# Literature Review
# Archive Collection Management of Cultural Artefacts: The Bleek and Lloyd Dictionaries

Lebogang Molwantoa
Department of Computer Science
University of Cape Town
mlwleb001@uct.ac.za

**ABSTRACT**

The Bleek and Lloyd Collection is a set of documents that preserve the history and culture of the early Bushman inhabitants of the Western Cape of South Africa. The Bleek and Lloyd dictionaries project aims to integrate a dictionary into the existing collection, which can be used as a reference by people accessing the collection. This project is an attempt at preserving the language of the Bushman people, whose culture is rapidly declining. This literature review discusses the need to present digitized artefacts in a dynamic, accessible and usable archive, ensuring that it remains intact for future generations. The main focus of this paper is on digital archive systems. A discussion of the available technologies used to preserve digital artefacts in an archive is given. XML-based archive solutions are recommended due to their extensibility and their expected long term preservation of data

## 1. INTRODUCTION

South Africa is a country that is rich in culture and heritage. It is thus important to preserve the culture and artefacts associated with the heritage, possibly in a digital format. Digital preservation also must consider issues of accessibility [8] – that is ensuring that the artefacts are stored in such a manner that they can be accessed easily by a novice and still provide an educational framework for people wanting to learn more about South Africa's history.

The Bleek and Lloyd Collection is a set of documents that preserve the language and culture of the |Xam and !Kun groups of Bushman people [1]. These documents have been meticulously scanned and hyperlinked to provide access to researchers all over the world to what is arguably one of the oldest known cultures. The digitisation of these documents is important in preserving the history and heritage of the Bushman people – especially when considering the rapid decline in Bushman culture as a result of Western influences.

The most recent part of this collection to be digitized is a set of approximately 40000 scans corresponding to a dictionary that may be used to interpret and understand the original texts. The aim of this project is to integrate this dictionary into the main collection so it can be used as a live reference by researchers, while also preserving its contents for future generations.

The proposed system that will implement the dictionary is split into 3 components to be integrated. These consist of an archive that will store images and metadata associated with the images, functionality to search and browse through the archive and a feature that allows for translation via image processing. This paper will focus on digital preservation archives and related technologies used to archive digital cultural artefacts.

The rest of this paper presents a high level view of digital preservation and cultural preservation in Section 2. Section 3 describes related systems and projects that are concerned with the preservation of cultural artefacts. Section 4 describes technologies that are used to build digital archives or digital repository systems that store and maintain cultural artefacts.

## 2. DIGITAL PRESERVATION

### 2.1 Definition

Digital preservation can be defined as the set of processes and activities that ensure continued access to information [8] - all kinds of records, scientific and cultural heritage - existing in digital formats. The preservation of digital entities requires data management technologies that are provided by digital libraries and data grids. Digital archives are dedicated to the long-term preservation of data with the directive to ensure that they capture and preserve the data in a manner such that it can be accessed and presented at any time [11].

### 2.2 Motivation for digital preservation

The preservation of data in a digital format is important in order to make content easily and readily available. Digital archives provide preservation environments that assure the authenticity and integrity of digital entities [7].

### 2.3 Strategies deployed in digital preservation

The Online Computer Library provided a 4-pronged approach in addressing the need to digitally preserve data [18]. Their approach was:

1. Assessing the risk presented by technology.
2. Providing access to the digital content.
3. Determining and attaching the appropriate metadata to the digital content.
4. Determining what type of digital format should be applied.

Other strategies that are employed when preserving data are [18]:

1. Refreshing – the transfer of data between two types of storage medium.
2. Migration – the transferring of data to a newer system environment.
3. Replication – the creation of duplicate copies of data on one or more systems.
4. Emulation –the replication of the functionality of an obsolete system.

### 2.4 Digital Preservation Standards

The Open Archival Information System (OAIS) model provides an architecture for conducting digital preservation research and experimentation [9]. The OAIS model consists of an organisation of people and systems whose mission is to ensure that information is preserved and is accessible to the community.

### 2.5 Language and cultural preservation

The preservation of cultural artefacts and digital documents forms a great portion of what is entailed in information preservation. A lot of research has been conducted on methods of preserving cultural heritage. Such initiatives will be discussed in section 3.

## 3. RELATED WORKS

Cultural heritage in many areas around the world is endangered, mainly due to the overwhelming influence of Western civilization, ideals and lifestyles [4]. A problem in a lot of cultures especially in Africa is that cultural heritage is not preserved and is in danger of being destroyed by degradation, inaccessibility or even natural disasters. Thus, there is a clear need to digitize and archive these cultural artefacts.

The CAMA (Contemporary African Music and Art Archive) is an archive that aims to digitally capture as much of contemporary Africa culture as possible [10]. This is done through the usage of camcorders, digital cameras and audio recorders. The CAMA project also aims to ensure that the archive is accessible to everyone as well as building a system which can present African art in a meaningful way [10].

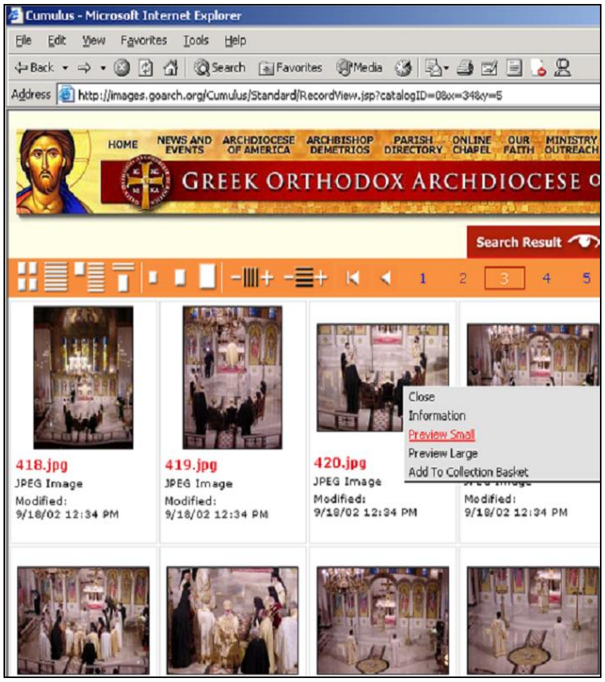
Many museums and libraries digitize their collections of historical artefacts to preserve them and also to make them accessible. A good example of this is the Armarius archive, which is an online document management system for ancient manuscripts [5]. The Armarius archive digitises historical documents in a dynamic archive that can be accessed by anyone. The Armarius archive does this by storage in a database, document structuring as well as providing a platform to access the collection. Some of the collections that are found in the archive are the Arabic ancient manuscripts found in Timbuktu, manuscripts from mathematicians in the 14th century as well as Syrian manuscripts. In addition, the archive uses an online annotation service.

The Travellers in the Middle East Archive (TIMEA) [6] is a digital archive that enables users to understand the explorations in the Middle East in the period between the 18th and 20th centuries. The TIMEA archive aims at enabling wide access to cultural heritage material while simultaneously promoting research skills amongst users of the archives – who are mainly historians and scholars. TIMEA is currently providing access to a growing collection of images and pager of encoded text. [6]. The archive integrates already existing technologies – GIS maps, digital asset management software called DSpace for texts and images as well as Connexions which encompasses contextual research and teaching material.

The Greek Orthodox Archdiocese of America (GOA) has a rich and varied collection of important artefacts that are in the form of historical iconography, art, letters and memorabilia [2]. Many of these artefacts are in a fragile state and cannot be handled by the many history scholars who wish to study them – an example is a lot of the church letters are written on very fragile onion skin paper. As a result, through the Department of Internet Ministries, the GOA has undertaken the project of digitising these artefacts with the main purpose of making them readily available for appropriate purposes. The artefacts are used mainly by theology scholars and historians who are interested in studying these artefacts. Figure 1 displays the Web interface available to a user who accesses the GOA archive.

There are many other initiatives that preserve languages and cultures. The Canadian Heritage Information Network's (CHIN) Virtual Museum is an online digital library that collects the contents of Canadian Museums and makes it available for the public to use [9].

North Carolina's Exploring Cultural Heritage Online (ECHO) [9] promotes the use of digital technologies in order to broaden and enhance access to the cultural heritage of the state of North Carolina as well as to encourage collaboration between all other states' cultural resource institutions. ECHO is an online portal to other online special collections of North Carolina's archives, museums and libraries [12].



418.jpg
JPEG Image
Modified:
9/18/02 12:34 PM

419.jpg
JPEG Image
Modified:
9/18/02 12:34 PM

420.jpg
JPEG Image
Modified:
9/18/02 12:34 PM

Modified:
9/18/02 12:34 PM

**Figure 1: Web interface for a user navigating the GOA archives**

The many archives in existence that preserve cultural artefacts further justify the need to preserve these artefacts as a means to ensure that they last for future generations. All of the above mentioned systems use archive management software and allow users to readily access the collection of digital cultural artefacts. Some systems such as TIMEA allow users to engage thoroughly with the ancient manuscripts through the integration of GIS maps and contextual material. However, none of these systems integrate a dictionary that can be used as a live reference by researchers and other people who access the archives.

It is on this very basis that the idea was formed of integrating an online dictionary as a live reference for scholars who access the Bleek and Lloyd Dictionaries. The integrating of this dictionary will be done through digital archives and in the next section a discussion of related software technologies is presented.

## 4. SOFTWARE TECHNOLOGIES

### 4.1 Digital Repository Systems

Hardware and software technologies evolve more rapidly than physical media decay, and this is one of the major challenges faced by archivists of digital information – a phenomenon referred to as technological obsolescence of the infrastructure that is used to access and present the information that is archived [11].

A digital repository system is software that is used to build a digital archive and provide services that help to manage and organise the repository. There are different forms of digital repository systems to manage the wide variety of digital objects in a way that is most suited to that object.

### 4.1.1 DSpace

DSpace is an open-source, cross-platform software package, written in Java that provides tools for the management of digital assets [13]. DSpace preserves and enables easy and open access to all types of digital content including text, images, moving images and data sets [13]. There are numerous benefits to using DSpace. These include:

- Providing a platform for long term storage of digital material.
- Reaching a wider audience through exposure to search engines such as Google.
- Having a persistent network identifier for your work that never changes or breaks.

DSpace is implemented in Java and JSP, using the Java Servlet API. It also supports the use of relational databases such as Oracle and PostgreSQL, as well as ApacheHTTPD for certificate support. It supports the Open Archives Initiative protocol for Metadata Harvesting which is a protocol developed by the Open Archives Initiative(OAI) [14] to collect the metadata descriptions of the records in an archive so that services can be built using metadata from many archives [14]. The latest stable release of DSpace is version 1.5.2.

### 4.1.2 Fedora

The Flexible Extensible Digital Object and Repository Architecture (Fedora) is a general-purpose, open-source digital object repository system [15]. Fedora is an open source system for the collection and management of different types of digital objects [15]. Fedora is built on the principle that the best way of integrating data and interfaces - as distinct modules - is by using the principles of interoperability and extensibility.

Fedora is not a complete application with all indexing, querying and discovery applications of a digital repository; it is merely a framework upon which other systems may be built. Fedora provides a general purpose management layer for the management of digital objects. The key features of the Fedora system are:

- The support of heterogeneous data types and being able to adapt to new data types.
- The ability to specify multiple content disseminations of digital objects.
- Associating rights management schemes with these disseminations.

The Fedora architecture is divided into four subsystems and a Web Services layer. The core subsystem layer consists of the management subsystem which manages all the operations on the digital objects and an access subsystem that implements the operations that are necessary for disseminating objects and discovering more information and behaviours for an object.

Fedora allows for the interchange between Fedora and XML-based applications and this mechanism facilitates archiving. Fedora supports the import and export of digital objects in a variety of XML formats.

### 4.1.3 Greenstone

Greenstone is a suite of software for building and distributing digital library collections. It provides way of organising information and publishing it on the Internet or on CD-ROM [15]. According to the Greenstone Digital

Library Software website, the aim of the software is to empower users, particularly in universities, libraries and other public service institutions, to build their own digital libraries.

Greenstone is easy to use and the usage of the system is made easier through the Greenstone Librarian Interface (GLI) [19] – which is shown in Figure 3. Greenstone is capable of building up multi-media digital documents such as text, PDF, audio and video [19]. The text, PDF, HTML and similar documents are converted into Greenstone Archive Format (GAF) which is an XML equivalent format.

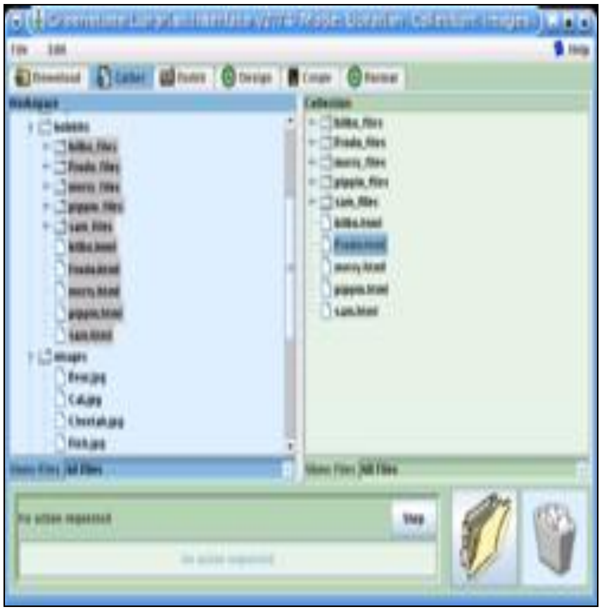


**Figure 2: Greenstone's Graphical Librarian Interface (GLI) v2.80**

A problem which was highlighted in a paper on the creation of an archive for the Bleek and Lloyd collection [1] is that Greenstone may present a portability problem as it requires some basic software installation, and this may become an issue when the system is meant to work on any arbitrary system.

### 4.2 Other Digital Library Technologies

#### 4.2.1 Relational Database Models

The relational database has a naturally close relationship with many Digital Library Systems such as DSpace and Fedora, which by default use mySQL and Postgres respectively[13],[15] to hold their primary metadata repositories. The relational database model provides a useful platform for a basic mechanism which enables insertion, removal and updating of an archive. A well established query structure as well as efficient existing operations on the database systems provide a good argument in favour of the use of database systems for metadata storage in digital archives.

But databases also present another range of problems. In a paper on Digital Libraries Without Databases: The Bleek and Lloyd Dictionaries[1], the author summarises some of the problems present in a database archive system, versus an XML-centric approach. A problems with a database system is that of the database not being platform-independent. This can be a problem on an arbitrary system, where the data needs to be extracted before it can be processed. In addition there is often a need for an administrator to run the database, which usually means a heavy reliance on this person.

#### 4.2.2 XML

An XML-based archival infrastructure complies with the notion of not requiring special access software and is open and simple to use for both humans and programs(via parsing). Because of user-defined tags and the fact that some documents contain some schema information in the structure of their parse trees, XML can be viewed as a generic and self-describing data format [11].

XML-centric solutions have been recommended in the preservation of heritage-based digital collections because of their expected long term method of preserving data [1]. The big concerns with using XML-based digital archives are the issues of scalability as well as the customisation of interfaces to these archives – allowing the user of a heritage-based digital archive to customise their interface to facilitate browsing.

#### 4.2.3 Java Content Repository (JCR) API

The Content Repository API for Java (JCR) is a specification for the Java platform that allows access to contents in a repository in a standard manner [17]. The JCR is defined as an object database with searching, storage and retrieval features. JCR can be found in content management systems as well as in the storage of metadata.

The data in the JCR is stored in a tree data structure consisting of Nodes with associated properties. Data is stored in the properties which hold arbitrary length binary data and strings. Queries in JCR are performed using XPath and it also has the ability to support some standard form of SQL.

## 5. CONCLUSION

This paper has looked at the available archive systems that are used to preserve cultural artefacts. This has paper has also looked at the available technologies that are used to implement digital repositories and found that XML-based solutions are recommended due to their extensibility and their expected long term preservation of data – which is important in building archives to preserve cultural artefacts. Based on the findings in the literature review, an XML-based repository system like Fedora will most likely be used for the implementation of the digital archive.

## 6. REFERENCES

[1] Suleman, H. 2007. Digital Libraries Without Databases: The Bleek and Lloyd Collection. In Kovács, László, Norbert Fuhr and Carlo Meghini, Eds. Proceedings Research and Advanced Technology for Digital Libraries, 11th European Conference (ECDL 2007), pages 392-403, Budapest, Hungary.

[2] Nicolakis, T., Pizano, C. E., Prumo, B., and Webb, M. 2003. Protecting digital archives at the Greek Orthodox Archdiocese of America. In *Proceedings of the 3rd ACM Workshop on Digital Rights Management* (Washington, DC, USA, October 27 - 27, 2003). DRM '03. ACM, New York, NY, 13-26.

[3] Natu, S. and Mendonca, J. 2003. Digital asset management using a native XML database implementation. In *Proceedings of the 4th Conference on information Technology Curriculum* (Lafayette, Indiana, USA, October 16 - 18, 2003). CITC4 '03. ACM, New York, NY, 237-241.

[4] Liu, J. and Tseng, M. 2004. Mediating team work for digital heritage archiving. In *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries* (Tuscon, AZ, USA, June 07 - 11, 2004). JCDL '04. ACM, New York, NY, 259-268.

[5] Doumat, R., Egyed-Zsigmond, E., Pinon, J., and Csiszar, E. 2008. Online ancient documents: Armarius. In *Proceeding of the Eighth ACM Symposium on Document Engineering* (Sao Paulo, Brazil, September 16 - 19, 2008). DocEng '08. ACM, New York, NY, 127-130.

[6] Spiro, L., Wise, M., Henry, G., Bearden, C., Byrd, S., Garza, E., and Decker, M. 2006. Enabling exploration: travelers in the middle east archive. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries* (Chapel Hill, NC, USA, June 11 - 15, 2006). JCDL '06. ACM, New York, NY, 163-164.

[7] Zorich, D., 2003, A Survey of Digital Cultural Heritage Initiatives and Their Sustainability Concerns, Council On Library and Information Resources, Washington D.C.

[8] Moore, R. W. and Marciano, R. 2005. Building preservation environments. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries* (Denver, CO, USA, June 07 - 11, 2005). JCDL '05. ACM, New York, NY, 424-424.

[9] Ray, J., Dale, R., Moore, R., Reich, V., Underwood, W., and McCray, A. T. 2002. Panel on digital preservation. In *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries* (Portland, Oregon, USA, July 14 - 18, 2002). JCDL '02. ACM, New York, NY, 365-367.

[10] Marsen, G., Malan, K., and Blake, E. 2002. Using digital technology to access and store African art. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems* (Minneapolis, Minnesota, USA, April 20 - 25, 2002). CHI '02. ACM, New York, NY, 528-529.

[11] Ludäscher, B., Marciano, R., and Moore, R. 2001. Preservation of digital data with self-validating, self-instantiating knowledge-based archives. *SIGMOD Rec.* 30, 3 (Sep. 2001), 54-63.

[12] North Carolina ECHO, Exploring Cultural Heritage Online, http://www.ncecho.org/about/index.shtml, Date Last Viewed: 2009-05-01

[13] Baudoin, P., M. Branschofsky. 2004.
MIT's DSpace experience: a case study.

[14] Open Archives Initiative Protocol for Metadata Harvesting, http://www.openarchives.org/pmh/
Last Viewed: 2009-05-01

[15] Greenstone Digital Library Software http://www.greenstone.org/
Last Viewed: 2009-05-01

[16] Suleman, H., Feng, K., and Marsden, G.. 2006. Customising Interfaces to Service-Oriented Digital Library Systems. In *Proceedings 9th International Conference on the Asian Digital Library*, Kyoto, Japan.

[17] The Java Community Process Programme, http://jcp.org/en/jsr/detail?id=283
Last Viewed: 2009-05-01

[18] OCLC Global Gateway [OCLC]
http://www.oclc.org/uk/en/global/default.htm
Last Viewed: 2009-05-04

[19] Greenstone (Software) – Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Greenstone_
(software)
Last Viewed: 2009-05-06