Digital Heritage Preservation

NOOSRAT HOSSAIN, University of Cape Town

The need for digital cultural heritage preservation has grown with the increase in digital content. One the better applications of digital heritage preservation is archive data collection. This in turns calls for action the best practices when implementing a digital archive. We need to decide on the best tool to use and the best architecture to implement. Which services to offer and how to ensure sustainability. There are three architectures investigated, Peer-to-Peer, Grid based and Layered Architecture. Peer-to-Peer consists of independent digital libraries all communicating with a global hub. Grid Based has several separate entities that communicate in a hierarchy. Layered architecture would be the most common and most efficient architecture to implement, with three simple layers. There are a variety of tools for different user needs, DSpace and EPrints for Documents, some simple implementations like Google Cultural Institute or Omeka, and a few more complicated like Fedora and Invenio who rather than offering all the services allows the user to decide what the end-user needs are and customize their own services. The choice of tool depends on the experience of the administrator, little experience may prefer simple in-tool services but a more complex system may need a tool that is extendible. To get an idea of how to present the archive and what services to offer there are many existing archives; even several South African archives. The importance of cultural heritage preservation is that of great concern for South Africa and the rest of the world.

Additional Key Words and Phrases: Cultural Heritage Preservation, Digital Libraries, Digital Repositories, African Heritage Preservation, Digital Archives.

1. INTRODUCTION

The Three Archives project entails migrating three heritage archives holding various artefacts about Cape Town's history. There has been a rapid growth in the volume of digital content over the years. The need for cultural heritage preservation arises due to the possible damage of such artefacts from weather, pollution and even the use by the general public [Navrud and Ready 2002]. If this is the case there must be a way to both preserve cultural heritage artefacts without limiting access to interested parties and the general public.

There are costs to making cultural objects publically available due to security overhead, transport and change of location. Digitization of data has the ability to store large amounts of data in small amounts of physical space, ensuring public availability on demand globally and reduces the risk of loss due to unforeseen events [Vilbrandt et al. 2004].

There are various ways of digitally preserving cultural heritage. Initially this is done by digitizing cultural artifacts like images and texts and reconstructing lost artifacts like paintings and architecture. After having collected all this content, the way to digitally preserve this data, as will be discussed in this paper, is by archiving this data in a digital cultural heritage archive or digital library [Vilbrandt et al. 2004].

2. ARCHITECTURE OF A DIGITAL ARCHIVE

This section will be discussing the architectures implemented when creating a digital heritage archive, the basic principles when constructing the archive and constraints.

2.1. PEER TO PEER ARCHITECTURE

A Peer-to-Peer architecture for digital libraries consists of independent digital libraries hosted locally that want they're information to be distributed globally. [Podnar et al. 2006]

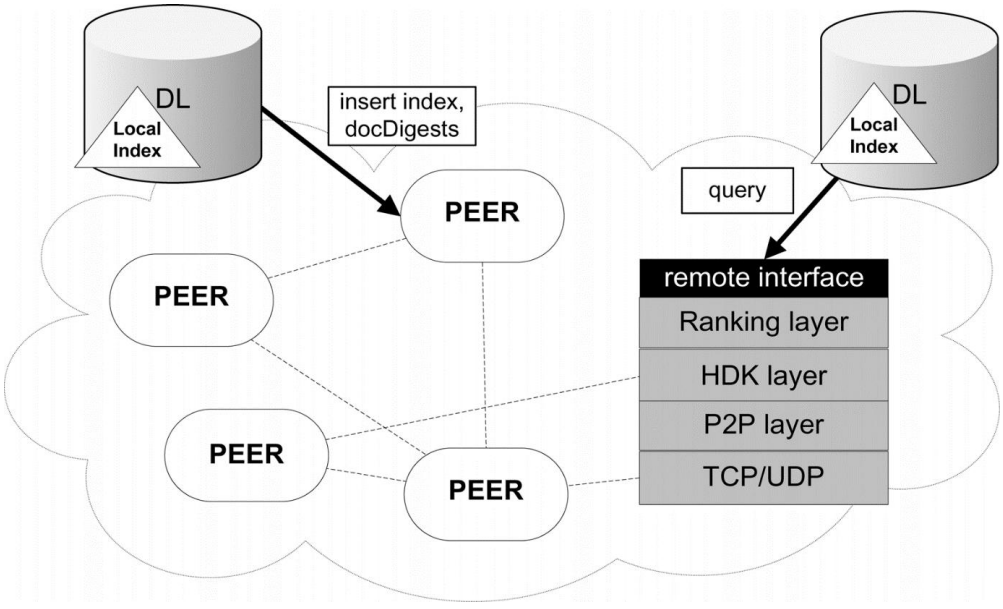

Figure 1: Overview of Peer-to-Peer Architecture for Digital Libraries [Podnar et al. 2006]

In the figure above we see a separate layered architecture where some peers connect to specific layers for specific services. A transport layer to host communication, a P2P layer storing queries in a hash table, the HDK layer maps the queries to keys and the ranking layer ranks the documents in the data libraries. This kind of architecture is mostly used for document storage and mainly offers one type of service, that being simple storage and search [Podnar et al. 2006].

2.2. GRID BASED ARCHITECTURE

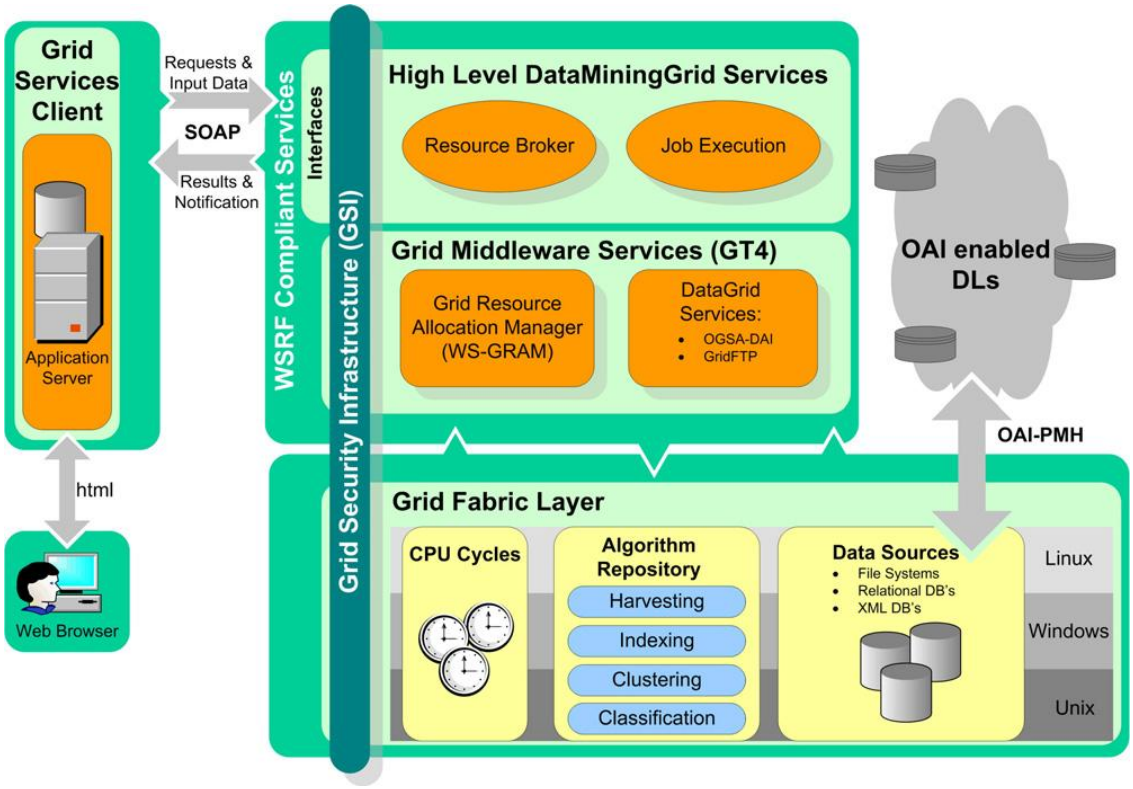

Figure 2: Architecture of a grid-based personalized FDL system [Trnkoczy et al. 2006]

The above diagram depicts how a grid based architecture for a digital library would work. The user interface is in the form of a webpage. An application server connects the user to the grid services. The top level of the grid handles two services: the Resource Broker which matches a request to a job and the Job Execution Service which initiates and monitors the execution of Jobs. Jobs represent the tasks a user can perform with the digital library. The middle level consists of the execution and data management services. The bottom level holds all the resources: Computational, Storage and Software resources. The storage resource connects to the World Wide Web to retrieve data from digital libraries [Trnkoczy et al. 2006].

2.3. LAYERED ARCHITECTURE

Though these architectures vary amongst different archives, they are often structured in a layered architecture with the following components: the repository layer (which stores the data), the service layer, and the user interface (controlling the user interaction). The diagram below depicts such an architecture [Phiri and Suleman 2015].

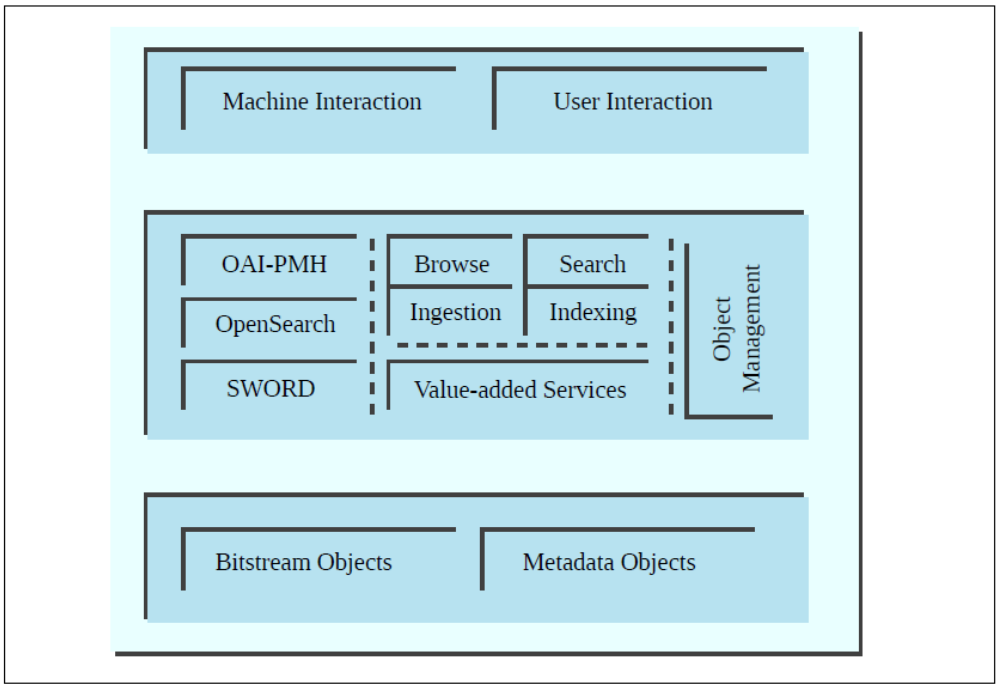

Figure 3: High level architecture of a typical Digital Library System [Phiri and Suleman 2015]

INTERFACE

When designing this layer it needs to be considered what components are connected to which back end element. Service-oriented Architecture is used in this case: this allows for systems to interconnect with other systems [Suleman et al. 2006]. To implement this kind of architecture a foundational set of services from existing platforms should be used.

The interface layer, also referred to as the user experience layer, can be further split into three sections: Services, Flows and Pages[Suleman et al. 2006].

Services
Connects pages associated with specific services to the service layer.

Flows
Controls navigation. Connects and assigns a flow structure to a specified list of pages.

Pages
Controls the visual design of each page[Suleman et al. 2006]

There are typically two main interfaces provided in a heritage archive system: Curator and end-user. Each interface offers different services and has different permissions. Curator interfaces will typically provide more administrative tasks like uploading and organizing content and the end-user interface will provide a means to view and download the content [Phiri and Suleman 2015].

SERVICES

The service layer holds the services required to access and use the digital content [Phiri and Suleman 2015]. This acts as the layer that would connect the user to the data. Many repositories offer a standard default set of services that offer the basic functionality one would expect to find. Some users may want a different set of services depending on what the library entails. This is where service-oriented architecture can be implemented [Suleman et al. 2006]. For example; when the portability of the content is in question often archives set up Web-accessible services making it easily and publically accessible on a range of platforms.

Examples of services include: Search, Browse, Object Management [Phiri and Suleman 2015], Exhibit and Compare [2015].

REPOSITORY

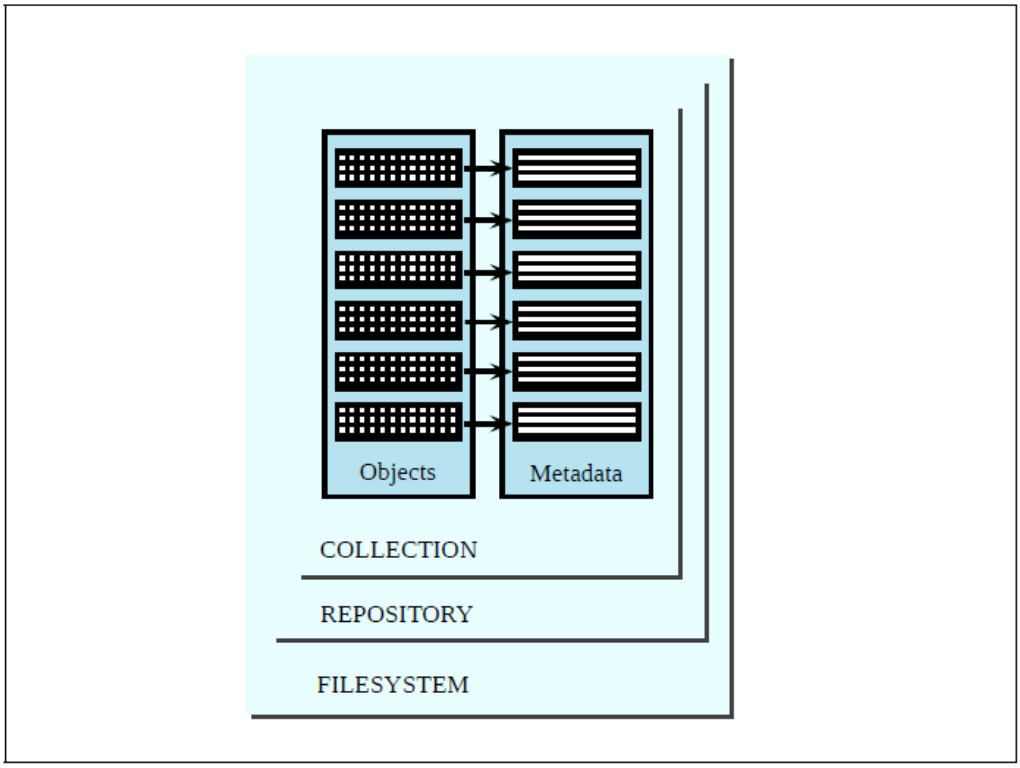

Figure 4: Repository Object Organisation

The figure above depicts how the repository layer of this architecture would be set up. The repository contains individual digital objects each linked to its individual set of metadata. The repository is stored using a typical native file system. This file system needs to ensure ease of access to the data objects; the data should be easily available and modifiable. This can be a file-based repository, a database-based repository or web-based [Phiri and Suleman 2015].

The Digital Object

A digital object is a data structure composed of digital material, or data, and a unique identifier called a handle[Kahn and Wilensky 1995]. The diagram below depicts the structure of a typical digital object:

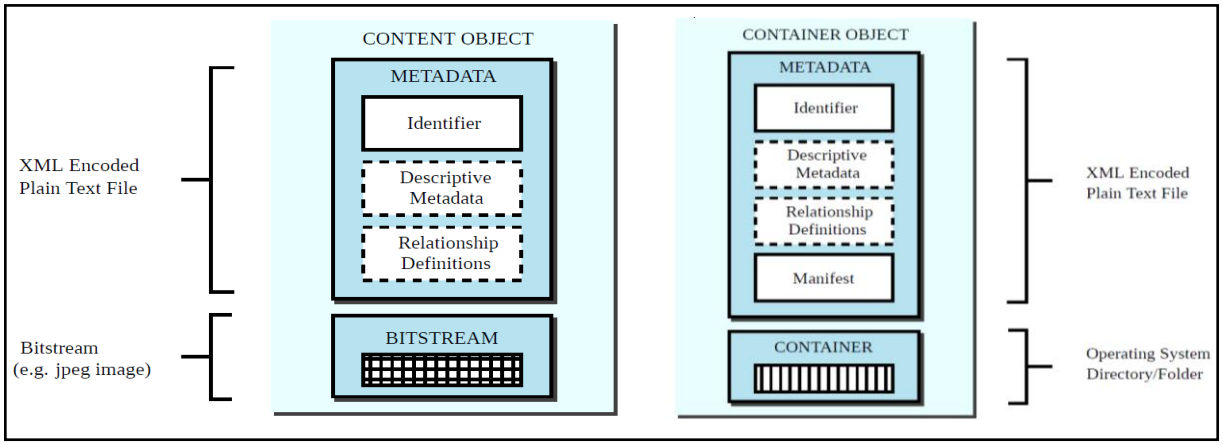

Figure 5: Digital Object Structure [Phiri and Suleman 2015]

A digital object consists of two components; the content object and the container object. The content object is the digital object stored in the repository. The container object includes a manifest; a detailed log containing other objects within the repository [Phiri and Suleman 2015].

Metadata

Metadata is information that uniquely identifies an object; describing it in more detail [Phiri and Suleman 2015]. Metadata is broken up into key and structural[Arms, William Y and Blanchi, Christophe and Overly 1997]. There are a variety of metadata forms, therefore a framework should be put in place to ensure interoperability; for example The Warwick Framework [Daniel Jr et al. 1998]. The Warwick Framework is a simple, structured framework and uses the Dublin Core metadata element set; a standard in which fifteen properties are used to represent or describe a data object [Dcmi 1995].

Data/Bit Streams

The bit stream refers to the content type.

2.4. CONSTRAINTS

There are a few constraints when designing cultural heritage archives; one being scalability. Scalability refers to whether a system is able to expand due to an increasing load. As mentioned, the amount of digital content is ever growing, so a digital archive need to be scalable in order for it to be sustainable. A cultural heritage archive needs to ensure preservation. It needs to be portable and content should be able to be easily shared [Phiri and Suleman 2015].

2.5. PRINCIPLES

A set of guidelines has been set when designing digital archives. They are as follows: Tools and services should be software independent. The library should be able to easily integrate new services and data types. Standardising in accordance with community and international norms can facilitate interoperability. It is best to use the minimum of external software; simplifying overall design and making it easier to manage. The structure needs to be kept simple for anyone to use as current administrators will not be with the specific project forever. Data should be structured and organisation. Lastly, the system must account for access in situations with least possible resources available

[Phiri and Suleman 2015]. Following these Principles can ensure sustainability for the archive.

3. USABILITY

Usability can be evaluated in different ways depending on which interface is being evaluated. It can be evaluated using three components: Content, Functionality and User interface [Körber and Suleman 2008]. Otherwise, usability can defined by how easily the user can find information. i.e. navigation, which will be explored further, is a key component of a digital archive's usability.

3.1. ADMINISTRATIVE/CURATIVE

Installation and configuration are the stepping stones to providing a digital library [Körber and Suleman 2008]. These are two vital stages in the operating of digital archives and should be done as intended.

3.2. END-USER

Usability for the end user is measured with different metrics [Blandford and Buchanan 2003]. One of these metrics is goal achievement, if the system does what the user intends. Learnability measures how well a user can figure out how to use the system after being exposed to it for the first time. Another metric is how well a user can recover after coming into an error. Lastly, end-user usability measures the overall user experience and enjoyment and how well the system fits in with the context [Blandford and Buchanan 2003].

4. TOOLS

A selection of tools can be used as a foundational platform for archival management. Each has different aspects and components that satisfy user needs. DSpace is an open source dynamic repository created by MIT used to manage digital resources to be used as is or with modification to suit user needs [Smith et al. 2003]. It is implemented in java and works just as a UNIX like system. DSpace follows a layered architecture similar to the one described above; an application layer handling user interaction, the business logic layer has an extra fixed service of managing users and user authentication, and the storage layer [Tansley et al. 2003]. DSpace was intended for use by academic institutions and other similar organisations. DSpace offers many of its own services and in many aspects is unable to fully integrate with other systems. Each layer is completely separate and each layer can only communicate with the one immediately below it.

A similar tool is the EPrints system. Many of the features are the same but EPrints is optimized. DSpace provides a platform for long term preservation whereas EPrints provides access to documents provided by authors. DSpace draws on the design of EPrints, EPrints being one of the earlier archiving tools. EPrints has a desirable interoperability in providing cross archive access [Tansley et al. 2003].

A platform that can allow for more customisation and integration is the open source integrated digital library system Invenio. Invenio implements a modular design, unlike DSpace, which enables it to perform a large variety of requirements [Caffaro and Kaplun 2011]. Each layer was ensured to be flexible. Invenio can store up to 10MB records [Caffaro and Kaplun 2011].

Many of the tools discussed are most often not stand alone and the administrator will need sufficient knowledge in the architecture and development of these types of platforms to fully utilize these tools. Google cultural institute provides a way for anybody with access to the internet to create their own heritage archives. The organisation provides a large-scale system for collecting, archiving, organizing, and interacting with digital assets of cultural material [Seales, W Brent and Crossan, Steve and Yoshitake, Mark and Girgin 2013]

Other simple archiving tools exist. Omeka is an open source collection management system aimed at smaller heritage projects and institutions with lack of technology and budget [Scheinfeldt 2008]. Metadata is a vital tool in representing digital objects, Omeka libraries incorporates the Dublin core set but can also allow for administrators to customize a metadata vocabulary. Omeka provides features like building online exhibits [Kucsma et al. 2010]. Omeka provides modular software architecture allowing extension through plugins to provide specific services [Scheinfeldt 2008].

When a system wants to offer several services and general frigid, unable to be customised, plug-in is not enough. Flexibility is key; after being tested in the field Fedora was found to accommodate a diverse set of information management problems. Unlike DSpace and EPrints primarily used for institutional repositories applications laid on top of fedora were complex digital library collection [Lagoze et al. 2005].

5. EXISTING ARCHIVES

5.1. LOCAL

The preservation of African culture is that of utmost importance to our society now and in the future [Suleman 2008]. The main concern for Africa preservation, unlike the rest of the world, has to do more with access to and archiving of heritage culture rather than the long term preservation. Though the concerns for African preservation are similar to that of the rest of the world there are specific concerns that affect African heritage: deterioration of artefacts, most notable verbal as many languages are not spoken anymore, the lack of funding and access to resource hinder the ability to preserve cultural artefacts and lastly the skill level in the continent is insufficient for the execution of successful heritage preservation [Suleman 2008]. Local successes need to be investigated.

Some examples of digital libraries in an African context are the Nelson Mandela Centre of Memory archive and the Digital Innovation South Africa Archive. The Nelson Mandela Centre of Memory had an abundance of digitized material documenting the life and times of former president Nelson Mandela. The Google Cultural Institute was brought on board to build the Nelson Mandela archive. They provide an online platform use to preserve and promote culture through galleries, virtual tours and user customisation [2015]. The system allows the end-user to contribute through the share stories service provided and is able to organise digital exhibits of the content which includes a narrative as seen in the picture below.

Figure 6: Nelson Mandela Archive exhibition service [2015]

This archive tries to ensure usability by providing a constant available help tool that shows the user how to navigate around the interface. The interface unfortunately has one flaw. The ambiguous home button takes the user to the Nelson Mandela Centre of Memory site rather than the archive home page [2015].

Traditionally data libraries focused mostly on the storage of digital artefacts. As time moved forward the storage of such artefacts was no longer the issue and data library systems began to move away from traditional databases. Some of the tools mentioned above like EPrints and DSpace still use this type of framework for their storage layer. The Bleek and Lloyd collection is a good example of an archive that has moved on from the database type framework [Suleman 2007]. The Bleek and Lloyd collection is a set of books and drawings documenting the language and culture of bushman groups in Southern Africa. These documents were scanned and had metadata generated for each item. The metadata was recorded for each item in an excel spreadsheet then converted to XML [Suleman 2007].

The aim for most digital libraries, more especially heritage archives, is to make content more easily available to the public. This is a constraint when designing archives, an existing project that struggled due to this constraint is the Digital Innovation South Africa(DISA) Project [Pickover 2008]. The Aim of the DISA Project was to build an archive to preserve fragile South African cultural artefacts regarding the struggle and to make these artefacts available to a wide audience [Saunders 2005]. DISA has become part of a larger heritage archive called ALUKA, an international digital archive focusing on African culture, which implemented a subscription based system where users had to pay to view content [Pickover 2008]. ALUKA has recently been acquired by JSTOR digital Library database [ITHAKA 2015].

All these archives are of great importance to the African community and any future works can take example form these.

5.2. INTERNATIONAL

With the rapid growth of technologies overseas and the constant improvements it is possible to learn a few things from existing international archives. A typical example is Europeana, one of the most well-known archives internationally. To the general public Europeana is seen as a very popular, having up to 10 million hits very hour when it started [Purday 2009], portal into the large cultural heritage of Europe [Concordia et al. 2010]. What many are unaware of is that Europeana provides an API service to the end user, where they can lay their own portal over the Europeana framework [Concordia et al. 2010].

Similar to Europeana, the Australian library Trove provides an API to its users to draw of information from trove in different ways [Holley 2010]. Trove itself does not store any of the content, only metadata. Trove links; like a search engine. Trove has a variety of sources from 1000 different libraries, museums and galleries. Trove also allows for public engagement and interaction with the site like crowdsourcing and user to user communication [Holley 2010].

Heritage archives don't have to be vast and full of variety, many hone in one specific historical artefact, a good example of this is the Beowulf manuscripts. The Beowulf manuscripts is a collection of a single poem by an English book seller Richard Price. The project of digitising the collection began in 1993

Unfortunately many heritage artefacts have yet to be digitized and archive; like the Timbuktu manuscripts. The manuscripts are currently located in with many lost or damaged due to insects, moisture, poor storage and general neglect. There are currently 408 collections and if not for the damage and disappearance there would have been millions [Haidara 2008]. Currently there are physical libraries holding the manuscripts, mostly in the form of families providing them for education purpose [Haidara and Taore 2008]. There is a clear trend of loss and damage, this is why the concept of digital heritage archives has become so important. If Timbuktu do not digitize the artefacts it runs the risk of being lost for good.

6. CONCLUSIONS

There is no single perfect way to create a digital heritage archive. A combination of tools and practices can produce the most efficient system dependent on the user needs. To begin the user needs and usability of the system should be considered as a basis of decision making. What services being offered and how should be considered and which tools offer which possibility of services and amount of customisation available. The layered architecture would be the simplest and most efficient architecture to implement. Depending on how experienced the administrator is any of the tools described could be used to make an efficient archive. For more customisation Fedora or Omeka allow more room for innovation. The Nelson Mandela archive is a good example of a well implemented archive and it was made by the Google Cultural Institute, so a simple user-friendly archiving tool can offer even the most experienced digital archivist an efficient basis for creating their archive. Digital Heritage preservation has become of great importance, especially within an African context. the best practices need to be used to ensure sustainability and prevent the risk of the cultural heritage being lost forever.

REFERENCES

Anon. 2015a. About - Nelson Mandela Centre of Memory. (2015). Retrieved April 28, 2015 from http://archive.nelsonmandela.org/project-info

Anon. 2015b. Explore - Google Cultural Institute. (2015). https://www.google.com/culturalinstitute/home

Edward A. Arms, William Y and Blanchi, Christophe and Overly. 1997. An architecture for information in digital libraries. *D-Lib Mag.* 3, 2 (1997).

A. Blandford and G. Buchanan. 2003. Usability of digital libraries: a source of creative tensions with technical developments. *IEEE-CS Tech. Comm. Digit. Libr. on-line Newsl.* 1, 1 (2003).

Jérôme Caffaro and Samuele Kaplun. 2011. *Invenio: A modern digital library system for grey literature*,

C. Concordia, S. Gradmann, and S. Siebinga. 2010. Not just another portal, not just another digital library: A portrait of Europeana as an application program interface. *IFLA J.* 36, 1 (April 2010), 61–69. DOI:http://dx.doi.org/10.1177/0340035209360764

R. Daniel Jr, Carl Lagoze, and S.D. Payette. 1998. A Metadata Architecture for Digital Libraries. In *Proceedings of the Advances in Digital Libraries Conference*. IEEE Computer Society, 276–288. DOI:http://dx.doi.org/10.1109/ADL.1998.670428

Dublin Core Metadata Initiative Dcmi. 1995. Dublin Core Metadata Element Set. *Online Comput. Libr. Cent.* (1995).

Abdel Kader Haidara. 2008. The state of manuscripts in Mali and efforts to preserve them. In *The Meanings of Timbuktu*. Cape Town: Open UCT, 265–269.

Ismaël Diadié Haidara and Haoua Taore. 2008. The private libraries of Timbuktu. In *The Meanings of Timbuktu*. Cape Town: Open UCT, 271–276.

Rose Holley. 2010. Trove : Innovation in Access to Information in Australia. *Ariadne* , 64 (2010), 1–9.

ITHAKA. 2015. Home - ALUKA. (2015). https://www.aluka.org/

Robert Kahn and Robert Wilensky. 1995. A Framework for Distributed Digital Object Services. *Int. J. Digit. Libr.* 6, cnri.dlib/tn95-1 (1995), 115–123.

Nils Körber and Hussein Suleman. 2008. Usability of digital repository software: A study of DSpace installation and configuration. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* 5362 LNCS (2008), 31–40. DOI:http://dx.doi.org/10.1007/978-3-540-89533-6-4

Jason Kucsma, Kevin Reiss, and Angela Sidman. 2010. Using Omeka to Build Digital Collections: The METRO Case Study. *D-Lib Mag.* 16, 3 (2010), 2.

Carl Lagoze, Sandy Payette, Edwin Shin, and Chris Wilper. 2005. Fedora: an architecture for complex objects and their relationships. *Int. J. Digit. Libr.* 6, 2 (December 2005), 124–138. DOI:http://dx.doi.org/10.1007/s00799-005-0130-3

S. Navrud and RC Ready. 2002. *Valuing cultural heritage: applying environmental valuation techniques to historic buildings, monuments and artifacts*, Edward Elgar Publishing.

Lighton Phiri and Hussein Suleman. 2015. Chapter 6 Managing cultural heritage : information systems architecture. In *Cultural Heritage Information: Access and Management*. Great Britain: Facet Publishing, 113–133.

M. Pickover. 2008. The DISA Project. Packaging South African heritage as a continuing resource: content, access, ownership and ideology. *IFLA J.* 34, 2 (June 2008), 192–197. DOI:http://dx.doi.org/10.1177/0340035208092177

Ivana Podnar, Toan Luu, Martin Rajman, Fabius Klemm, and Karl Aberer. 2006. A Peer-to-Peer Architecture for Information Retrieval Across Digital Library Collections. In *10th European Conference on Research and Advanced Technology for Digital Libraries*. Springer, 14–25.

Jon Purday. 2009. Think culture: Europeana.eu from concept to construction. *Electron. Libr.* 27, 6 (November 2009), 919–937. DOI:http://dx.doi.org/10.1108/02640470911004039

Christopher Saunders. 2005. Digital imaging South Africa (DISA): a case study. *Progr. Electron. Libr. Inf. Syst.* 39, 4 (2005), 345–352. DOI:http://dx.doi.org/10.1108/00330330510627962

Tom Scheinfeldt. 2008. Omeka: Open Source Web Publishing for Research, Collections and Exhibitions. *Open Source Bus. Resour.* , December 2008 (2008).

Sertan Seales, W Brent and Crossan, Steve and Yoshitake, Mark and Girgin. 2013. From assets to stories via the Google Cultural Institute Platform. In *2013 IEEE International Conference on Big Data*. Santa Carla: IEEE, 71–76.

MacKenzie Smith et al. 2003. DSpace. *D-Lib Mag.* 9, 1 (January 2003). DOI:http://dx.doi.org/10.1045/january2003-smith

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*. 1–9.

Hussein Suleman. 2007. Digital libraries without databases: The Bleek and Lloyd collection. *Res. Adv. Technol. Digit. Libr. Proc.* 4675 (2007), 392–403. DOI:http://dx.doi.org/10.1007/978-3-540-74851-9_33

Hussein Suleman, Kevin Feng, and Gary Marsden. 2006. *Customising Interfaces to Service-Oriented Digital Library Systems*,

Robert Tansley et al. 2003. The DSpace institutional digital repository system: current functionality. In *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*. Washington: IEEE Computer Society, 87–97.

Jernej Trnkoczy, Žiga˘. Turk, and Vlado Stankovski. 2006. A grid-based architecture for personalized federation of digital libraries. *Libr. Collect. Acquis. Tech. Serv.* 30, 3-4 (2006), 139–153. DOI:http://dx.doi.org/10.1016/j.lcats.2006.12.004

C. Vilbrandt et al. 2004. Cultural Heritage Preservation Using Constructive Shape Modeling. *Comput. Graph. Forum* 23, 1 (March 2004), 25–41. DOI:http://dx.doi.org/10.1111/j.1467-8659.2004.00003.x