.3 OAI REQUESTS

Data providers are repositories in which structured metadata is exposed through the OAI-PMH [8]. Service Providers send requests for the OAI-PMH service to collect the metadata. The OAI-PMH is a set of six verbs or requests that are invoked in the HTTP context [9]. The protocol uses the HTTP POST or GET methods [8]. The use of requests is to simplify the configuration of OAI-PMH compliant repositories for data providers by using readily available Web tools. All requests have the following structure:

base-url – comprises of the Internet host and the HTTP server port functioning as a repository. Optionally, with an OAI-PMH request path specified by the HTTP server.

keyword arguments – comprises of a list of key-value pairs. At a minimum, each OAI-PMH request has one key-value pair specifying the name of the request.

Responses to all requests to the OAI-PMH are encoded in XML. Each response contains the protocol request that generated it, which allows responses to be processed in machine batch. The requests are listed in Table 1 with their descriptions.

Table 1: OAI-PMH Requests and Descriptions

Protocol Request	Description
GetRecord	This verb returns a single record (metadata) from an item in a repository. The required arguments specify the identifier, or key, of the record being requested and the metadata format that should be used in the document.
Identify	This verb returns repository information. The response schema specifies information the verb Identify will return: a human readable name for the repository; the base URL of the repository; the version of the OAI-PMH supported by the repository and the e-mail address of the administrator of the repository.
ListIdentifier	This verb returns record identifiers which can be extracted from a repository. Optional arguments require identifiers to be chosen based on their association in a certain repository, or based on their alteration, development, or deletion within a given date period.
ListMetadataFormats	This verb returns the available metadata formats from a repository. An optional argument restricts the request for a particular record to the available formats.
ListRecords	This verb harvests repository records. Optional arguments require harvesting selection to be chosen based on their association of records in a certain repository, or based on their alteration, development, or deletion within a given date period.
ListSets	This verb returns the structure of the sets within a repository.

3.1.4 OAI HARVESTING

In the OAI-PMH, the process of harvesting is the periodic transfer of updated metadata from one machine to another [15]. Metadata harvested will be stored in a single location. The single location of the metadata has a minimal storage cost compared to the digital objects themselves [15]. The OAI-PMH harvesting services facilitate effective interoperability between content repositories by presenting specific aggregated community metadata through end-user search portals [19].

Notably, the NETD harvester component collects metadata records from a set of remote Electronic Theses and Dissertations (ETD) repositories [13]. Initially, a full harvest of each remote repository is made where all metadata records are retrieved and stored in the central repository. Subsequently, the harvester performs incremental harvesting where only metadata records are collected which have been added since the last harvest date. Incremental harvesting minimises data transfer and bandwidth usage. Retrieved metadata records are then stored directly in the shared repository data store. Metadata records are validated so as not to insert malformed records into the database. Time intervals may be set for the harvesting process to increase efficiency. The NETD system incorporates an administrative end-user interface, provided to monitor the harvesting process and manage the list of remote repositories.

4. METADATA AGGREGATORS

This section reviews how four metadata aggregator systems that can be adapted for a national heritage portal were developed.

4.1 NSDL

The National Science Digital Library (NSDL) is an American metadata aggregator that collects Web-based educational resources from Science, Technology, Engineering and Mathematics (STEM) data provider repositories to provide easy access to teaching and learning resources to educators, learners and the general public [15]. The metadata collected allows users to use a Web portal to perform structured searches and have open access to these resources.

Technically, the architecture of the NSDL is a service-oriented architecture as individual components communicate with a single defined protocol - the OAI-PMH [15]. The OAI-PMH uses a central metadata repository to store metadata harvested from distributed data provider repositories. The NSDL system stores both standard unqualified DC metadata and NSDL-specific variants and indexed full-text content. Metadata formats have been chosen to ensure that there is a low interoperability barrier [8]. Inserting metadata into the repository can be achieved either by entering metadata directly into the database, by harvesting metadata from data providers using the OAI-PMH, or by collecting metadata through Web-crawling [15]. The metadata of the repository is then shared with service providers, such as the local search service, via the OAI-PMH interface.

4.2 EUROPEANA

Europeana is a large-scale metadata aggregator that collects millions of digital European cultural and scientific heritage resources [2] from a variety of European cultural heritage institutions, such as libraries, museums, archives and galleries, to provide public access through

Web portals [3]. Users can search through archives and access resources through either Web portals or added third-party application services. The operational service of Europeana is based on the aggregation and exploitation of metadata for digitised objects from different contexts as shown in Figure 4.

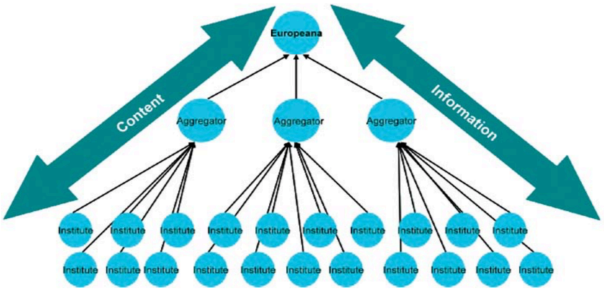

Figure 4: Aggregators in the Europeana organisational model [2]

Due to design complexity, Europeana was built at a high cost [3]. The design required the provision of seamless, efficient services on top of the aggregation. The solution to these challenges and the difficult issues of data integration came at the cost of time and technical expertise. Full-time staff were required to develop large-scale infrastructures and workflows for data aggregation, ingestion, indexing, standardisation and publishing [2].

Similarly, to the digital architecture of other distributed libraries such as the NSDL system, digital content is stored in the repositories of the distributed data providers. Still the metadata is saved and exchanged with a central Web portal [15], providing users with simple access and exploration of the heritage of the different European communities. Europeana's Web portal stores metadata as a networked data system based on Resource Description Framework (RDF) data model [15]. Thus, the rich relationships between and among objects and collections can be exploited by search and browse services.

Fundamentally, the data is either harvested using the OAI-PMH or manually entered into the central repository using a given Application Programming Interface (API) [15]. In this way, the core data architecture varies from the NSDL system, but the core network architecture is the same.

4.3 NDLTD Union Catalogue

The Networked Digital Library of Theses and Dissertations (NDLTD) operates a Union Catalogue/archive metadata aggregator, which gathers metadata publicising ETD resources from ETD repositories around the world [15]. This is a single mechanism to aggregate all ETD to provide NDLTD-wide services, such as searching [14]. Metadata harvesting was chosen as a simpler and more reliable solution compared to previous unsuccessful attempts of using federated search. Federated searching increases the complexity of the protocol system by transmitting search queries to all data providers.

Importantly, the OAI-PMH allows streaming of metadata from multiple repositories to a single location [14], ultimately standardising digital library interoperability [1] of distributed systems and collecting metadata from different data providers [14]. The creation of a single repository that stores metadata collected from OAI-compliant remote sites encourages NDLTD members to contribute from their host sites and allows sites to collect data via the OAI-PMH if core user services such as browsing, searching, or annotation are provided by the sites. Implementing the OAI-PMH on all NDLTD members archives allows user-level services to access all participating NDLTD members' metadata as shown in Figure 5. A central repository enables services to be developed at a central location without having to collect metadata multiple times, reducing the number of failure points to exactly one - an advantage as it manages to deal better with a single or replicated point than with arbitrary and distributed failure points [14].

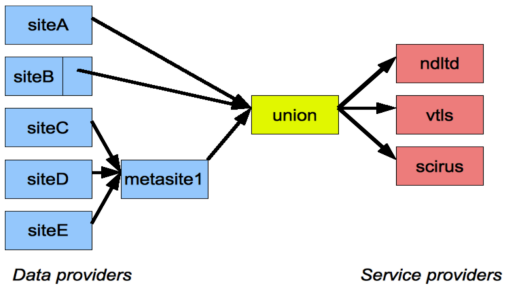

Figure 5: Data and service provider network of the NDLTD [15]

In the case of the OAI-PMH, the NDLTD Union Catalogue uses a harvester to harvest ETD metadata, the original data is then Web crawled and stored on a local machine [1]. Storing data as a local copy on a local machine is a safety measure in the event that the data publisher cannot provide data. The metadata in the repository is in DC or ETD-MS formats. ETD-MS metadata format is ETD-specific [15]. An ongoing challenge facing NDLTD is deduplication, as some sites are overlapping organisations. Metadata is only harvested if not harvested since the last harvest [1]. This scheme is an efficient mechanism for updating the mirror site or mirroring any data collection provided that the data collection has an associated OAI-PMH compliant server.

Similarly, to Europeana [3] and the NSDL system [15], metadata is harvested periodically from partner sites using OAI-PMH. Sites may represent either a single association, a country/regional project or an international collaboration. The NDLTD Union Catalogue architectural model [15] is a generic model for internationally focused digital library systems as it ignores the source of the metadata [8] and the linguistic and cultural differences in higher education systems around the world [15].

Using the OAI-PMH, services are again provided at a higher level by independent service providers receiving a stream of metadata from the Union Catalogue [15]. The Union Catalogue is able provide the service of a harvester and a data provider, by implementing the OAI-PMH as the harvested metadata that resides internally is republished through its own OAI data provider interface [14]. In 2011, both VTLS and Scirus provided metadata discovery interfaces [15].

4.4 NETD

The National Electronic Theses and Dissertations (NETD) portal is a South African metadata aggregator that provides access to a collection of local ETD [13]. This platform also facilitates the development of ETD projects across universities by organising, managing, and recording ETD. This portal is a custom multi-tiered, simple, repository component architecture chosen to improve system interoperability by interconnecting to larger repository systems such as NDLTD and similar organisations. It has been established to provide cross-archive services to the various digital library collections of ETD. A disadvantage of the portal's international services is that local institutions, such as universities, cannot fully duplicate these services [13].

However, the national archive provides comprehensiveness, sustainability, and expansion by collecting metadata from distributed source archives [13]. To publish this metadata on a central Web portal, the OAI-PMH was implemented. This system consists of three separable layers as shown in Figure 6: harvester, repository, and Web portal to support future expansion and scalability.

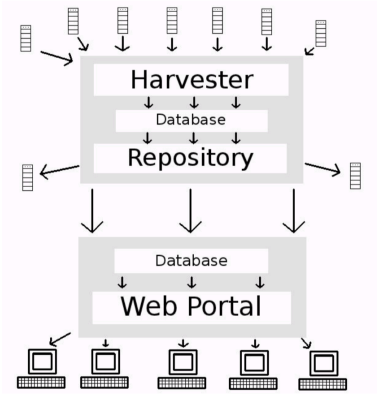

Figure 6: The NETD system architecture [13]

Remotely, repositories are harvested for ETD metadata records [13]. The local database stores verified metadata records that can be accessed through the OAI-PMH. The central repository is indexed by a search engine that provides a search service via a human-usable Web interface. The system architecture enables metadata results to be searched and browsed in a single Web portal. The results of a search or browse service are metadata containing a link to the original repository and the original documents. Using the OAI-PMH, the system can be harvested from potentially larger repositories like the NDLTD system. This simplifies the harvesting for larger repositories of a country or region. The advantage of this is that management is decentralised and closer to the responsible entities, with automatic propagation of metadata to higher levels [13].

Previously, similar work attempts at a unified NDLTD search system [15] implemented a federated search system [13]. A problem of this approach is that the system transformed and propagated user queries to each participating site, thus a search request would be made on each distributed site. Multiple results returned to the system from the search of each site, which was not merged. Merging was not a simple task because each site provided results in a different form of an HTML page. Another problem of this approach is that the system's performance relied on each participating site being reachable, for results to be obtained. Any delay in the system's search results was due to the time the participating site took to reply [13].

By implementing the OAI-PMH, the NETD system solved most of the problems faced by a federated search system [13]. The protocol allows less interaction between remote sites in a distributed discovery environment by harvesting metadata into a central repository, as opposed to remote metadata searching for each search query. The NDLTD Union Catalogue [15] and the NETD portal data transfer mechanism are based on the operation of service providers collecting and localising ETD metadata from the OAI-PMH compliant repositories for use [13]. The advantage of this approach is that it uses less bandwidth and provides greater stability than a federated search system. The Web portal then uses the metadata and other repository services to provide a unified discovery interface across all repositories harvested.

Evidently, the NETD system architecture is a component model in which each component is separated and operates independently, maintaining a balance between component size and function, and component management [13]. All components were written in Java using Java Servlet Web technology hosted on Tomcat servers in Ubuntu Linux. An advantage is that this approach offers a more scalable service as a whole. Components can be moved from a single server to their own server depending on the data collection size. Another advantage is that a component can be replaced rather than rewriting the entire system if the system needs to be updated or changed to suit a particular environment. Another important advantage of this architecture is that it avoids a single point of failure of the system [13].

5. DATA PROVIDER ARCHIVE SYSTEMS

This section reviews two digital archives that have expressed interest in the provision of data for a national heritage portal.

5.1 FHYA

The Five Hundred Year Archive (FHYA) is a digital archive that focuses on collecting digital artefacts and metadata linked to South Africa's pre-colonial past. This project is based locally at the University of Cape Town Archives and Public Culture Centre [12]. For the digital repository, a low-cost design was explored as the archive is hosted in a low-resource providing environment, mitigating maintenance costs and reducing the need for technical skills, thus helping to reduce system costs. The repository architecture focuses on a series of folders for digital artefacts and metadata.

Firstly, a script ingests a metadata spreadsheet of digital objects, converting metadata to individual XML files in the metadata directory [12]. Thereafter, a second script generates an HTML representation of each metadata file using an XML Stylesheet Transformation with thumbnails and pointers to the digital object; the thumbnails are integrated into the JavaScript slider to display a series of images in a fixed space. A JavaScript-based faceted search engine generates inverted files that enable search and browsing. These operations

produce a static website with the functionality of a read-only Web archive, that additionally enables comments to be stored in a collection displayed to the user on a regenerated digital object page. Dynamic page generation is not needed as adding a comment alters the static repository.

In order to maintain minimalism, the system uses short scripts instead of multi-layer abstractions, Web-based or agent architectures [12]. Read-only access is through direct access to HTML files without Web application mediation. JavaScript applications read static files to provide the search service. Only archive contributions require an online archive version and an Internet connection; otherwise a read-only archive can be distributed on offline media. The data and metadata stored in the repository are arranged in a hierarchical directory structure. Due to pre-processing data collection building, the system provides quick access to items as new items are ingested, as the system remains in a static state. As a result, the low-resource digital repository architecture specifically allows faster migration to other software systems not planned for low-resource environments [12].

5.2 THE BLEEK AND LLOYD COLLECTION

The Bleek and Llyod Collection is a portable digital archive commissioned by the Lucy Llyod Archive Resource and Exhibition Centre [16]. The Bleek and Lloyd series contains artefacts such as journals and sketches that link to and record the culture and history of |xam and !kun speakers in Southwest Africa [18]. After digitisation, digital artefacts and metadata were integrated into a system that allows full access. The multimedia set of journals and drawing photos are accessible digitally or on DVD-ROM and can be used irrespective of operating system or device architecture [16]. Firstly, all metadata was created, edited and converted to XML in Microsoft Excel [18]. Then a completely in-browser query service was created by implementing AJAX [16].

Greenstone was the first technological solution proposed, since it is the first digital library toolkit with possible distribution of CD-ROMs [18]. Greenstone needed applications to be installed on the target device and did not work on all devices. The second technology solution considered was to build an XSLT stylesheet to convert XML files to PDF format using XSL Formatting Objects (XSL-FO). This worked well for small collection subsets as PDF documents have an integrated search facility, and digital objects could be embedded in browsable pages. However, as the size of the collection increased, PDF files became unmanageable and most PDF viewers would significantly slow down due to file size. This choice was also not feasible [18].

Ultimately, the technology decided on is a static XHTML Website [18]. Using XSLT stylesheets, XML metadata has been converted to XHTML. A minor setback with this strategy was one of scalability and that was sufficiently overcome by using a combination of strategies such as dividing 16000 XHTML documents into batches to prevent them from being produced at once, using indices and keys to speed up XPath query resolution and logically grouping objects according to different parameters using the Muenchian process [18]. Another issue was that static Websites could not provide Internet users with services like search engines. The use of an AJAX-based query system addressed this.

Just as an information retrieval system, a set of inverted indices has been created and stored with the XHTML files generated [18]. Pre-generated inverted files allow the AJAX application to provide the end-user with the ability to search XHTML pages without any server-side dynamic operations. The system also functions whether the collection is delivered on a CD-ROM, on the Web or on a local drive, attracting and encouraging researchers and academics from different environments to use this data [18].

6. SUMMARY

Table 2: Comparison of Metadata Web Portals

System Name	Domain	Type of System	Cross-Archive Discovery	Client/Server-Side Services	Technology used for providing services	Complexity/Cost
Bleek and Llyod Collection	|xam and !kun speakers Heritage	Digital Archive	No	Client-Side	AJAX	Low
Europeana	European Heritage	Metadata Aggregator	Yes	Server-Side	OAI-PMH	High
FHYA	Pre-Colonial South African Heritage	Digital Archive	No	Client-Side	JavaScript	Low
NDLTD Union Catalogue	ETD	Metadata Aggregator	Yes	Server-Side	VTLS & Scirus	Low
NETD	ETD	Metadata Aggregator	Yes	Server-Side	Lucene	Low
NSDL	Science, Technology, Engineering and Mathematics	Metadata Aggregator	Yes	Server-Side	OAI-PMH	High

From observing Table 2, a common distinction between digital archive systems and metadata aggregators is that metadata aggregators, such as the NSDL, Europeana, the NETD, and the NDLTD Union Catalogue, typically store metadata collected from multiple remote repositories, whereas digital archives, such as the CALJAX, the FHYA and the Bleek & Llyod Collection, generally store a single collection of the original digital objects and associated metadata. This literature review looked at these two variations of digital libraries in terms of the technology used to develop them, their architecture, their implementation and how they provided search and browse services for end-users to either a single collection or multiple archives through a Web portal.

From the discussion, improvements and limitations of existing technologies and systems provided insight into the project in order to create a central metadata aggregator Web portal, that collects and provides access to data, from data providers in a lightweight implementation. The low-cost OAI-PMH and AJAX Web technology, has shown from the systems and architectures discussed, that they are suitable and useful for the construction of a national heritage portal from the collection of metadata. Both the form of metadata and the size of the collection are factors in the ability of the metadata aggregator to provide any offline services or storage on a browser for end-users. The OAI-PMH has shown to be a significant link in the transfer, preservation and display of cultural heritage metadata from a variety of heritage archives. Metadata aggregation is a feasible solution to provide cross-archive searching and browsing, as the local NETD system is an example that this type of metadata Web portal for the ETD domain, was possible.

REFERENCES

[1] Anil Bazaz and Edward Fox, 2003. Preservation of ETDs on NDLTD, NCSTRL Computer Science Technical Reports at Virginia Tech, Virginia Tech, Blacksburg, VA, available at: http://eprints.cs.vt.edu/archive/00000651/01/oai_ndltd2.pdf.

[2] Sally Chambers and Wouter Schallier, 2010. Bringing Research Libraries into Europeana: Establishing a Library-Domain Aggregator, Liber Quarterly.

[3] Nuno Freire, Antoine Isaac, Glen Robson, John Howard, and Hugo Manguinhas, 2017. A survey of Web technology for metadata aggregation in cultural heritage. Information Services & Use, Vol. 37, Issue 4, 425–436.

[4] Nuno Freire, Rene Voorburg, Roland Cornelissen, Sjors de Valk, Enno Meijers, E and Antoine Isaac, 2019. Aggregation of Linked Data in the Cultural Heritage Domain: A Case Study in the Europeana Network. Information, 10, 252.

[5] Jesse James Garrett, 2005. Ajax: a new approach to Web applications. http://www.adaptivepath.com/ideas/essays/archives/000385.php

[6] Nikos Houssos, Kostas Stamatis, Vangelis Banos, Sarantos Kapidakis, Emmanouel Garoufallou, Alexandros Koulouris, 2011. Implementing enhanced OAI-PMH requirements for Europeana. Research and Advanced Technology for Digital Libraries, 396-407.

[7] Unmil P. Karadkar, Geoffrey A Potter and Shengwei Wang, 2016. Visualizing Published Metadata in Large Aggregations, Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries.

[8] Carl Lagoze and Herbet Van de Sompel, 2001. The Open Archives Initiative: Building a lowbarrier interoperability framework, Proceedings at Joint Conference on Digital Libraries, Roanoke, VA.

[9] Carl Lagoze, Herbert Van de Sompel, Michael Nelson, and Simon Warner, 2002. The Open Archives Initiative Protocol for Metadata Harvesting, Open Archives Initiative. Available http://www.openarchives.org/OAI/openarchivesprotocol.html

[10] Lighton Phiri, Kyle Williams, Miles Robinson, Stuart Hammar, and Hussein Suleman, 2012. Bonolo: A General Digital Library System for File-Based Collections. In: Proceedings of the 14th International Conference on Asia-Pacific Digital Libraries. Ed. by Hsin-Hsi Chen and Gobinda Chowdhury. Berlin, Heidelberg: Springer Berlin / Heidelberg, 49–58. DOI: 10.1007/ 978-3-642-34752-8_6.

[11] Hussein Suleman, Marc Bowes, Matthew Hirst and Suraj Subrun. Hybrid online-offline digital collections, 2010. Proceedings of the 2010 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on - SAICSIT '10, ACM Press, 421-425.

[12] Hussein Suleman, 2019. Reflections on Design Principles for a Digital Repository in a Low Resource Environment. Proceedings of HistoInformatics Workshop 2019, 13 September 2019, Oslo, Norway, CEUR.

[13] Hussein Suleman, Tatenda Chipeperekwa and Lawrence Webley, 2011. Creating a National Electronic Thesis and Dissertation Portal in South Africa. Proceedings of 14th International Symposium on Electronic Theses and Dissertations.

[14] Hussein Suleman, Edward A Fox , 2002. Towards Universal Accessibility of ETDs: Building the NDLTD Union Archive, Proceedings of The Fifth International Symposium on Electronic Theses and Dissertations (ETD 2002), Provo, Utah, USA, Available at http://www.wvu.edu/~thesis/proceedings.html

[15] Hussein Suleman, 2012. Design and architecture of digital libraries. Facet Publishing.

[16] Hussein Suleman, 2007. In-Browser Digital Library Services, Proceedings of Research and Advanced Technology for Digital Libraries, 11th European Conference (ECDL 2007), 16-19 September 2007, Budapest, Hungary, 462-465.

[17] Hussein Suleman, 2019. Investigating the effectiveness of client-side search/browse without a network connection, Proceedings of 21st International Conference on Asia-Pacific Digital Libraries (ICADL), Kuala Lumpur, Malaysia, Springer.

[18] Hussein Suleman, 2008. An African Perspective on Digital Preservation, in Post-Proceedings of International Workshop on Digital Preservation of Heritage and Research Issues in Archiving and Retrieval, Kolkata, India, 29-31 October 2007. Available http://www.husseinsspace.com/research/publications/iwdph_2007_african.pdf

[19] Sarah Shreeves, Joanne Kaczmarek, Timothy Cole, 2003. Harvesting cultural heritage metadata using the OAI protocol. Library High Tech 21, 2, 159-169

[20] Malcom Wolski, Joe Young, Joanne Morris, Lance De Vine and Robyn Rebollo, 2010. Metadata Aggregation - A Critical Component of Research Infrastructure for the Future, eResearch Australasia 2010 (Vol. 2010). Gold Coast, QLD: University of Queensland.