Creating a Centralized South African Cultural Heritage Web Portal

Ashil Ramjee
University of Cape Town
Cape Town, South Africa
rmjash002@myuct.ac.za

ABSTRACT

The act of digital cultural heritage preservation is essential for the future. Efforts have been made to create digital archives in South Africa, however, these are not interlinked in any way. Portals such as Europeana in other countries show how it is possible to create a centralized portal. Therefore, the creation of one centralized South African Web portal would be beneficial for users of cultural artifacts. With the emergence of new technologies such as the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), the creation of a portal is feasible in the context of South Africa. This literature review gives a brief overview of the idea of metadata and three metadata schemas. Furthermore, it discusses portal architectures with different core characteristics. These include federated search systems, digital libraries implemented with metadata harvesting and combinations of both. Lastly, digital libraries, an important aspect in the creation of a portal are discussed. This includes various architectures designs and the services they provide.

KEYWORDS

cultural, heritage, portal, metadata, metadata harvesting, search, digital libraries

1. INTRODUCTION

Cultural heritage and the preservation of its artifacts play an important role in benefiting future generations. Artifacts vary from books and writing to drawings and paintings. These artifacts are not only useful to historians but rather the general public which includes governments and schools. Furthermore, social, environmental, and political aspects are influenced by cultural heritage which may directly affect the nature of decisions taken by the government [1].

Traditional means of physically storing tangible cultural heritage artifacts is outdated due to artifact deterioration, however, technological advancements such as digitization provide a means to store, search and analyze these artifacts more efficiently and reliably [2]. XML databases, the internet and Web 2.0 technologies are building the necessary environment to preserve information [23]. Countries are preserving their cultural heritage artifacts by implementing digital archive systems with the use of digital libraries and other services. South Africa is no exception and existing archive systems such as the Bleek and Lloyd collection and Digital Innovation South Africa both document different cultural areas of South Africa [14]. However, no South African archive integrates all these collections into one centralized repository. Europeana is an example of how archives can be accessed through one centralized portal [3]. The problem with South Africa is that its socioeconomic and economic problems bring upon limitations in the creation, maintenance, and accessibility of these systems. Firstly, gaining appropriate funding is difficult due to less interest in incorporating digitization into cultural preservation [2]. Secondly, many people are digitally illiterate due to the lack of skills and education which affects the constant collection of data and maintenance of these systems [2]. Thirdly, areas with limited access to the Internet and unstable connections bring up challenges related to accessibility. Furthermore, areas with constant Internet access are bandwidth constrained due to bandwidth being very expensive [2].

Therefore, this paper will be exploring ways in which a successful centralized South African cultural heritage portal prototype can be created by discussing metadata standards, overviewing different portal architectures, and expanding on digital libraries while considering South Africa’s limitations. Furthermore, specific technicalities needed to create the portal will be mentioned. These include but are not limited to interoperability, the design of the portal, metadata, metasearching, metadata harvesting, digital libraries and the services they provide. After discussing the ideas mentioned in this paper, a brief summary outlining the main aspects of creating a portal will be given.

2. METADATA

Remote cultural archives all vary in terms of their architecture therefore their heterogeneous metadata also varies in terms of structure [11]. Furthermore, digital cultural artifacts vary from images, videos, books, text, or combinations of each other. Moreover, remote archives use different formats such as the Dublin Core or MARCXML to structure metadata [11]. This section explores ways of structuring metadata with the use of standardized formats. In addition, it shows how some formats cater for complexity in terms of the metadata.

2.1. Extensible Markup Language (XML)

XML, a subset of Standard Generalized Markup Language (SGML) and with a similar structure to HTML can be used to describe metadata with varying complexity due to its variety of tags [13]. The use of an XML based schema allows for different document types with different tags for markup with varying sizes [13]. Therefore, in terms of storing cultural artifacts, an XML schema may be a suitable candidate due to its capabilities of storing different information. Furthermore, with the use of OAI-PMH, metadata from remote sites with different standards may be harvested into XML while keeping the integrity of the metadata intact. Interestingly, once harvested, storing the metadata without databases is possible [14]. This is the case with the Bleek and Lloyd Collection (BLC) which uses an XML schema to store its data. CDWA Lite, an XML based schema that describes cultural heritage data shows how XML can be used to address more complex metadata [15]. It shows how an XML markup can be drawn up and tailored towards specific complex data. Both Dublin core and the Resource Description Framework are XML based schemas [12][13]. With the use of best practices from other XML schemas, a cultural heritage schema can be marked up for a South African cultural heritage portal.

2.2. Dublin Core (DC)

DC which was originally developed to define Web resources is one the most common metadata standards which are compliant with OAI-PMH [16]. Its 15-element set is comprised of well-researched resource elements that describe the data in an acceptable simplified manner [12]. Core design features include simplicity, semantic interoperability in the sense that DC can be used with many different disciplines. Furthermore, with the flexibility and modularity, DC allows you to add additional structure and semantics if needed [12]. In terms of creating a portal, data collected from either a federated search system or metadata harvesting could be filtered and displayed using DC to reduce complexity. However, not all resource descriptions are simplified and may require a standard more complex. DC is however constantly updating their standards and introducing new capabilities.

2.3. Resource Description Framework (RDF)

The RDF is a relatively new metadata standard with a similar objective as DC which provides resources descriptions for the internet [13]. Similarly, to DC, its core design offers semantic interoperability [13]. The XML based schema includes features such as namespace found in XML which allows different standards to be used within RDF [13]. Therefore, different standards can be applied to the RDF which in turn means that the RDF can now hold more than one metadata standard [13]. This allows certain rules from specific standards to be adhered to while to keeping the data integrity intact. In terms of the portal, this could be useful with the storage of data collected. For example, if DC records were harvested, RDF capabilities allow the DC rules to be used therefore the DC tags and naming scheme will be used when showing the record after it has been searched or browsed [13]. RDF capabilities can be adopted into other standards including DC, which is promising, considering that remote archives vary in terms of architectures and metadata standards. Europeana is an example of a portal that extensively uses RDF as a metadata schema [3].

3. PORTAL ARCHITECTURES

In previous efforts, portals can be created in many ways with similar core characteristics. To avoid the cost associated with creating a digital library from scratch, toolkits such as DSpace, Atom or Fedora could be used [14]. However, Section 4 will explain a more cost-effective digital library that can be implemented instead.

Firstly, the most common and cost-effective approach would be to create a union catalog which can be done by implementing a digital library with a single Web interface for data providers to contribute data [17]. A simple Web interface directly linked to the repository would be used by data providers to add their collections to the repository. Data providers are also able to send their metadata in the correct schema for service providers to add to the collection instead of using the web interface. However, data providers are forced to use a single metadata schema which is set by the service provider therefore metadata integrity for most collections would not be kept intact [17]. INFLIBNET, a union catalogue that was created to centralize academic libraries in India uses a similar approach as mentioned above, however, the service provider manages the conversion of data [18]. The basic architecture of the system comprises of three tiers. For the first tier, unconverted data is received from data providers and manually converted to one format, however, there are inconsistencies with the data integrity. For the second tier, the converted data is uploaded to a UniCat system that manages the workflow of the SQL database system. UniCat performs validity checks, data integrity checks, removes duplicates and finally adds the records to the database. Tier three consists of the software for searching, displaying, saving, or downloading the files. Contradictorily, the database does not conform to one metadata standard which presents concerns with regards to complexity, uniformity, and consistency [18]. Moreover, developers of the system were working with the XML protocol, Z39.50 protocol and a data harvesting tool which suggests that the architecture of the system was outdated and could be improved [18].

Portals can also be created by implementing a federated search engine [9]. This can be done by sending a query to multiple registered sites or clients and returning results to the server. Each client site would have its search syntax which would need to be added to the database of the system. Commonly, the Z39.50 protocol is used as it facilitates the use of queries and merging of results after executed the queries [9]. A system which multilingually searches through heterogeneous collections uses a similar architecture [9]. The creators based the architecture on the term “simple search” which emphasizes that most users enter single words or small phrases when searching [9]. Primarily, the system comprises of a Searchable Database Markup Language (SearchDB-ML), a translation protocol and a federated searcher [9]. Firstly, the SearchDB-ML is used to store the search syntax of each site. Secondly, the translation protocol facilitates the act of creating a single search query to search through multilingual search interfaces. Thirdly, the Java-based federated searcher sends queries to client sites which respond by returning results that are merged. Along with the other types of problems these systems face, this approach primarily relies on a stable Internet connection between the clients and servers when searching therefore not catering for low resource environments [14]. In addition, maintenance and scalability would be cumbersome since each site search syntax would need to be added to the search syntax database.

Contrastingly, a more feasible approach to creating a portal would be to use the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) to harvest metadata from remote archives, store it in a local repository and implement a digital library [7]. The overall architecture of the portal is multi-tiered and split into three sub-systems, namely, the metadata harvester, the repository and the web interface that incorporates digital services in a user-friendly manner [7]. However, the repository and Web interface could be implemented as a digital library providing the essential search and browse services. These sub-systems need to be interconnected and should allow safe transference of the harvested metadata between them. The OAI-PMH enables this interconnection by providing interfaces for each sub-system, therefore, providing an interoperability framework [7]. The use of OAI-PMH provides a reliable standardized protocol that most remote archives support therefore making harvesting less complex and timeous [7]. This approach is based on the system architecture of the National Electronic Thesis and Dissertation Portal in South Africa (NETD) [7].

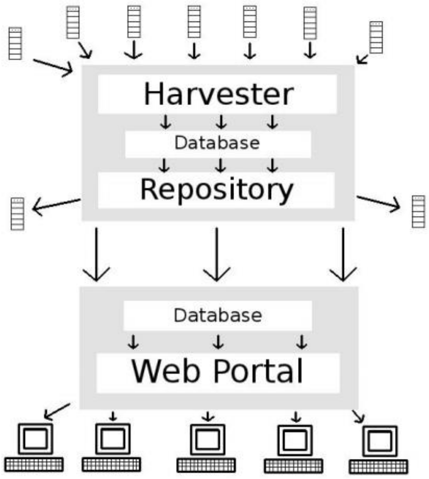

Figure 1. NETD Architecture [7]

As shown in Figure 1, the harvester provides the repository with metadata [7]. The Web portal fulfils the search and browse services by using the metadata from a local repository. In terms of addressing the limitations, this architecture uses less bandwidth due to incremental harvests that add to the local repository therefore metadata is only harvested if it is new or modified. Furthermore, the use of the standardized OAI-PMH reduces complexity, ultimately increasing the ease of maintainability and reducing the funding needed. In addition, the adoption of this architecture was based on reusability and scalability [7].

Interestingly, both approaches could be combined into a system that entails parts of a federated search system and metadata harvesting. For example, MARIAN, a system which provides interoperable federated digital library services, addresses the interoperability problems of a federated search system by using the OAI-PMH and other protocols [10]. This allows a search of the locally harvested data and federated searches on other sites to be merged [10]. However, although worth mentioning, this system is very complex and South Africa's digital environment makes it infeasible.

The Networked Digital Library of Theses and Dissertations (NDLTD) was formally implemented as a federated search system that provided services that could search through Electronic Theses and Dissertations (ETD) [9]. However, due the increase in scale and the introduction of the OAI-PMH, the NDLTD is now a union catalogue that uses metadata harvesting as a method of retrieving metadata [8]. The purpose of the NDLTD is to use the OAI-PMH to harvest ETDs from data provider collections and store them in its local repository while adhering to one metadata format, namely the Dublin Core format as required by the OAI protocol [8]. In addition, the standard ETD format is also used. Furthermore, with the use of the OAI protocol, NDLTD uses the OAI data provider interface to add support for service providers who would like to harvest form the union archive. Therefore, the NDLTD acts as a data provider for service-orientated archives and a service provider for ETD collections [8].

Europeana, a centralized cultural heritage portal that interlinks many Europeans archives and resources is an example of a large-scale portal, consequently costing a large amount to create and maintain [3]. Data is either incrementally harvested with the use of OAI-PMH or added through an interface. The costs are explained by its scale and its ability to cater for rich data that other portals or digital libraries cannot [3]. The Resource Description framework schema (RDF) is used to cater for complex metadata types by using simple triple statements [3]. With the use of RDF, APIs and the OAI Object Reuse and Exchange framework (OAI-ORE), it is able to provide the means of supporting and displaying complex digital objects. It uses metadata aggregation which essentially collects Web resources [3]. OAI-ORE allows Web resources to be identified by using aggregation resources which are described by resource maps. In addition, it models its data using an object-centric and event-centric approach [3]. Europeana takes the crown in terms of complexity however it shows that with the right amount of funding and knowledge, a system on this scale is possible [3]. Features such as the use of RDF could be beneficial when creating a portal in the context of South Africa.

The next section discusses digital libraries which constitute a large portion of the creation of a national portal.

4. DIGITAL LIBRARIES

Digital libraries are a set of services that facilitate the storage, management, and access to digital objects [4]. In terms of cultural heritage, digital objects or metadata are simply pieces of structured data that comprise of a unique identifier and other data relating to the cultural artifacts [4]. Consequently, the collection and storage of data constitute a large proportion of complexity in terms of implementing a digital library within a portal. This section gives an overview of the design principles, methods of collecting metadata, architectures, services, and various toolkits that may be used when implementing a digital library.

4.1. Design

Users of the system include the general public therefore the design needs to be simple to install and manage [5]. Generally, the success of any system does not entirely depend on aesthetics. Instead, it depends on the user experience therefore there are design principles that should be considered that interconnect both design and usability. The user interface of a digital library plays an important role within the design [23]. Some user interface design principles include simplicity, the familiarity of the interface, informative feedback, error prevention, multimedia and profile-based support, and multilingual support [23].

Pandey [5] outlines the following design principles for a digital library as a whole. The design should be service-driven, must support open architecture, cater for scalability, preserve data by making sure data is persistent and private. Additionally, the design should be practical and flexible. However, these principles are the minimum standard to be adhered to when designing a digital library. Furthermore, they do not address the limitations mentioned in the introduction of this paper. On the other hand, Suleman's [6] paper reflecting on design principles is geared towards a low resource environment and is therefore more feasible in the context of South Africa. Based on the Bleek and Lloyd collection, the following principles were derived [6][14]. With *minimalism*, the idea is to use short scripts instead of other multi-tiered or layered architectures. The *imposition on users* allows already assigned identifiers to be reused. Without the use of direct access to HTML files and Javascript applications, *no internet* is required for search and browse service. *Simple preservation* is maintained by using XML and stylesheets. If Web browsers cannot view a certain data format, the file will be downloadable therefore introducing the principle of *Any Objects*. *Flexibility* allows the reuse of search and browse services. *Superimposed information* allows comments on data objects to be stored separately, therefore, keeping data intact. *Hierarchical organization* supports hierarchical metadata storage. A system that can run on any Web browser regardless of the operating system is *platform agnostic*. Lastly, *collection binding* allows the preprocessing of data.

4.2. Retrieving metadata and Architectures

Digital libraries are able to retrieve their data in various ways. These include contributions to repositories or union catalogs and metadata harvesting. Certain digital library architectures support these methods by using certain protocols. Therefore, this sub-section discusses methods of retrieving data. In addition, it discusses different implementations of digital library architectures.

4.2.1. Retrieving Metadata

Using a traditional union catalog which would be accessible to registered sites would be the simplest way of collecting data and implementing a digital library [17]. In terms of simplicity, registered sites would contribute to one web interface which supports a single metadata schema which reduces complexity and increases interoperability within the system [17]. In terms of cost-effectiveness, it shifts the costs of converting data to the data providers since they need to adhere to the schema specified by the repository host [17]. Although ideal in some situations, collecting metadata in this way has its limitations. Data providers are forced to leave out certain elements if they do not fit into the metadata schema [17]. Therefore, in the context of cultural heritage artifacts, many elements would be left out due to the complexity in some collections.

The emergence of OAI-PMH allows digital libraries to retrieve metadata in an automated, incremental, and safe process [17]. The act of metadata harvesting involves the incremental collection of metadata from different remote archives with the use of scripts [17]. In addition, the metadata collected is either converted to another schema and then stored in a single repository or stored as-is in the repository [17]. This is done with the use of mapping and crosswalks [17]. However, service providers are able to edit the schema by adding or removing elements, but this does reduce interoperability. In its simplest form, OAI-PMH implements an unqualified Dublin Core schema as a standard to be used for data providers [17]. With supported OAI-PMH repositories, the OAI-PMH framework harvests the DC records from the repository, examines each record while drawing out the Identifier which in turn holds the URL and other information needed to locate the resource. Furthermore, without using OAI-PMH, the resource is collected and added to the server providers local repository [16]. Records are created or modified on remote archives frequently, therefore they need to be harvested again in order to be up to date. One approach to doing this is by using the OAI-PMH date stamps to inform the harvester that the records need to be harvested again. However, it has been identified that there are flaws with using the dc.identifier and date stamps [16]. Firstly, the dc.identifier tag does not always point to the network location of a resource for some metadata schemas, therefore, the conversion to unqualified Dublin core brings upon complications. Secondly, the use of date stamps is not always foolproof in the sense that there may be missed updates or unnecessary downloads [16].

Data providers using more complex metadata schemas need to map their elements to the unqualified Dublin Core to achieve compliance [17]. Van De Sompel, et al's paper on the use of OAI-PMH for harvesting shows how it is possible to harvest more complex metadata with the use of extensions external to the OAI-PMH framework [16]. In addition, it has identified that many complex metadata formats share characteristics such as the use of wrapper XML documents, therefore, these could be used as a basis to fit them into the OAI-PMH data model [16]. It devised a way in which the OAI-PMH framework could be used to harvest complex metadata formats while addressing the flaws of using the dc.identifier and date stamps. To put it in simple terms the OAI-PMH framework would be used in the same way with added processes [16]. Firstly, the complex records are checked if they supported by the harvester. Secondly, if they are supported, the records are harvested while addressing the flaws of the date stamp. A parser is then used to extract bitstreams and references from each record which are used for identification and to trigger harvesting. Lastly, another process collects bitstreams which are then used to store records in the service provider's database. Although this solution is specific to a certain project, it shows how service providers are not limited to only using the standard unqualified Dublin Core schema.

Metadata harvesting is a promising protocol that offers interoperability. In addition, the idea of incremental harvests is cost-effective in the sense that less bandwidth is being used. Therefore, it may be feasible to use it for collecting data when implementing a digital library.

4.2.2. Architectures

In section 4.1, design principles for a low-cost digital library were discussed. These principles were based on the architecture of the Bleek and Lloyd collection which essentially uses XML files instead of a database to store its data [14]. The architecture addresses the concerns of using a database-driven repository

[6][14]. In addition, it shows that a typical client-server architecture is not needed to create a digital library [6]. The basic architecture comprises of XML, XSLT, XHTML, and other scripts and could be considered as a portable architecture [14]. Firstly, XML generated Excel sheets were cleaned into one XML file. XHTML pages are then generated from the XML file. Indexing using XSLT keys is used therefore search and browse services can be executed. In addition, the architecture makes use of an Ajax-based search system with the use of the XHTML pages.

Similarly, to the Bleek and Lloyd collection, The Five Hundred Year Archive (FHYA) uses the same repository architecture [6]. However, the archive used a Javascript-based faceted search engine to fulfil the search and browse services. Web scripts update the static repository when changes are made. The extensive use of XML has its advantages; however, it does not address the concerns of scalability which may be essential in when creating a national portal [14].

Contrastingly, a more complicated architecture, namely, Fedora, based on the use of digital objects is an example of a multi-layered architecture or a Web-based architecture [20]. XML encoded digital objects store metadata and other resources. In order to store metadata and provide Web services, the repository comprises of three layers, namely the Web services Exposure Layer, the Core Subsystem Layer, and the Storage Layer [20]. Firstly, the Web Services layer provides three services that provide an interface for the creation and management of the digital objects (Management Service), an interface for accessing individual digital objects and their data (Access Service), and lastly a lightweight Access Service (Access-Lite-Service). Secondly, the Core Subsystem Layer validates and keeps the digital object integrity intact. In addition, the system reveals behaviors of objects. Lastly, the Storage Layer facilitates the actions of reading, writing, and deletion of data. Digital objects include a Dublin Core record in addition to other metadata schemas. Therefore, OAI-PMH is supported and harvesting can be used to fill the repository.

4.3. Services

Digital libraries are not only storage systems for information. They also provide information retrieval services [21]. Fundamental services include search, browsing and indexing, however, there are a variety of services that digital libraries provide. This subsection explains three services, namely, indexing, search and browse.

4.3.1. Indexing

Firstly, with the use of indexing, the search and browse services can be executed more accurately and efficiently [22]. Annika, et al’s paper on low-cost semantic enhancement of metadata and indexing outlines a few effective low-cost strategies of indexing [22]. Two of these ways include lexical indexing and semantic indexing. Lexical indexing is a full-text indexing strategy that examines the test and creates index entries of the location of where each term occurs [22]. On the other hand, semantic searching uses mining from other sources to match concepts in the text [22]. If concepts match, they are indexed to full-text indexes. The paper also introduces the idea of enhancing lexical searching through semantics by using concept labels [22].

4.3.2. Search and Browse

There are a several ways of executing the search service. As discussed before, federated searches with the use of search queries and the Z39.50 protocol can be used to search within a digital library [9]. The translation protocol used to search through multilingual heterogenous collections could be adopted to address the issue of multilingual searching [9][21]. To display the data, search queries need to be sent to each site and metadata is returned in their different standards [17]. Depending on the architecture, the data can either be shown in its different standard or can be merged and converted into one uniform standard. To fulfil these operations, certain protocols need to be used. One of the most common protocols, Z39.50, a client-server protocol, allows queries to be sent simultaneously to sites for searching [17]. In addition, it provides indexing services. However, sites need to also support Z39.50 for the search and indexing services to work. Similarly, to the TCP/IP protocol, Z39.50 is based on the idea of clients and servers. In essence, the client communicates with a server or computer by sending a search query [17]. The client computer then returns its results in its own syntax and order which the service provider must handle. In addition, the server may need to convert the metadata to a single metadata schema. Sites that support the Z39.50 protocol make it easier for federated search systems, however, other problems such as displaying the data in terms of relevance, date collected, or categories need to be considered [17]. Semantic mapping and the use of crosswalks allow elements with similar grouping to be identified therefore allowing search queries to be used with merged results [17]. Although using its own protocol which was based on Z39.50 and other protocols, a federated search engine system was made to search through heterogeneous collections [9]. As mentioned in Section 2, this system uses a search syntax database which addresses the problem of many websites not supporting Z39.50. Another possible way of dealing with the limitations of Z39.50 would be to use XML Gateway which provides services that allow queries and results to be formatted in XML [17]. Following Z39.50, Z39.50 International: Next Generation (ZING) was introduced which expands the functionality of Z39.50 [17]. Metasearching does have its advantages, however, in terms of costs and context metadata harvesting may be more suitable for the creation of a portal.

With the use of indexing, the simplest, although not very accurate simple search uses single words or phrases to search through a digital library [9][22]. On the other hand, an advanced search introduces filters and parameters for searching [22]. The use of Boolean logic allows advanced searches to be conducted [21]. These filters and parameters include relevance, proximity to similar phrases, date of publication, languages, and case sensitivity. In most cases, users do not know what they are searching for which may be problematic [22]. In addition, spelling and other mistakes contribute to wrong searches. Therefore, users may prefer to browse a digital library which can be done by browsing the indexes [21]. In addition, with the use of grouping of indexes, browsing can be done by viewing fields of study and authors for example.

4.4. Toolkits

To reduce the complexity of creating a digital library from scratch, toolkits can be used to implement a digital library instead while providing the main search, browse and other services. However, toolkits require expertise a good internet connection, therefore, they may not be suitable [6]. This section discusses three toolkits, namely, DSpace, Fedora and Omeka.

4.4.1. DSpace

DSpace is simply an open-source toolkit that allows users to create repositories [19]. In addition, it manages and provides services for these repositories. However, it mainly tailored to be used by institutional repositories. The DSpace software package consists of four levels, each relating to specific user needs [19]. The first level provides search and browse services to the general public with the use of an interface. The second and third levels provide an interface for contributors and community administrators to create, upload, edit or delete records depending on their accessibility privileges. The fourth level allows developers to customize and maintain the software. An as-is implementation of DSpace uses Dublin Core as a metadata standard, however, users of the toolkit are able to customize it to some extent [19]. DSpace also provides plugins that can be used for OAI metadata harvesting. However, there are limitations to using DSpace. Although based on Dublin Core, the DSpace naming scheme does not conform to the naming scheme used by Dublin Core therefore cross-mapping of metadata can cause inconsistencies and incomplete records [19]. Therefore, it may be infeasible to use such a package for cultural artifacts due to the complexity of the data.

4.4.2. Fedora

The Fedora toolkit is another open-source repository system that provides web services [20]. Unlike DSpace, Fedoras core architectural functionality is built upon the idea of object models [20]. These objects models are able to hold digital resources and data regarding these resources. In addition, they interlink the data to customized software and services in order to provide the Web services in the way that the user specifies. The repository part of the package allows access to these object models. Metadata is represented with the use of datastreams which are encoded in XML [20]. With its customization functionality, Fedora is able to provide specific services to Web browser. With the use of the namespace feature and XML, Fedora is compatible with many metadata standards [20]. However, this contributes to complexity therefore Fedora would be harder to implement than DSpace. Its object models include a Dublin Core record therefore it provides support for OAI-PMH. In essence, the package provides the implementer with an OAI-PMH supported repository, Web Services to provide search, browse and indexing services, and an interface for users to use these services [20].

4.4.3. Omeka

Omeka is an open-source collection management toolkit and unlike DSpace and Fedora, it is specifically tailored to storing and managing cultural heritage collections [24]. It is favorable due to its view on metadata representation which allows system administrators to customize the metadata schema [24]. Furthermore, the system is easy to install and maintain which addresses the concerns of South Africa’s digital environment. However, as a server-based application, Omeka can only be run on a Linux distribution [24]. Omeka uses CSV files as a means of collecting its data. With the use of plug-ins, it can used as an OAI-PMH repository and can facilitate harvesting of metadata [24]. Other plug-ins and features include support for XHTML, RSS, and search and browse services. Therefore, out of all three toolkits mentioned, Omeka could be used as a starting point for the implementation of a digital library within a national heritage portal.

6. SUMMARY

This paper has introduced the concept of creating a national heritage portal in South Africa while mentioning the limitations its environment entails.

Section 2 gave an analysis of metadata schemas such as XML which both Dublin Core and RDF are based on [12][13]. Both Dublin Core and RDF offer semantic interoperability. RDF, which is used by Europeana is a promising schema that can be used with complex metadata [3].

Section 3 discussed various portal architectures ranging from union catalogues, federated search systems and metadata harvesting repositories. Union catalogues provide the easiest and most cost-effective means of creating a portal, however, it was noted that there are many limitations such as data integrity compromises and inconsistencies [18]. Federated search systems, a client-server architecture uses search queries which are sent to registered sites which consequently returns results [9]. However, similarly to other client-server applications, a stable Internet connection is needed, therefore, it does not address the limitations of South Africa. With the introduction of OAI-PMH, a multi-tiered architecture that incrementally harvests metadata from data providers was discussed [7]. Furthermore, it addresses some of the limitations concerned by providing means of interoperability. MARIAN, a system that combines both federated searching and metadata harvesting was mentioned, however, the architecture is very complex [10]. Next, two notable portals, namely the NDLTD and Europeana were overviewed [3][8]. Europeana is a prime example of a portal created with scalability considered [3].

The last section gave an overview of the design, methods of collecting metadata and architectures, services, and toolkits that can be used to implement a digital library, an important component in the creation of portals. User-interface principles and low-cost digital library principles that were derived for the Bleek and Lloyd collection were mentioned [6][14]. Methods of collecting metadata include data providers that can contribute data to one union catalogue with a required specific schema [17]. In addition, metadata harvesting which uses the OAI-PMH to incrementally harvest data can be used [16]. OAI-PMH addresses the issue of Internet access and limited bandwidth by performing incremental harvests. Furthermore, the low-cost Bleek and Lloyd digital library architecture was discussed [14]. The architecture made extensive use of XML, XSLT, XHTML [14]. Moreover, Fedora, an architecture based on the use of digital objects outlined a multi-layered architecture [20]. Next, services such as searching, browsing, and indexing were explained. In this regard, indexing techniques such as lexical and semantic indexing are favorable [22]. With the use of the Z39.50 protocol, metasearching can be used to retrieve results after sending search queries to sites [17]. The paper ended off by comparably overviewing the functionality of three toolkits, namely, DSpace, Fedora and Omeka, which could be used to implement a digital library.

REFERENCES

[1] Michele Pickover, Dale Peters. 2002. DISA: An African Perspective on Digital Technology. *Innovation, 24.*

[2] Hussein Suleman. 2011. An African Perspective on Digital Preservation. In *Multimedia Information Extraction and Digital Heritage Preservation*, 295-306. https://doi.org/10.1142/9789814307260_0016

[3] Martin Doerr, Stefan Gradmann, Steffen Hennicke, Antoine Issac, Carlo Meghini, Herbert van de Sompel. 2010. The Europeana data model (edm). In *World Library and Information Congress: 76th IFLA general conference and assembly.* 10-15.

[4] William Y. Arms, Christophe Blanchi, Edward A. Overly. 1997. An architecture for information in digital libraries. In *D-lib magazine,* 3(2). http://mirror.dlib.org/dlib/february97/cnri/02arms1.html

[5] Richa Pandey. 2003. Digital library architecture. In *DRTC Workshop on Digital Libraries: theory and practice.http://dlissu.pbworks.com/w/file/fetch/44829234/B_architecture.pdf*

[6] Hussein Suleman. 2019. Reflections on Design Principles for a Digital Repository in a Low Resource Environment. In *Proceedings of HistoInformatics Workshop 2019*, 13 September 2019, Oslo, Norway, CEUR. http://pubs.cs.uct.ac.za/id/eprint/1331/1/ho_2019_lowresource.pdf

[7] Lawrence Webley, Tatenda Chipeperekwa, Hussein Suleman. 2011. Creating a national electronic thesis and dissertation portal in South Africa. http://pubs.cs.uct.ac.za/id/eprint/748/1/etd2011_webley.pdf

[8] Hussein Suleman, Edward A. Fox. Towards universal accessibility of ETDs: building the NDLTD union archive. In *Fifth International Symposium on Electronic Theses and Dissertations (ETD2002), Provo, Utah, USA* (Vol. 30).

[9] James Powell, Edward A.Fox. 1998. Multilingual Federated Searching Across Heterogeneous Collections. In *D-Lib Magazine.* 4(8). September 1998. http://www.dlib.org/dlib/september98/powell/09powell.html

[10] Marcos André Gonçalves, Robert K. France, Edward A. Fox, and Tamas E. Doszkocs. 2000. MARIAN Searching and Querying across Heterogeneous Federated Digital Libraries. In *DELOS*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.21.2223&rep=rep1&type=pdf

[11] Herbert Van de Sompel, Michael L.Nelson, Carl Lagoze, Simeon Warner. 2004. Resource Harvesting Within the OAI-PMH Framework. *D-lib magazine.* 10(12). https://digitalcommons.odu.edu/cgi/viewcontent.cgi?article=1002&context=computerscience_fac_pubs

[12] Stuart Weibel. 1997. The Dublin Core: A simple Content Description Model for electronic resources. *Bulletin of the American Society for Information Science and Technology*. 24(1), 9-11. https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/bult.70

[13] Judith R. Ahronheim. 1998. Descriptive Metadata: Emerging Standards. *Journal of academic librarianship*, 24(5), 395-403.

[14] Hussein Suleman. 2007. Digital libraries without databases: The bleek and lloyd collection. In *International Conference on Theory and Practice of Digital Libraries*. 392-403. Springer, Berlin, Heidelberg

[15] Regine Stein. Erin Coburn. 2008. CDWA Lite and museumdat: New developments in metadata standards for cultural heritage information. In *Proceedings of the 2008 Annual Conference of CIDOC* 15-18. https://pdfs.semanticscholar.org/5673/ced82275df25749a62297cf412b31b59621a.pdf

[16] Herbert Van de Sompel, Michael L.Nelson, Carl Lagoze, Simeon Warner. 2004. Resource Harvesting within the OAI-PMH framework. *D-lib magazine*, *10*(12). https://digitalcommons.odu.edu/cgi/viewcontent.cgi?article=1002&context=computerscience_fac_pubs

[17] Mary S. Woodley. 2008. Crosswalks, metadata harvesting, federated searching, metasearching: Using metadata to connect users and information. Getty Research Institute. http://scholarworks.csun.edu/handle/10211.2/2001

[18] Prem Chand, Suresh K. Chauhan. 2008. The union catalogue of academic libraries in India: an initiative by INFLIBNET. *Interlending & Document Supply.*

[19] Mary Kurtz. 2010. Dublin Core, DSpace, and a brief analysis of three university repositories. *Information technology and libraries*, 29(1), 40-46. https://ejournals.bc.edu/index.php/ital/article/view/3157

[20] Thorton Staples, Ross Wayland, Sandra Payette. 2003. The Fedora Project. *D-Lib Magazine*, 9(4), 1082-9873. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.458.4750&rep=rep1&type=pdf

[21] Alastair G. Smith. Search features of digital libraries. *Information Research*, *5*(3). http://informationr.net/ir/5-3/paper73.html

[22] Annika Hinze, Sally Jo Cunningham, David Bainbridge, J. Stephen Downie. 2016. Low-cost semantic enhancement to digital library metadata and indexing: Simple yet effective strategies. In *2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*. 93-102. IEEE

[23] Hanumat G. Sastry, Lokanatha C. Reddy. User interface design principles for digital libraries. *International Journal of Web Applications*, *1*(2), 86-91. http://dirf.org/ijwa/v1n20109.pdf

[24] Jason Kucsma, Kevin Reiss, Angela Sidman. 2010. Using Omeka to build digital collections: The METRO case study. *D-Lib magazine*. 16(3/4), 1-11.