An African Perspective on Digital Preservation

Hussein Suleman
Department of Computer Science, University of Cape Town

Abstract

The preservation of culture, language and history is recognised as being of crucial importance for future and current generations. Africa, especially, has the unique problem of needing to archive many collections of crucial significance to the world, yet without the skills or funds that often are utilised in this pursuit. The Bleek and Lloyd collection, that documents Bushman culture, and the DISA collection, that documents the South African liberation struggle, are examples of such projects. In both cases, and in other similar projects, there is a growing need for approaches to digital preservation, tools and techniques that better suit the needs and environment of developing countries. Some very specific solutions have emerged to solve problems related to Internet bandwidth and information accessibility, and these will be discussed, but these problems are only recently being acknowledged. Current digitisation efforts still outstrip efforts to organise information. Future technical solutions will therefore need to be found to maintain the pace of preservation as well as ensure ongoing access to all the peoples of the world.

Keywords

Africa; digital preservation; developing world; curation; dissemination

Motivation for Preservation

Digital preservation is motivated by different factors in different parts of the world. UNESCO states in its charter on preservation of digital heritage that access to heritage is one of the motivators for preservation and further stresses that cultural heritage should be made accessible to all the people of the world (UNESCO 2003).

In the global north-west (Europe and USA) there have been numerous projects related to different aspects of preservation. PREMIS develops metadata to support preservation (Caplan and Guenther 2005). PLANETS focuses on preservation planning and tools to support it (Becker et al 2007). CASPAR is about a European infrastructure for preservation based on the Open Archival Information System model (Giaretta 2006). Trustworthiness and auditing of repositories is the central theme for yet other efforts (OCLC 2007). These are among some of the key issues for heritage preservation. However, there are other very specific and practical concerns that affect developing countries but not the northern counterparts.

Africa, in particular, is made up almost solely of developing countries. While these are fairly heterogeneous, there are some problems and solutions that transcend country borders to create a shared context for heritage preservation activities. In this context, a prime motivator for heritage preservation is the large number of significant collections that are of great value and in need of curation and dissemination. Unlike the LOCKSS focus on dark archives and more recent work on automated migration (Rosenthal et al 2005), heritage preservation in these cases focuses more closely on archiving and access.

A number of example collections are presented along with the motivation for preservation in each case.

The Mapungubwe Collection (SouthAfrica.info 2008; Huffman 2000) is a collection of archaeological artefacts from the northern parts of South Africa that indicate the existence of a sophisticated culture in the region during the Iron Age. According to researchers, the community in this area included the

notions of class and evidence indicates that there were artisans and craftspeople within the community. While the civilization did not flourish beyond the 1300s, it is now recognised as the earliest advanced civilization in Southern Africa. This find was made during the Apartheid era of South African history and the information was intentionally kept from people because it was evidence that Black South Africans were not descendant from savages. Thus, it is vitally important that this collection be preserved and made accessible to people to correct the misinformation of the past.

The Timbuktu Manuscripts (Minicka 2006) are a collection of approximately 700000 documents found in Mali that document science and medicine as studied and practiced in the local community around the year 1300. This is a significant piece of evidence of a modern written knowledge system in Africa before European colonisation. As such, the manuscripts are currently being physically and digitally preserved.

The Kirby Collection (University of Cape Town 2007) contains African musical instruments from the early 1900s in South Africa. This provides a useful non-textual insight into local cultures. Many instruments that form part of this collection are already obsolete.

Digital Imaging South Africa at the University of Kwazulu-Natal (Pickover and Peters 2002) has as its flagship project a collection of magazines and newspapers that document the struggle for liberation in South Africa. This collection forms as alternative source of information on historical events. Unlike the previous examples, this collection is completely digitized and available online with browse and search interfaces.

The Bleek and Lloyd archive (Skotnes 2007) includes artefacts and documents related to Bushman culture in the south-west parts of Africa. The Bushman groups are recognised as being among the oldest cultures in the world, so whatever is known about their culture needs to be preserved. A large part of this culture was transmitted orally, and the there are few storytellers in modern 21st century society, hence the urgency to document the culture. The Bleek and Lloyd collection includes notebooks with transcriptions of stories and dictionaries as well as annotated drawings, all of which serve at the very least as a guide to the rock art in the region.

Finally, there are numerous new collections being created within institutional repositories and digital archives in public institutions and institutions of higher learning. While the need to preserve at-risk information is paramount, the new knowledge that is being created on an ongoing basis must be part of any digital preservation strategy.

Why an African Perspective

An African perspective on preservation ought not to be different from other perspectives. However, digital preservation is often discussed in terms of technology, infrastructure and practices in the global north-west. Africa is largely composed of developing nations and thus has particular problems, the solutions to which can potentially influence global practices. The following are examples of issues that need to be considered in the African context.

Artefact Deterioration

Some documents and storage media (e.g., in the Timbuktu Manuscripts) are rapidly deteriorating and oral history (e.g., in the Bushman cultures) is no longer being maintained in modern society. As a result, there is an urgent need to preserve information about ancient cultures which would otherwise be lost. There was no written language in many parts of Africa. Early researchers, like Wilhelm Bleek, transcribed oral histories to paper but these documents are themselves in danger of deterioration.

Rewriting History

We now know that there were significant ancient civilizations in many parts of Africa. Up until recent

times, this information was suppressed by colonial governments to maintain intellectual superiority over local people. Part of the process of rebuilding a post-colonial world includes educating people about the important contributions of every culture – for this the untainted history must be restored. Unfortunately, in many instances there is little evidence of this history left and preservation of what is left is therefore more critical.

Skills and Education

Archivists in African institutions are arguably not as technically skilled as their counterparts in other parts of the world. The availability of computer systems in some parts of the continent has the effect that curators of information do not receive sufficient training in electronic systems. Digital media is not the norm for many forms of communication and information storage.

The level of education of the general population in many African countries also is a problem. The number of literate individuals, as well as the number of individuals with access to a computer and the Internet is lower than elsewhere in the world. This creates a challenge for digital preservation both in terms of collection building, especially for end-user submissions, and dissemination. Novel solutions are needed for both these problems to make digital archives effective.

Funding

Typically, there is little funding for digital preservation, and heritage preservation in general, because of other priorities in many countries. Many preservation projects receive funding from external international agencies (e.g., Mellon Foundation, Ford Foundation) but there are sometimes restrictions placed by the agencies that limit who the data may be shared with.

In general, there is a need to do more with fewer resources – sharing and reuse of resources are critical parts of any solution.

Internet Bandwidth (Digital Divide)

The digital divide is still a major hindrance. In many parts of Africa there is little access to computers and the Internet. In those parts where there is Internet access, the resources, such as bandwidth, are severely limited or extremely expensive. As of December 2007, only 4.7% of Africans have access to the Internet, which is well below every other continent (Internet World Stats 2007).

Some digital preservation systems, such as LOCKSS, have questionable applicability. In the case of LOCKSS, a group of sites collaboratively maintain the integrity of collections. LOCKSS, however, does not cater for unstable and irregular bandwidth availability – its algorithms will not make the most efficient use of bandwidth and may exacerbate problems at sites with poor bandwidth.

All online archives need to make use of bandwidth in a way that is both minimal and cognizant of the differences among sites.

Why not a Global Perspective

Africa is not special in terms of the problems that affect digital preservation or preservation of heritage. Some communities in other developing countries face many of the same problems and many communities face some of the same problems. However, most communities, be they developing world or not, can benefit from the solution to these problems – e.g., greater efficiency, increased robustness of systems, lower costs.

Techniques and Solutions

The following techniques and solutions have been adopted in African projects to preserve heritage, in an attempt to deal with the identified problems.

Lightweight Systems

In order to build a lightweight system, simplicity is a key criterion. If the systems and processes are easy to implement, they will be easy to maintain and migrate in future when other systems are put into place.

XML can be used as an underlying data storage mechanism. While this creates some inefficiencies in accessing data collections – as opposed to databases – there are numerous advantages. These include that XML is human-readable, structured and can be stored in flat files which do not require software intermediaries.

Archives should be minimalist to reduce the overhead of maintaining systems. In many cases, metadata is being managed in spreadsheets as this is a simple technique which does not require that data capturers are re-trained and there is no need for a metadata repository until the data needs to be made discoverable.

Static collections of digital objects are easier to preserve than dynamically-managed or mediated objects because only the data needs to be preserved and not the service (de Lusenet 2002). The files can simply be copied and archived. In contrast, a service that mediates access to digital objects must be preserved along with the manifestations of the digital objects.

Repositories can be made multi-purpose so that a single installation of the core software systems can be shared across multiple conceptual repositories. This serves as an enabler of collection development where the hardware, software and human resources are shared among collections. This approach is sometimes frowned upon but works well where resources are limited and archiving software (e.g., Dspace) allows for distinct sub-divisions within the collections.

Skills also can be shared, thus creating less work within a community. Communities of practice in the same geographical area have resulted in training and support networks in repository management, especially from the Open Access perspective.

Bandwidth

Collections can be made accessible over different media. Greenstone supports the creation of CD-ROM versions of its collections (Witten and Bainbridge 2003). In additions, collections can be shared over a network or on removable media such as USB drives. These approaches reduce the need for bandwidth or eliminate it altogether. In addition, preserving the bits in such cases is a matter of copying files rather than invoking software services – the former can be accomplished using any backup and archiving tools.

In order to minimize bandwidth, client-side services can be created that run within the end-user's Web browser e.g., using Asynchronous JavaScript and XML (AJAX). Such services can perform a wide range of useful tasks, with only minimal interaction with the server instead of loading of new pages for each request. Services could be invoked off a local device, thus operating without any form of Internet connection.

Case Study: Bleek and Lloyd Collection

The Bleek and Lloyd collection includes a set of books and drawings documenting the near-extinct culture of the |xam and !kun Bushman groups of Southern Africa. They are widely acknowledged as

descendents of the first people on earth and elements of their culture are of interest to researchers and scholars.

The notebooks that this collection revolves around were produced by Wilhelm Bleek, Lucy Lloyd and others in the late 1800s in Cape Town. They were the result of interviews with prisoners from the local prison, who belonged to various Bushman groups. These stories and other pieces of information were transcribed into over 14000 pages of text (see Figure 1) with the original text in many cases juxtaposed with translations thereof. Thus, the notebooks serve as a Rosetta stone of the Bushman language and culture! Accompanying the notebooks are a collections of approximately 800 annotated drawings that, for example, can help to explain the meaning of Bushman art, particularly rock art.
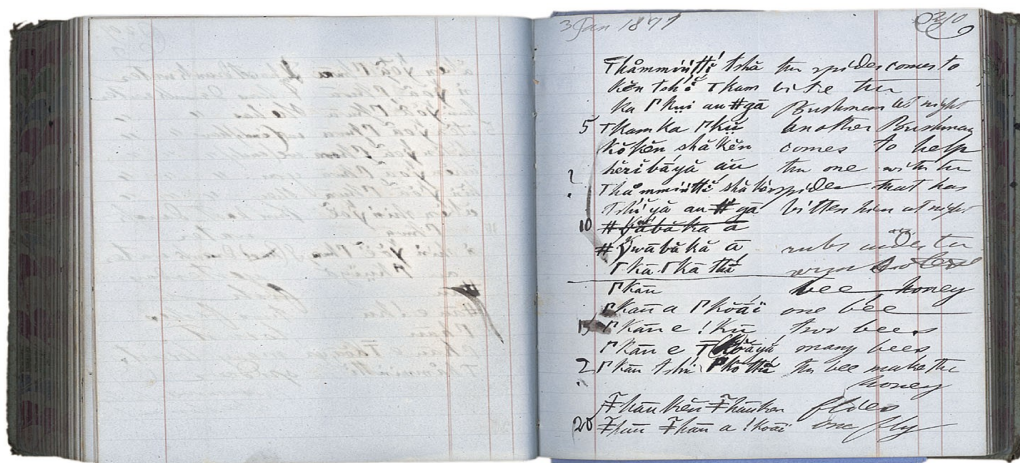

Figure 1: Sample page from Bleek and Lloyd notebook

This collection has been placed by UNESCO on its Memory of the World register and it is jointly curated by a group that includes the University of Cape Town, National Library of South Africa, Iziko-South African Museum and the University of South Africa. Digitization of the collection was funded by the Mellon Foundation and De Beers and this effort was led by the Michaelis School of Fine Arts at the University of Cape Town.

After digitization, the digital objects and metadata were to be assembled into a system that would allow the greatest possible access. In particular, the following were deemed as requirements:

- The collection needed to be accessible from a wide range of storage devices: over the Web, off a CD-ROM drive, off a local or network drive, etc.

- The collection needed to be accessible irrespective of the hardware and software platform (Mac/MS-Windows/Linux/etc.).

- The system needed to be usable with minimum end-user barriers (such as requiring software installation).

- The system needed to comply with standards wherever possible.

All metadata was first created and edited in Microsoft Excel, and thereafter converted to XML.

The first software solution considered was Greenstone, as it is the only digital library toolkit with a CD-ROM distribution option. Unfortunately, Greenstone requires software to be installed on the target computer and it does not work on all platforms. The second software option considered was to create an XML Stylesheet Language Transformation (XSLT) stylesheet to convert the source XML files into PDF using the XSL Formatting Objects (XSL-FO) standard. This worked well for small subsets of the collections as PDF documents have a built-in search facility and the digital objects could be embedded in browsable pages. However, as the collection size was increased, the PDF files became

unmanageable and most PDF viewers would slow down drastically because of the size of the files. Thus, this option too was not feasible.

The technology decided upon was a static XHTML website. The XML metadata was translated into XHTML using XSLT stylesheets. One stumbling block with this approach was that of scalability and this was suitably resolved using a combination of the following techniques (Suleman 2007a):

- The 16000 XHTML documents were split into batches to avoid generating all of them at once. The XSLT engine would consume memory proportional to the number of documents being generated, so this technique placed a bound on the resources required when generating the website.

- Non-trivial XSLT stylesheets using large XML source documents resulted in very slow XPath query resolution. To resolve this, indices and keys were used to speed up XPath queries – in a manner very similar to database indices.

- XSLT does not support queries that cross document boundaries so all the data had to appear in a single XML document.

- To create browseable listings of objects within the collection, it was necessary to group items logically according to various criteria (like author). Grouping is not common in XSLT but the Muenchian Method (Tennison 2007) allows one to efficiently group items by exploiting indices and the semantics of XSLT set operations.

The other major problem was that static websites cannot offer users services such as search engines. This was resolved by using an AJAX-based query system (Suleman 2007b). As in any typical information retrieval system, there are a set of inverted indices – these are created immediately after the XHTML files are generated and stored along with them. Then, when the XHTML pages are viewed, a small AJAX application (see Figure 2) allows the end-user to search using the pre-generated inverted files, without any server-side dynamic operations. This system works identically, therefore equally well, if the collection is distributed on CD-ROM, is on the Web or is on a local drive.
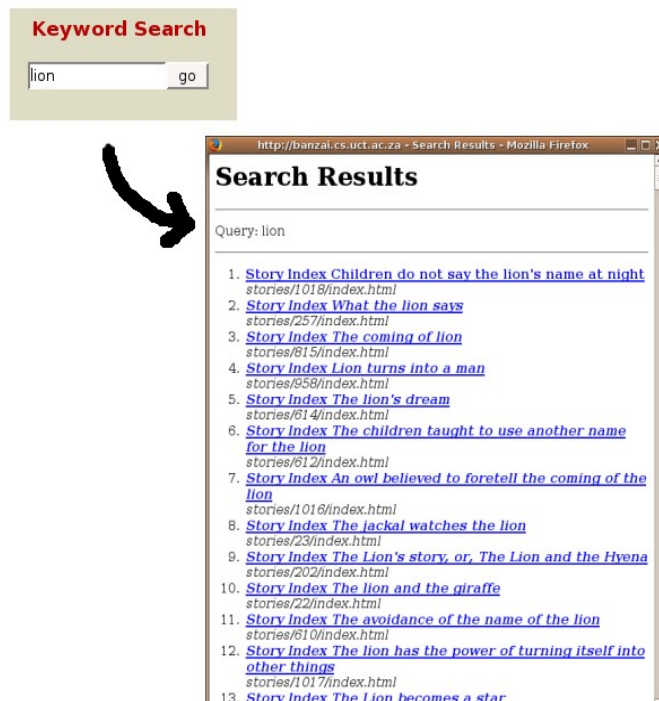

Figure 2: Search engine interface in Bleek and Lloyd collection

Future Challenges

The preservation of heritage collections is faced with many peculiar problems in developing countries. However, experience with current projects indicates that most of these problems can be resolved with appropriate and innovative use of new technologies.

One of the remaining challenges is how to scale the preservation effort itself. Few collections have been curated to date, usually at institutions with some capacity or funding. However, the techniques and tools need to be reusable so that heritage collection curation is as simple as the development of Open Access collections.

The tools for heritage collection management are still somewhat of a hurdle. Greenstone is a popular and well-developed tool so this could possibly be extended to cater for collections with special requirements. Alternatively, special tools could be developed or adapted to manage metadata and digital objects such that they can fulfill the needs of curators and users in Africa who may submit to the collections.

Conclusions

Digital preservation of heritage is an ongoing activity in Africa, with much interest but not enough capacity or resources. African projects typically highlight the special needs of developing countries and the innovative solutions that must be put into place. With such solutions, it is indeed possible to digitize and disseminate heritage collections, but this is not necessarily a repeatable process. Some effort is still needed to create tools and processes that exploit emerging technology, resulting in simple solutions that work across a wide range of contexts.

References

Becker, C., Kolar, G., Küng, J., and Rauber, A. 2007
Preserving Interactive Multimedia Art: A Case Study in Preservation Planning
In *Proceedings of Tenth International Conference on Asian Digital Libraries, ICADL 2007, Hanoi, Vietnam, 10-13 December*
http://publik.tuwien.ac.at/files/pub-inf_5204.pdf (Accessed 16 March 2008)

Caplan, P. and Guenther, R. 2005
Practical Preservation: the PREMIS Experience
Library Trends, 54(1)
http://www.loc.gov/standards/premis/caplan_guenther-librarytrends.pdf (Accessed 16 March 2008)

de Lusenet, Y. 2002
Preservation of digital heritage, Draft discussion paper prepared for UNESCO, European Commission on Preservation and Access
Place: UNESCO
http://www.knaw.nl/ecpa/PUBL/unesco.html (Accessed 16 March 2008)

Giaretta, D. 2006
CASPAR and a European Infrastructure for Digital Preservation
ERCIM News, 66
http://www.ercim.org/publication/Ercim_News/enw66/giaretta.html (Accessed 16 March 2008)

Huffman, T. N. 2000
Mapungubwe and the Origins of the Zimbabwe Culture, pp 14-29
Goodwin Series, 8, African Naissance: The Limpopo Valley 1000 Years Ago

Internet World Stats. 2007
World Internet Users

http://www.internetworldstats.com/stats.htm (Accessed 16 March 2008)

Minicka, M. 2006
Safegaurding Africa's Literary Heritage : Timbuktu rare manuscripts project
In *Proceedings LIASA WCHELIG Winter Colloquium : Collaboration for success, Cape Town, South Africa*
http://eprints.rclis.org/archive/00006729/ (Accessed 16 March 2008)

OCLC. 2007
Trustworthy Repositories Audit & Certification: Criteria and Checklist
Place: OCLC
http://www.crl.edu/PDF/trac.pdf (Accessed 16 March 2008)

Pickover, M. and Peters, D. 2002
DISA: an African Perspective on Digital Technology
Innovation, 24

Rosenthal, D. S. H., Lipkis, T., Robertson, T. S., and Morabito, S. 2005
Transparent Format Migration of Preserved Web Content
D-Lib Magazine, 11(1)
http://www.dlib.org//dlib/january05/rosenthal/01rosenthal.html (Accessed 16 March 2008)

Skotnes, P. 2007
Claim to the Country: The Archive of Wilhelm Bleek and Lucy Lloyd
Auckland Park: Jacana. 388pp

SouthAfrica.info. 2008
Mapungubwe: SA's lost city of gold
http://www.safrica.info/ess_info/sa_glance/history/mapungubwe.htm (Accessed 16 March 2008)

Suleman, H. 2007
Digital Libraries Without Databases: The Bleek and Lloyd Collection, pp 392-403
In *Proceedings of Research and Advanced Technology for Digital Libraries, 11th European Conference (ECDL 2007), Budapest, Hungary, 16-19 September*, edited by Kovacs, L., Norbert F. and Meghini, C.
http://pubs.cs.uct.ac.za/archive/00000433/01/ecdl_2007_dlwd.pdf (Accessed 16 March 2008)

Suleman, H. 2007
in-Browser Digital Library Services, pp 462-465
In *Proceedings of Research and Advanced Technology for Digital Libraries, 11th European Conference (ECDL 2007), Budapest, Hungary, 16-19 September*, edited by Kovacs, L., Norbert F. and Meghini, C.
http://pubs.cs.uct.ac.za/archive/00000434/01/ecdl_2007_ajax.pdf (Accessed 16 March 2008)

Tennison, J. 2007
Using the Muenchian Method
http://www.jenitennison.com/xslt/grouping/muenchian.html (Accessed 16 March 2008)

UNESCO. 2003
Charter on the Preservation of the Digital Heritage
Place: UNESCO
http://portal.unesco.org/ci/en/files/13367/10700115911Charter_en.pdf/Charter_en.pdf (Accessed 16 March 2008)

University of Cape Town. 2007
The Percival R. Kirby Collection of musical instruments
http://web.uct.ac.za/depts/sacm/kirby.html (Accessed on 16 March 2008)

Witten, I. and Bainbridge, D. 2003

How to Build a Digital Library
San Francisco: Morgan Kauffman. 518pp