In Search of Simplicity: Redesigning the Digital Bleek and Lloyd

Lighton Phiri and Hussein Suleman

Department of Computer Science

University of Cape Town

Private Bag X3, Rondebosch, 7701

Cape Town, South Africa

{lphiri,hussein}@cs.uct.ac.za

Abstract

The Digital Bleek and Lloyd is a collection of digitized historical artefacts on the Bushman people of Southern Africa. The underlying software was initially designed to enable access from as many people as possible so usage requirements were minimal – it was not even necessary to use a Web server or database. However, the system was not focused on preservation, extensibility or reusability. In this article it is argued that such desirable attributes could manifest themselves in a natural evolution of the Bleek and Lloyd software system in the direction of greater simplicity. A case study demonstrates that this is indeed feasible in the case of the Digital Bleek and Lloyd and potentially more generally applicable in digital libraries.

Keywords

Digital libraries, digital preservation, developing world, extensibility, reusability, simplicity

1 Introduction

Storage of information in digital form is increasingly becoming popular due to the flexibility and cost-effective nature of storage media. This data proliferation of both born-digital and digitised information has effectively increased the amount of digital information that is at risk of loss due to obsolescence of hardware and software, and physical damage. The process of managing this content and its associated representations over time to guarantee their long-term usability is known as digital preservation. It collectively requires the establishment and implementation of a reliable preservation policy and well-coordinated procedures to ensure that the digital objects are preserved against hardware and/or software obsolescence, media failure and any other potential physical threats posed on the digital objects [14].

The preservation of digital information requires constant attention due to the fast pace at which technology is advancing and because of the delicate nature of digital storage media that is used to store digital content. A typical digital preservation procedure involves activities such as: regularly backing up of data to a different location to ensure its availability in the event of a disaster; refreshing of media by copying the digital information from one media to another; migration of digital content after a change in hardware and/or software technology used; and development of disaster recovery plans. However, in most developing countries, data producers such as cultural heritage organisations pay insufficient attention to issues surrounding digital preservation. Their digital collections are often stored in digital library and content management systems that are primarily set up for management of assets and external visibility. Activities such as migration to other platforms are assigned a lower priority. In this article, it is argued that a change in the fundamental architecture of the digital library system can potentially improve the preservability of a collection.

Simpler architectures for digital library systems may also make it easier to interconnect systems and extend or modify the features of a specific system. In developing countries, extensions and modifications are relatively rare, given the lack of human and machine resources. Reducing the resource requirements may result in more digital library systems with improved services. Thus, this article presents some elements of an alternative architecture for digital library systems that is arguably easier to manipulate and therefore supports more use cases and long term preservation.

In our earlier work [19], we highlighted some issues that hinder widespread preservation projects in Africa. A number of possible techniques and solutions currently adopted in African projects to preserve heritage were also presented. Overall, simplicity is proposed as a potential key criterion for building lightweight systems that are easy to maintain and migrate, effectively reducing overall maintenance costs. In order to further our investigation on simplicity, a recent project [17] involved building a prototype lightweight, distributable, generic repository management and access system. The design approach centred on simplicity is potentially useful in areas with limited resources and/or expensive Internet bandwidth.

This article focuses on a case study where the Digital Bleek and Lloyd collection was transformed into a simpler form to gain all the potential advantages listed above. The following sections present background work, details of the case study, analysis of the case study and, finally, concluding remarks.

2 Related Work

The 1996 Task Force Report [21] on Archiving of Digital Information initiated formal research in digital preservation. It is particularly important because it identified the core challenges of digital preservation for organisations that manage digital collections. A number of initiatives have since emerged to support digital preservation. The Open Archival Information System (OAIS) reference model [5] provides a formal and comprehensive overview of digital preservation and was approved by the International Standards Organisation (ISO) as the ISO 14721 standard. It has since been used as a basis for most tools that currently facilitate preservation. Preservation Metadata: Implementation Strategies (PREMIS) is a working group, composed of international experts in metadata to support digital preservation activities [13]. PREMIS is essentially a metadata community initiative that recommends best practices for preservation metadata. The PREMIS working group published a data dictionary and an XML schema to support the implementation of the data dictionary in digital archiving systems. The National Digital Information Infrastructure and Preservation Programme (NDIIPP) is a programme, led by the Library of Congress, to archive and provide access to digital resources. NDIIPP is a precursor to the National Digital Stewardship Alliance (NDSA), an organisation focused on improving preservation standards and practices [10].

In the repository software development community, some projects have emerged as enablers of distributed preservation. Lots Of Copies Keep Stuff Safe (LOCKSS) is an international community initiative that provides libraries with digital preservation tools and support to facilitate the preservation of authorised copies of digital content [12]. DuraCloud is a managed cloud service that provides organisations with the necessary cloud subscription services to make it possible to safely store digital content in the cloud by easily multiplexing digital content storage with different providers [6]. MetaArchive is a distributed alliance of university libraries, archives and research centres that works towards supporting distributed digital preservation practices. The MetaArchive corporation uses a technical framework that is based on the LOCKSS software to ingest digital content into the geographically distributed network [15]. However, these hosted services require payment of subscription fees periodically and work on the assumption that participating organisations have adequate Internet bandwidth.

Standalone repository software tools are commonly used to facilitate digital preservation. Greenstone [1] and Omeka [8] are commonly used by cultural heritage organisations for creation of digital collections. Greenstone is particularly popular in developing countries due to its ability to create digital collection archives that can potentially be distributed on a CD-ROM. DSpace [20] and EPrints [7] are popular among academic institutions where they are commonly used to build institutional repository systems. However, the downside of using such standalone tools is that the preservation and extensibility features built into them require individuals with full developer-level knowledge of the software packages.

In contrast, the Library of Congress developed the BagIt specification that specifies a hierarchical file packaging format for storage and transfer of digital content [2]. The packaging format is specifically designed to facilitate the storage and transfer of digital content for long term preservation. The specification has resulted in the development of a number of utility tools that are used to facilitate digital preservation [11]. This concept of flexibility applies equally well to services as it does to data. The Dienst [9], ODL [16] and OpenDLib [3] projects all sought to decrease management and effort, without compromising on features, by introducing flexible Web-based component architectures. In these cases, systems were created by interconnecting a suite of simple components to provide the necessary level of complexity of a fully-fledged digital library system.

The projects described above demonstrate the feasibility of preservation-oriented data stores and flexible service architectures, both premised on simplicity as an architectural principle. This article now merges, extends and generalises these ideas to create a minimal framework for simple digital library systems/collections, operationalised in the Digital Bleek and Lloyd collection

3 Case Study: The Digital Bleek and Lloyd Collection

The Bleek and Lloyd collection is a 19th century compilation of notebooks and drawings comprising of linguistic and ethnographic work of Lucy Lloyd and Wilhelm Bleek on the life of the |Xam and !Kun Bushmen people of Southern Africa. In 2003, the Lucy Lloyd Archive and Research centre at the University of Cape Town embarked on a large scale digitisation project and all the artefacts were scanned and corresponding representation information generated.

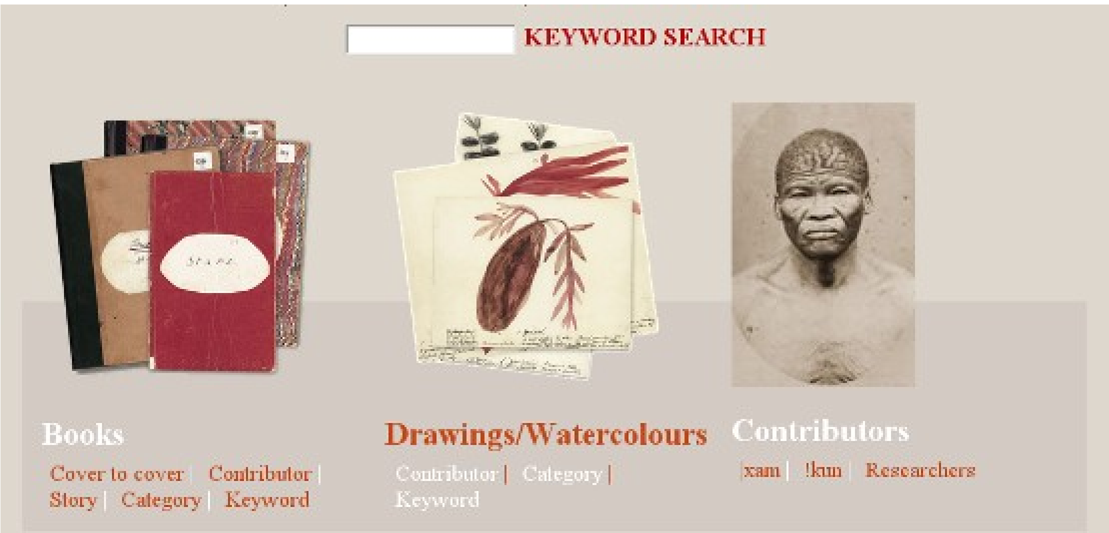


Figure 1. The Digital Bleek and Lloyd Collection

The scanned images formed the basis for the implementation of a Digital Library System that is used to store, manage and preserve digital objects, effectively making the digital collection accessible online.

Explicit user requirements necessitated the implementation of an XML-centric solution that is accessible on a wide range of storage devices (CD-ROM drives, USB drives and local network drives), is hardware independent and also requires minimal software installation. The collection is Web-based and accessible online [4], as shown in Figure 1. The implementation of the digital collection was based on pre-generated hyperlinked static XML pages from XML source data using XSLT. An Ajax-based search component was integrated with the static pages, coupled with a JavaScript routine to perform query operations [18].

The collection's repository architecture was recently redesigned in an attempt to facilitate the long term preservation of the digital content and also to make it possible to easily integrate the system with new collections of recently digitised content and new services.

3.1 High Level System Architecture

The high-level architecture of the system, as shown in Figure 2, is composed of three main layers: the client layer is made up of the user interface through which end users interact with the system; the service layer is composed of services that facilitate discovery of information; and the repository layer is used as a storage facility for the digital objects. The repository stores both the digital content and metadata objects on the filesystem, using a hierarchical filesystem structure, with the metadata objects stored in XML plain text files.

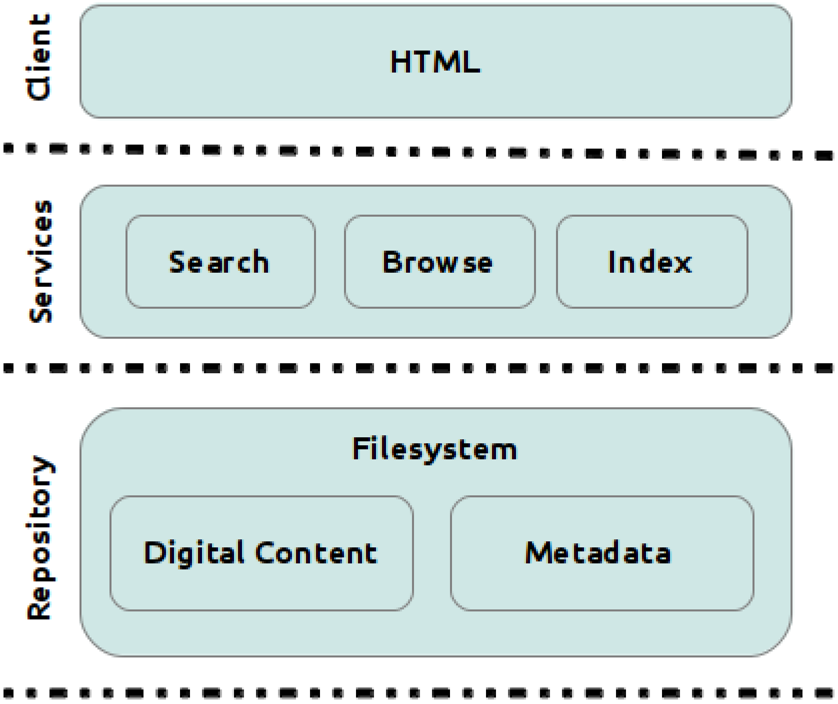


Figure 2. The Digital Bleek and Lloyd System Architecture

3.2 Collection Overview

Table 1 presents an overview of part of the Bleek and Lloyd collection; more content is still being digitized and it is anticipated that these digital objects would eventually have to be integrated with the current digital collection.

Table 1. Overview of the Bleek and Lloyd Collection

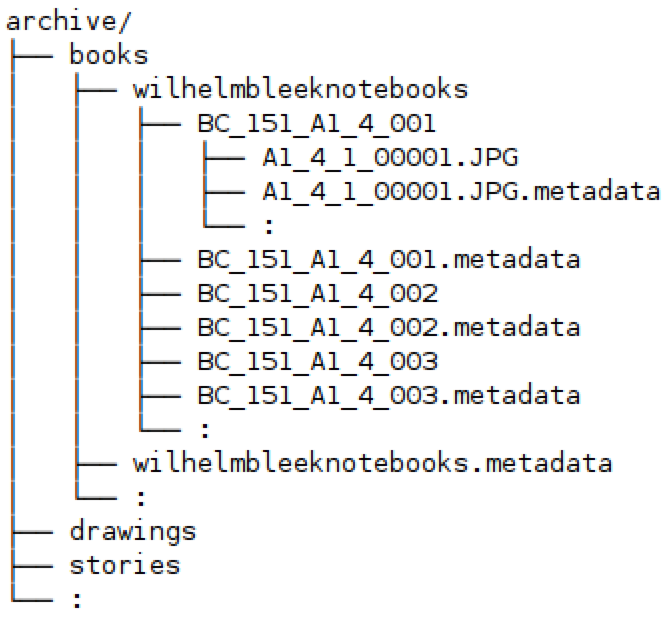
Collection theme	Historical artefacts; museum objects
Material types	Digitised
Collection size	6.2 GB
Content type	image/jpeg
Total number of digital objects	18,924

3.3 Data Model

The digital objects – digital content and metadata objects – are both contained in the filesystem, within hierarchical container structures – filesystem directories. The hierarchical structures conform to the original collection structure that implicitly defines the relationship between the actual digital content and collection types. Figure 3 shows part of the books collection structure of the Bleek and Lloyd collection. The hierarchical structure is important as it helps to preserve the semantics of the digital content. It also helps in the overall management of the repository as bulk actions can be performed on objects contained within the same container. The other advantage of using a hierarchical model is that it helps to implicitly define relationships between disparate objects; in the case of the digital Bleek and Lloyd collection, it helps to define the relationship between actual digital content (e.g., the images) and virtual objects (e.g., the stories).

3.4 Digital Objects

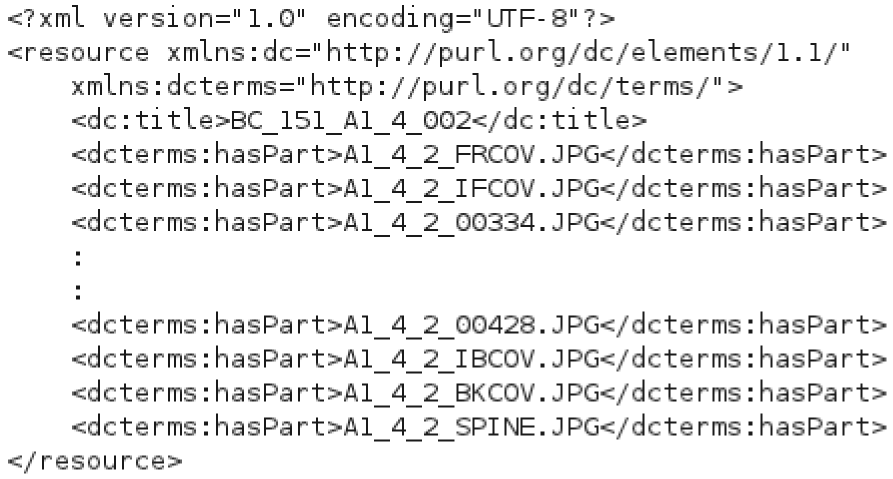
The metadata objects are stored as plain XML text files on the filesystem, alongside the corresponding objects that they describe. The XML text files are encoded using qualified Dublin Core [22].



```
archive/
├── books
│   ├── wilhelmbleeknotebooks
│   │   ├── BC_151_A1_4_001
│   │   │   ├── A1_4_1_00001.JPG
│   │   │   ├── A1_4_1_00001.JPG.metadata
│   │   │   └── :
│   │   ├── BC_151_A1_4_001.metadata
│   │   ├── BC_151_A1_4_002
│   │   ├── BC_151_A1_4_002.metadata
│   │   ├── BC_151_A1_4_003
│   │   ├── BC_151_A1_4_003.metadata
│   │   └── :
│   ├── wilhelmbleeknotebooks.metadata
│   └── :
├── drawings
├── stories
└── :
```

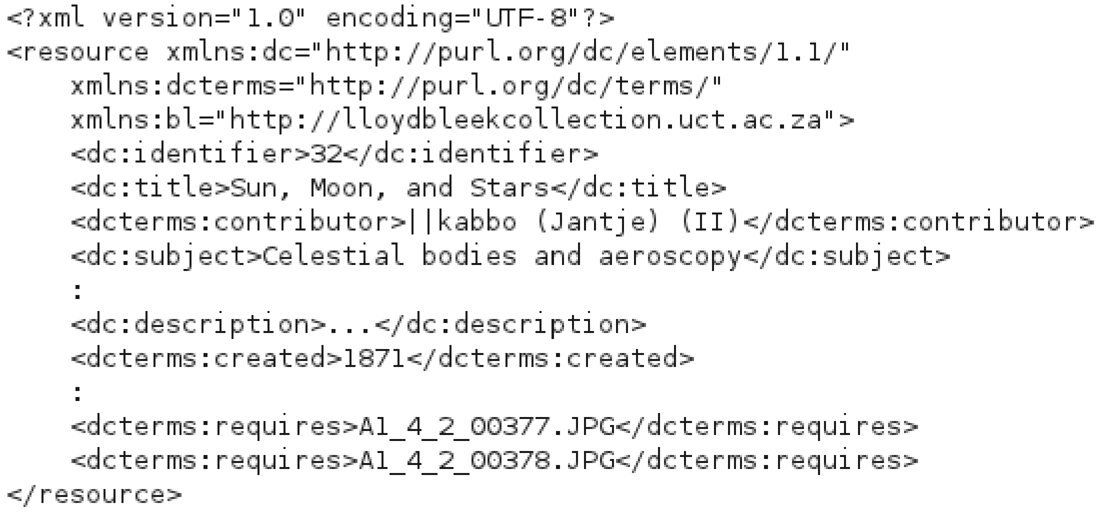
Figure 3. The Repository Hierarchical Directory Structure

This design strategy helps to simplify the preservation of the collection by making is easy for the collection to be migrated to other hardware and/or software platforms. The use of plain text files as storage for metadata, as opposed to the traditional mode of using a database, is especially useful as it makes it possible for the digital objects to be easily moved to any type of hardware or software platform with very little effort. The use of the 'cCopy' utility command present in virtually any operating sSystem is sufficient to move the entire collection. The explicit use of Dublin Core as the metadata scheme ensures conformance to international standards and widespread understanding of the metadata. Physically storing metadata files with the actual objects makes it easier for the representation information to be moved together with the objects when moving to a different hardware and/or software platform.



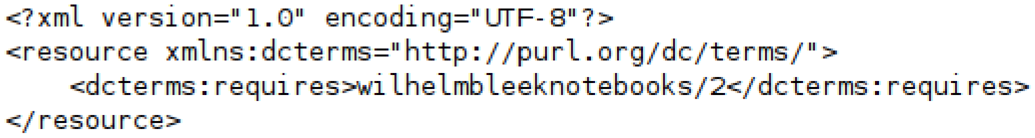
```xml
<?xml version="1.0" encoding="UTF-8"?>
<resource xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:dcterms="http://purl.org/dc/terms/">
    <dc:title>BC_151_A1_4_002</dc:title>
    <dcterms:hasPart>A1_4_2_FRCOV.JPG</dcterms:hasPart>
    <dcterms:hasPart>A1_4_2_IFCOV.JPG</dcterms:hasPart>
    <dcterms:hasPart>A1_4_2_00334.JPG</dcterms:hasPart>
    :
    :
    <dcterms:hasPart>A1_4_2_00428.JPG</dcterms:hasPart>
    <dcterms:hasPart>A1_4_2_IBCOV.JPG</dcterms:hasPart>
    <dcterms:hasPart>A1_4_2_BKCOV.JPG</dcterms:hasPart>
    <dcterms:hasPart>A1_4_2_SPINE.JPG</dcterms:hasPart>
</resource>
```

Figure 4. A Sample Container Object XML Metadata File



```xml
<?xml version="1.0" encoding="UTF-8"?>
<resource xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:dcterms="http://purl.org/dc/terms/"
    xmlns:bl="http://lloydbleekcollection.uct.ac.za">
    <dc:identifier>32</dc:identifier>
    <dc:title>Sun, Moon, and Stars</dc:title>
    <dcterms:contributor>||kabbo (Jantje) (II)</dcterms:contributor>
    <dc:subject>Celestial bodies and aeroscopy</dc:subject>
    :
    <dc:description>...</dc:description>
    <dcterms:created>1871</dcterms:created>
    :
    <dcterms:requires>A1_4_2_00377.JPG</dcterms:requires>
    <dcterms:requires>A1_4_2_00378.JPG</dcterms:requires>
</resource>
```

Figure 5. A Sample Virtual Object XML Metadata File



```xml
<?xml version="1.0" encoding="UTF-8"?>
<resource xmlns:dcterms="http://purl.org/dc/terms/">
    <dcterms:requires>wilhelmbleeknotebooks/2</dcterms:requires>
</resource>
```

Figure 6. A Sample Digital Content XML Metadata File

There are three distinct types of metadata objects: container metadata objects define the composition of a directory that can be composed of other container metadata objects and/or digital content; virtual objects help define the semantics of the images and collection as a whole by describing the logical association of the images and collection; and digital content metadata objects describe the actual images. The container-based metadata files are a manifest of the objects contained within them. Figures 4, 5 and 6 show sample metadata files for container objects, virtual objects and digital content objects respectively.

3.5 End User Access

The collection front-end makes use of statically-generated HTML files, with an AJAX-enabled search feature for online/off-line searchability of information. It is generated by applying XSLT style sheets to all the collection metadata objects. This technological decision removes the dependency of a Web server and/or Application server for rendering of the collection content. The only software component that would inevitably be required is a Web browser, effectively making it possible for the collection to be stored and accessed from other hardware storage devices such as CD-ROMs, USB drives and local network drives.

4 Discussion

The simple techniques and solutions used to redesign the digital Bleek and Lloyd collection have obvious advantages, particularly to organisations with limited resources. The solution allows for the seamless moving of the collection to any hardware and/or software platform and allows for the recreation of the collection in disaster recovery scenarios. The advantage of this is a cost-effective preservation strategy resulting from reduced reliance on expensive, technically-inclined, human resources; and a simplified migration process that any computer novice would be capable of performing.

The use of plain text for storage of metadata, as opposed to using a database, facilitates the portability of system. The filesystem is a very efficient place to store data; in fact, that is where many databases store their data files. Most importantly, however, the digital collection uses the simplest possible retrieval mechanism (the reading of a file in a directory location) and the data is seldom modified. The use of the filesystem for storage of the metadata makes it the optimal storage location.

Hardware and software obsolescence could easily be overcome as the collection can easily be migrated to a different platform. Additionally, plain text files are highly like to be supported in future software versions. There is no guarantee that most of the currently existing proprietary database formats will be supported in the near future or if future software versions will be backwards compatible, but text files in hierarchical directories are probably going to be understandable.

The major challenge, however, would be to explore the scalability of such simple techniques when applied to large and complex collections. Another challenge is perhaps brought about by the use of static HTML files. The static collection necessitates complete regeneration of the entire digital collection when additional content is added to the collection. This is clearly not optimal as it would make it challenging to operate constantly changing collections.

5 Conclusions

Simplicity in the design of digital library systems is arguably a better bet for preservation of crucial cultural heritage collections. At the same time, simple architectures can support extensibility and flexibility of services. These ideas were tested by converting the Digital Bleek and Lloyd to conform to these principles. The new system is wholly a replacement for the prior system, but with a digital object structure that enables simpler preservation over time and enables easier integration of new services. Thus, as a proof-of-concept, it has been shown that it is possible to create a production-quality digital library system based on fundamental notions of simplicity.

This is, however, only a first step in providing sufficient evidence that a simpler architectural design will work for digital library systems in general.

6 Future Work

This simple approach to the architecture of digital library systems can be extended and generalised further.

For the most part, the solution presented is somewhat specific to the Bleek and Lloyd collection; there is thus a need to generalise this approach to make it applicable to a wide range of collections. This would require formally identifying design decisions that may inevitably result in a simple, lightweight system that is easy to manage. This is particularly important as it would highlight how simple repository storage and service architectures can be designed.

A comprehensive evaluation of such an approach is also required to identify the advantages and disadvantages that such solutions have when compared to well-established architectural patterns. It would particularly be useful to establish the scalability of simpler architectures.

A number of existing solutions are implemented with out-of-the-box support for essential features like extensibility and interoperability. It would thus be imperative to explicitly support such features. It may also be necessary to explicitly integrate the current solution with tools and services that facilitate preservation, such as content conversion, logging and integrity checking.

7 References

1. Bainbridge, D., Buchanan, G., Mcpherson, J., Jones, S., Mahoui, A., and Witten, I.H. Greenstone : A platform for distributed digital library applications. *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*, (2001), 137-148.

2. Boyko, A., Kunze, J., Littman, J., Madden, L., and Vargas, B. The BagIt File Packaging Format (V0.97). *Retrieved September 7*, 2011, 26. http://www.ietf.org/internet-drafts/draft-kunze-bagit-06.txt.

3. Castelli, D., Pagano, P., and Thanos, C. OpenDLib: an infrastructure for new generation digital libraries. *International Journal on Digital Libraries 4*, 1 (2004), 45-47.

4. Centre of Curating the Archive. The Digital Bleek and Lloyd. http://lloydbleekcollection.cs.uct.ac.za/.

5. Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System (OAIS)*. Issue, Washington, DC, 2002.

6. DuraSpace. DuraCloud. 2012. http://duracloud.org/.

7. Gutteridge, C. GNU EPrints 2 Overview. *11th Panhellenic Academic Libraries Conference*, (2002).

8. Kucsma, J., Reiss, K., and Sidman, A. Using Omeka to Build Digital Collections: The METRO Case Study. *D-Lib Magazine 16*, 3/4 (2010).

9. Lagoze, C. and Davis, J.R. Dienst: an architecture for distributed document libraries. *Communications of the ACM 38*, 4 (1995), 47.

10. Library of Congress. Digital Preservation (Library of Congress). 2012. http://www.digitalpreservation.gov/.

11. Library of Congress. NDIIPP Partner Tools and Services Inventory. 2012. http://www.digitalpreservation.gov/tools/.

12. Maniatis, P., Roussopoulos, M., and Giuli, T. The LOCKSS peer-to-peer digital preservation system. *on Computer Systems (23*, 1 (2005), 2-50.

13. Premis Editorial Committee. PREMIS PREMIS Data Dictionary for Preservation Metadata. *Preservation*, 2008.

14. RLG/OCLC Working Group on Digital Archive Attributes. *Trusted Digital Repositories: Attributes and Responsibilities*. Mountain View, CA, 2002.

15. Skinner, K. and Gore, E. The MetaArchive Cooperative: chronicles in distributed digital preservation. *Proceedings of the 2010 Roadmap for Digital Preservation Interoperability Framework Workshop*, ACM (2010), 8.

16. Suleman, H. and Fox, E.A. A Framework for Building Open Digital Libraries. *D-Lib Magazine 7*, 12 (2001).

17. Suleman, H., Bowes, M., Hirst, M., and Subrun, S. Hybrid online-offline digital collections. *Proceedings of the 2010 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on - SAICSIT ’10*, ACM Press (2010), 421-425.

18. Suleman, H. Digital Libraries Without Databases: The Bleek and Lloyd Collection. *Proceedings of Research and Advanced Technology for Digital Libraries, 11th European Conference (ECDL 2007)*, Springer Berlin / Heidelberg (2007), 392-403.

19. Suleman, H. An African Perspective on Digital Preservation Keywords Motivation for Preservation Why an African Perspective. *Africa*, (2008), 1-9.

20. Tansley, R., Bass, M., Stuve, D., et al. The DSpace institutional digital repository system: current functionality. *2003 Joint Conference on Digital Libraries, 2003. Proceedings.*, IEEE Comput. Soc (2003), 87-97.

21. Waters, D. and Garret, J. *Preserving Digital Information. Report of the Task Force on Archiving of Digital Information.* 1996.

22. Weibel, S., Kunze, J., Lagoze, C., and Wolf, M. Dublin Core Metadata for Resource Discovery. *Internet Engineering Task Force RFC 2413*, 1998. http://www.hjp.at/doc/rfc/rfc2413.html.