Original: English **CI/INF/UAP/2007/WS/2**

MEMORY OF THE WORLD

Towards an Open Source Repository and Preservation System

Recommendations on the Implementation of an Open Source Digital Archival and Preservation System and on Related Software Development

By

Kevin Bradley
National Library of Australia
UNESCO Memory of the World Sub-Committee on Technology

Junran Lei,
Australian Partnership for Sustainable Repositories,
Chris Blackall,
Australian Partnership for Sustainable Repositories

United Nations Educational,
Scientific and Cultural Organization Paris, June 2007

Recommended catalogue entry:

Towards an Open Source Repository and Preservation System / Prepared by Kevin Bradley, National Library of Australia, Junran Lei, Australian Partnership for Sustainable Repositories,Chris Blackall, Australian Partnership for Sustainable Repositories
Paris: UNESCO, 2007. – ii, 34 p, 30 cm – (CI/INF/UAP/2007/WS/2)

I - Title

II - UNESCO

This report was commissioned by UNESCO Memory of the World Programme and prepared with the support of the Australian Partnership for Sustainable Repositories (APSR) (http://www.apsr.edu.au). Kevin Bradley, National Library of Australia, led the project, edited the report and is ultimately responsible for its content. Junran Lei undertook the research and Chris Blackall supported the development of the structure of the document and coordinated APSR's work. All provided significant content and their valuable contributions are duly acknowledged.

Table of contents

Preface 1
Executive Summary and Recommendations 3
1. Introduction 5
1.1 Background and Aims of the Report 6
1.2 Scope of the Report 7
1.3 Assumptions and Constraints 8
1.3.1 Users: 8
1.3.2 Institutions: 8
1.3.3 Infrastructure: 9
1.3.4 Funding Cycles 9
1.3.5 System Characteristics: 9
1.3.6 Archival Storage and Backup Behaviour 10
2. Survey of Open Source Software 11
2.1 The Open Archival Information System (OAIS) and the Open Source Archival Repository 11
2.1.1 Ingest 12
2.1.2. Archival Storage 14
2.1.3. Data Management 19
2.1.4. Administration 20
2.1.5. Access 24
3 Other Issues and Systems 25
3.1 Shared Content and Exchange Formats 25
3.1.1 Preservation Distribution Information Packages 25
3.1.2 Standards Development Work 26
3.2 Training 26
3.3 Repository Software Comparison 26
3.3.1 DSpace 27
3.3.2 Fedora 27
3.3.3 Greenstone 28
3.4 Packaging and Distro 28
3.4.1 Packaging 28
3.4.1 Distro 29

Preface

Digital information and hence the need to preserve that information has been with us for many decades. The knowledge and the further development of handling and storing digital information was, however, the domain of a small group of specialists. They worked for a rather small spectrum of commercial clients such as banks and insurance companies but also for scientific fields such as astronomy, high energy physics, or meteorology. With the advent of microcomputers in the early 1980s, digital text and image processing became widespread and has not only conquered daily business and academic life but also the private domain.

With the development and expansion of the internet in the 1990s, archives, libraries and museums started to make digital copies if their documents and objects in order to use modern Information Communications Technologies (ICT) to make their contents easily available. For the first time in history, these activities have opened up the access to information and knowledge in truly democratic dimensions.

With increasing levels of digitisation, particularly in the cultural heritage sector, the quantity of digital files to be stored increased well beyond what was produced by the world of text processing and offices for which much of the hard- and software was designed. Specifically, audio and video archives need to store digital files in quantities which significantly exceeded the storage capacities of most of the conventional clients. Only a few cultural heritage institutions - mostly broadcast archives and national institutions - were able to afford professional digital storage technologies. The majority of smaller and financially less wealthy institutions had to choose consumer technology, like optical recordable discs, instead of professional storage technology.

In fulfilling its mandate to monitor matters related to the safeguarding of information and documents of all kinds, the Sub-Committee on Technology of the Memory of the World Programme has commissioned a study on the use of recordable optical discs as reliable digital carriers. The study warned of the danger of the sole use of this technology without extensive testing. This, in turn, is time consuming and costly. Since around 2000, prices for professional IT components for data storage have come within the financial reach of smaller institutions. The study encourages the use of professional technology which is applicable, even under unfavourable climatic conditions, everywhere in the world[1].

An important component to successfully storing digital data in greater amounts is software for the management of the digital preservation process. In contrast to hardware components and storage media, commercial software has maintained a high price level. To date, software developers have not reacted adequately to the market demands; a market which, since the 1990s, has grown well beyond the formerly exclusive clientele to a broad spectrum of cultural heritage institutions, including even private individuals[2]. Due to the lack of affordable commercial solutions, many of the necessary software tools are being developed as cooperative open source software projects, though these are still, to some extent, fragmentary.

Consequently, the Sub-Committee on Technology has commissioned this study in order to survey Open Source Repository and Preservation Software, to analyse existing gaps and to

1. Bradley, Kevin: Risks Associated with the Use of Recordable CDs and DVDs as Reliable Storage Media in Archival Collections - Strategies and Alternatives. UNESCO, Paris 2006 http://www.unesco.org/webworld/risk
2. Schüller, Dietrich: "Personal" Digital Mass Storage Systems - a Viable Solution for Small Archives and Developing Countries. In: Points of View, UNESCO 4/2001 http://www.unesco.org/webworld/points_of_views/schuller.shtml

make recommendations for the development and packaging of an Open Source Digital Preservation System.

As the outgoing Chair of the Sub-Committee I am pleased to present this study to the International Advisory Committee at its 8th Meeting in Pretoria, South Africa, 13-15 June 2007. Thanks go to the CI Sector of UNESCO and to the Australian Partnership for Sustainable Repositories (APSR) for supporting this study. Kevin Bradley and his team deserve our gratitude and compliments for their excellent work. The implementation of the proposed pilot projects and the further development of their recommendations will be in the hands of my successor Jonas Palm, for which I wish him and the Sub-Committee all success

Dietrich Schüller
Chair of the Sub-Committee on Technology
Vienna, 7.June 2007

Executive Summary and Recommendations

This report defines the requirements for a digital archival and preservation system using standard hardware and describes a set of open source software which could used to implement it. There are two aspects of this report that distinguish it from other approaches. One is the complete or holistic approach to digital preservation. The report recognises that a functioning preservation system must consider all aspects of a digital repositories; Ingest, Access, Administration, Data Management, Preservation Planning and Archival Storage, including storage media and management software. Secondly, the report argues that, for simple digital objects, the solution to digital preservation is relatively well understood, and that what is needed are affordable tools, technology and training in using those systems.

An assumption of the report is that there is no ultimate, permanent storage media, nor will there be in the foreseeable future. It is instead necessary to design systems to manage the inevitable change from system to system. The aim and emphasis in digital preservation is to build sustainable systems rather than permanent carriers.

Great national institutions and well funded archives have been working together with industry to develop some remarkable and innovative solutions. In situations where funds are less, though management of the visual, audio or audio visual heritage is still critical, a lower cost solution which will allow for the continual upgrade of systems is necessary. It is only in finding a solution to this problem that a sustainable approach will be found to meet the needs of many communities. If these preservation repositories are well designed these systems will take advantage of digital preservation solutions which are being developed while not incurring the cost of their creation.

The way open source communities, providers and distributors achieve their aims provides a model on how a sustainable archival system might work, be sustained, be upgraded and be developed as required. Similarly, many cultural institutions, archives and higher education institutions are participating in the open source software communities to influence the direction of the development of those softwares to their advantage, and ultimately to the advantage of the whole sector.

A fundamental finding of this report is that a simple, sustainable system that provides strategies to manage all the identified functions for digital preservation is necessary. It also finds that for simple discrete digital objects this is nearly possible. This report recommends that UNESCO supports the aggregation and development of an open source archival system, building on, and drawing together existing open source programs.

This report also recommends that UNESCO participates through its various committees, in open source software development on behalf of the countries, communities, and cultural institutions, who would benefit from a simple, yet sustainable, digital archival and preservation system. Specifically, these recommendations would include:

1. UNESCO establish a steering committee based in the MoW Sub Committee on Technology to support the development of a single package open source digital preservation and access repository
2. Support and resource a pilot project with a number of communities or institutions who can articulate their requirements and act as beta testers of such a system

3. Through that and other committees and projects, influence and support the development of specific software, as discussed in this report
4. Investigate the development of solutions to the system gaps noted in this report, particularly in the area of preservation planning and archival storage systems
5. Support the integration of a number of open source tools to develop a single package open source repository system based on existing open source platforms as described in this report
6. Encourage the development of federated and cooperative approaches through the adoption of standard data packages
7. Ensure that, low cost notwithstanding, the solution is based in international standards and best practice.
8. Support and expand existing training and education to include technical training in the envisaged system in parallel with work on intellectual property and cultural rights.
9. Liaise with existing open source distributors such as Ubuntu, or with development communities, such as the Australian Partnership for Sustainable Repositories (or other suitable) to support these aims.

Though informed decisions about the inclusion of suitable technologies were made, this report does not claim to be either comprehensive or proscriptive. Rather it demonstrates that a practical open source system for digital preservation could, with a little work, be constructed and that this would be of enormous benefit to communities and institutions all over the world.

1. Introduction

Finding a solution to all of the problems of digital preservation is a convoluted and difficult problem with many complications and variables. Many of the world's great cultural, educational and technical institutions are bringing quite substantial resources to bear on the problem. The solutions and systems which this effort will certainly and eventually bring into existence will solve the problem of preserving increasingly complex and multifunctional and multifaceted digital objects. These solutions will, by the very cooperative nature of their solution, be of benefit to all those concerned with the sustainability of digital information.

Related to this is a simpler, but much more pressing problem to be solved for which much of the technological systems already exist, and it is faced by many of the less grand, but no less important, cultural institutions throughout the world. It is the sustainable preservation of simple digital objects which document the social and cultural history of the communities and societies these institutions exist to represent. These objects are simple only in the sense that they tend to have meaning in themselves, not as part of a compound digital object. They are, in other words, discrete digital objects.

These types of objects are:

- Image
- Audio
- Text
- Video (sometimes called audio-visual or moving image).

The methodology for preserving these materials in digital file formats, as individual items, is well understood. Moving image is the most recent content type for which preservation of archival formats have been developed, but the other types of material, i.e. image, audio and text, have had stable formats for some years. The shared principles under which they have been developed are as follows.

- Create and store the content on a digital file in a format which does not apply any form of manipulation which causes data loss or loss of authenticity.
- Use a format which is widely implemented and supported, and preferably, though not necessarily, open or non-proprietary.
- Use a format that has a potentially long life (digitally speaking).
- Use a format that is most likely to have available migration pathways to the next format.
- Store enough metadata to be able to facilitate identification, access and preservation processes.
- Use a reliable storage format on at least two types of carrier.
- Make multiple copies, and check and verify them regularly.
- Plan to replace carriers and software as the market demands, and plan to migrate the content to the next type of reliable carrier.

The technologies which have practical implementation for these principles exist, however they are generally expensive and complex, or such low cost options as are available are fragmented and dispersed. The open source community has developed systems and support for systems which meet the needs of many users world wide, some of which are directly relevant to the aims of this project.

This report examines the availability of open source software using low cost hardware for the purposes of developing a digital archival and preservation system and makes recommendations on how these systems might be developed, implemented and supported in the open source environment.

1.1 Background and Aims of the Report

The UNESCO Memory of the World (MoW) Sub Committee on Technology (SCoT) celebrated the publication of *Risks Associated with the Use of Recordable CDs and DVDs as Reliable Storage Media in Archival Collections - Strategies and Alternatives,* (http://unesdoc.unesco.org/images/0014/001477/147782E.pdf) at its September 2006 meeting in Mexico City. At the same time the committee noted with concern the limiting factors associated with the hard disc and tape based storage systems, currently the most, and possibly only, viable approach to digital storage of large quantities of data in perpetuity. These limiting factors are primarily; cost, including the ongoing replacement cycle; and complexity, making such approaches problematic for use in small scale institutions, for whom recordable CDs and DVDs appear so attractive. This report, which is intended to describe a pathway which might address those concerns, was commissioned by that meeting, and funded by UNESCO MoW and the Australian Partnership for Sustainable Repositories (APSR).

The report is intended to provide actionable recommendations to UNESCO for the encouragement and facilitation of the development of a small-scale open-source digital preservation system(s). It is expected that, by adopting these recommendations, UNESCO could focus the development of such storage systems by encouraging its uptake in existing open source communities, or using some of the successful open source communities as a model for new communities. This development would build on the already extensive range of open source software available, tailoring it for the particular archival responsibilities of small to medium scale digital archives or collections.

A low cost reliable open source archival digital repository would be of significant value to participants in many UNESCO initiatives, including the Memory of the World Programme and the Intangible Cultural Heritage Programme.

This report identifies possible open source pathways for sustainable preservation and, more importantly, identifies the gaps in available technology and recommends a way for UNESCO to encourage digital communities to address that shortcoming. The report may be of some value to those intending to build a digital repository, but its intended audience are those who are responsible for, or are able to influence the direction of open source development so that a packagable and affordable sustainable digital archival storage system can be developed, supported and made available.

The methodology of the project has been to describe (as though to construct) a small-scale stand-alone digital storage system and repository using open source software. The system is intended to undertake all the normal archiving functions of a digital storage system. That is, ingest and task management, metadata extraction and management, preservation and backup storage.

The system should be constructed so that a sophisticated user could build complex capabilities on top of the supported system, while the basic user would still have the necessary functions to manage a collection of simple digital objects.

1.2 Scope of the Report

The need for reliable, sustainable, preservation standard, archival digital storage is critical amongst cultural institutions and those responsible for the management and maintenance of cultural materials. Extensive work has been, and continues to be undertaken to determine the requirements for a digital preservation approach. Standards work has established agreements which facilitate the approach, and high cost systems have been implemented by many of the well funded national cultural institutions which have demonstrated the practicality of undertaking digital preservation. It is, however, still a complex and expensive process.

The digital community as a whole recognise that any digital solution should be based in open standards and automated system because all digital solutions must address the issue of technical change. Digital preservation requires that at some time the technologies on which the information is stored, or in which form the information is encoded, will have to be migrated to a format, operating system or hardware. It is inevitable and unavoidable and most institutions manage this as a regular business requirement and replace there systems on a reasonably regular basis. The approach for digital preservation then, is not to build permanent systems, but rather to construct systems which will facilitate the management and preservation of the data content in the face of change.

There are a finite number of functions an archival digital repository must be able to perform in order for it to reliably and sustainably perform the purpose for which it is designed. These are defined in the Reference Model for an Open Archival Information System (OAIS) as; Data Management, Ingest, Access, Administration, Preservation Planning and Archival Storage.

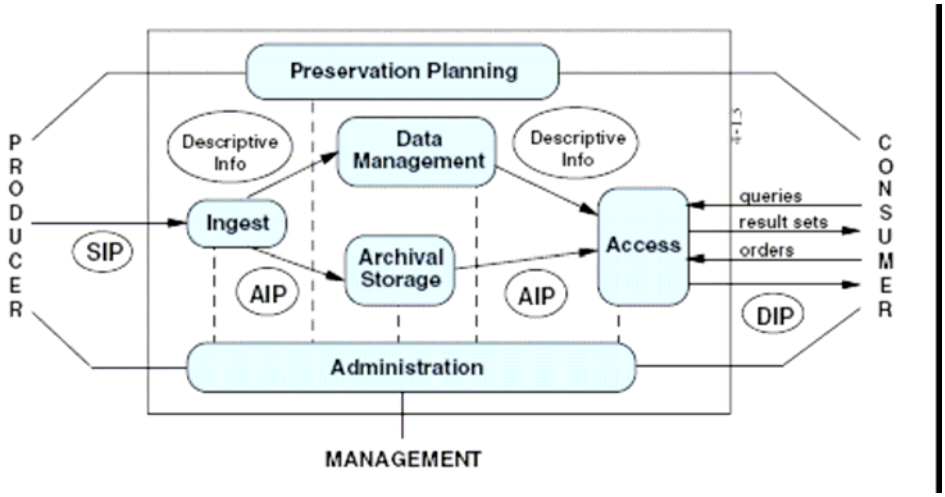


Figure 1 OAIS Functional entities

A number of open source initiatives address one or more of these functions, however, not all necessary functions have been fully addressed and no single open source or low cost system addresses all, or even most, of these issues. The consequence of this is that only commercial systems can currently meet all the requirements of a sustainable digital storage system, though many of these commercial systems take advantage of the open source developments. The end result of this is that only well funded or technically proficient archival institutions can afford

to undertake this approach. It also means that only such institutions can take advantage of the free and open development work being undertaken.

This report recommends that UNESCO supports the aggregation and development of an open source archival system, building on, and drawing together existing open source programs. This report also recommends that UNESCO supports and facilitates the development of an open source distributor who can provide support along the lines of existing providers for other desk top services.

1.3 Assumptions and Constraints

To facilitate this process the report assumes that the primary requirement of the imagined user is for a system which will manage and preserve digital *images*. This is partly because work in image storage is well developed, if fragmented, but primarily because the MoW SCoT have recognised that the most often expressed need is for image based systems. The results of this study, however, could just as easily be applied to text, audio or audio-visual materials. The main issue of the report, which is the viability of the establishment of a sustainable open source digital repository and preservation system, is adequately examined by this process and under this assumption.

The system is constructed making the following assumptions regarding the users, institutions, and the technical infrastructure.

1.3.1 Users:

There are three types of users of the system: Those who access the system for its content, those who create the content for the system and those who administer and maintain the system, and are responsible for caring for its content. It is these latter users, i.e. the curators, administrators and collection managers, who are referred to below.

It is assumed the users have some expertise with computers, i.e. they would have the skills to load self installing software using a wizard-like installation processes; they would be able to enter data and upload image files. They are assumed to have an understanding of the requirement to undertake digital preservation processes without necessarily having a deep understanding of the principles. It is expected that the users could follow online or phoned instructions as to software updates and would be able to follow automated prompts and respond by undertaking any required low level technical tasks.

1.3.2 Institutions:

The types of institutions that might take advantage of such a system could be very broad, as the storage and management of images is a matter of wide interest. For the purposes of this report the institution is assumed to be a small- to medium-size cultural heritage organization responsible for the management of image files, either created digitally or digitized.

1.3.3 Infrastructure:

The system is designed to be a standalone preservation repository, and not dependant on remote storage facilities or systems. The software used to run the repository could be loaded onto a physical carrier such as a CD or DVD, and installed onto a standalone system, and updated in the same way as required. In this regard, the necessary online bandwidth could be quite small. However, in order to gain benefit from downloadable software and open source support the institution should be connected to adequate internet connections to allow exchange of information and some data as allowable on web browsing and email.

"Technology, knowledge, government and economics … are inseparable and interdependent parts" (Rooney 1997). It is clear that no system consists solely of a technical solution on its own; rather it exists and functions within the socio-technical infrastructure that forms that society. This report recognises the nature of the interaction between society and technology without trying to define what that relationship should be to ensure a functioning archival system. It is sufficient to say that a socio-technical infrastructure should be capable of supporting a functioning and sustainable system.

1.3.4 Funding Cycles

The design assumption of this project is that the system is being implemented in an environment with low, but not non-existent, capital investment. This report examines how it is possible to build a low cost technology system, but cannot escape the conclusion that there must be some level of technical knowledge and recurrent resources, albeit at a low level, to make it sustainable. Regardless of the design complexity and robustness of the system, it will need to be replaced at some time or risk losing the content it manages. A recommendation on funding is critical.

"Digital preservation is as much an economic issue as a technical one. The requirements of ongoing sustainability demand at their base a source of reliable funding, necessary to ensure that the constant, albeit potentially low level, support for the sustainability of the digital content and its supporting repositories, technologies and systems can be maintained for as long as it is required. Such constant funding is not at all typical of the many communities that build these digital collections, many of which tend to be grant funded on an episodic basis. There is therefore a need to develop costing models for sustainability of digital materials according to the specific requirements of the various classes of content, access and sustainability." (Bradley 2004).

1.3.5 System Characteristics:

For the purposes of this report, it was specified that the system would have the following characteristics.

- The system will use open-source software and licenses where ever possible.
- It will be designed to manage images for the purposes of this test/analysis.
- Ability to ingest and verify TIFF or JPEG 2000, and generate jpeg distribution copies and thumbnails.
- Able to ingest common non archival camera formats (jpeg etc) and convert to standard forms and extract and/or maintain camera metadata.

- Able to create METS documents (SIP and DIP) for the exchange of Distribution packages
- Able to extract relevant metadata, complying with a set derived from z39.87, the necessary PREMIS metadata, MODS and/or DC data
- The hardware will consist of a simple pc/server, hard drive and tape back up (LTO2 or 3 tape drive or similar).
- The system may be a standalonesystem, or a server based system for access by a number of ingest systems
- The system will be of a size to manage between 1TB and 20TB of image data
- The system and the software will be constructed around the building blocks expressed in the Reference Model for an Open Archival Information System (OAIS).
- It will manage some form of data back up. The extent of backup will be determined as part of the project (i.e. incremental back up, individual item back up, version management etc)
- The system is intended to manage its own data back up and preservation processes.
- Though the system is intended to manage its own data back up and preservation processes it will be aware of other storage systems and will allow for the possibility of data exchange.

1.3.6 Archival Storage and Backup Behaviour

Principles of digital storage for the purposes of preservation

- There should be multiple copies. The system should support a number of duplicate copies of the same item.
- Remote copies. Copies should be remote from the main or original system and from each other. The greater the physical distance between copies the safer in the event of disaster.
- There should be copies on different types of media. If all the copies are on a single type of carrier, such as hard disc, the risk of a single failure mechanism destroying all the copies is great. The risk is spread by having different types of carriers. IT professionals commonly use data tape as the second (and subsequent) copy.

1.3.6.1 Life of Carriers and Systems

If a carrier is replaced within the life of the carrier it will last long enough. This is a truism. However, for practical purposes, it is reasonably true. The life of the tape is conservatively estimated to be five years, and this is considered a suitable operating lifespan by most institutions which store digital information.

It is almost certain that the system which supports access to and replay of the data tape will need to be replaced within 5 years. The forces of functional improvement, obsolescence, carrier failure and market imperatives combine to make this an inescapable fact of digital preservation. All systems must be designed to allow for data migration and system replacement (see SIP and DIP below).

Some carriers could possibly last longer than 5 years. However, if a tape is intended to be maintained for longer than 5 years, then the need for clean room conditions, filtered air

conditioning and other specialist technical structures grows proportionally greater. The longer the period is extended, the less reliable the individual carrier becomes and the greater the risk to content the system manages. The cost of providing a facility which supported very long term retention of the carriers would exceed the cost of replacing the system, especially if an appropriate low cost system is developed along the lines discussed in this report. On the other hand, keeping a system operating using office standard air conditioning and facilities is quite feasible for the projected 5 year life of the tape and system after which the data is migrated to the next (low-cost) system.

It is also worth noting that both IT management and digital preservation expertise agree that maintaining a functioning system for 5 years is practical, but beyond that the need for expertise and specialist knowledge to keep an obsolete operating system functioning, as well as maintaining unsupported hardware, makes this approach prohibitively expensive.

With regard to the life of the system, a conclusion of this report is that a system should be designed to support the data carriers for a finite period, that the reliability of the system should be framed for that finite period, and that the systems architecture, software and hardware should be sufficiently modular to allow upgrades and migration without risk to the content.

2. Survey of Open Source Software

2.1 The Open Archival Information System (OAIS) and the Open Source Archival Repository

The Reference Model for an Open Archival Information System (OAIS) is a widely adopted conceptual model for a digital repository and archival system. The OAIS reference model provides a common language and conceptual framework that digital library and preservation specialists now share. The framework has been adopted as an International Standard, ISO 14721:2003. Though some critics identify shortcoming in the detail of the OAIS, the concept of constructing repository architectures in a form that corresponds with the OAIS functional categories is critical to the development of modular storage systems with interoperable exchange of content. The following sections of the report adopt the major functional components of the OAIS reference model to assist in the analysis of the available software and to develop recommendations for necessary development.

As noted in the introduction, there are a finite number of functions an archival digital repository must be able to perform in order for it to reliably and sustainably perform the purpose for which it is designed. These are defined in the OAIS as; Data Management, Ingest, Access, Administration, Preservation Planning and Archival Storage and are considered in individually below.

The OAIS also defines the structure of the various information packages that are necessary for the management of the data, according to the place in the digital life cycle. These are the Submission Information Package (SIP), Dissemination Information Package (DIP) and Archival Information Package (AIP). A package is the conceptual parcel of the data and

relevant metadata and descriptive information necessary to the particular object. This object is conceptual only in the sense that the package contents may be dispersed in the system or collapsed into a single digital object. OAIS defines an information package as the Content Information and associated Preservation Description Information which is needed to aid in the preservation of the Content Information.

The SIP is an Information Package that is delivered to the system for ingest. It contains the data to be stored and all the necessary related metadata about object. The SIP is accepted into the system and used to create and AIP.

The AIP is an Information Package which is stored and preserved within the system. It is the information package the system stores, preserves and sustains.

The DIP is the information package created to distribute the digital content. There are three roles in this system. First is access, and this DIP would be in a form that the users can use and understand. Second is exchange for the purpose of distributing risk. An archival repository may choose to share parts of its content to other similar institutions, or with an organisation whose role is archival storage. In this case the DIP would contain all the relevant metadata necessary to undertake this role. The third is for distributing content to archives as a last resort. The scenario where a particular archive or institution no longer has the resources to maintain its collection is not difficult to imagine. A standard DIP for this purpose allows other similarly architected systems to undertake the role with the minimum of manual intervention.

2.1.1 Ingest

Ingest, in the OAIS model, is the process that accepts the content and all its related metadata (SIP), verifies the file, extracts the relevant data and prepares the Archival Information Packages for storage, and ensures that Archival Information Packages and their supporting Descriptive Information become established within the OAIS.

The requirement for the test system is that it will accept, verify and identify the standard image file formats (TIFF and JPEG2000), extract the relevant technical metadata according to Z39.87, extract and store camera or scanner information in the relevant systems, convert JPEG (pre 2000) to standard formats (TIFF), and allow for the inputting of descriptive information according to agreed standards.

2.1.1.1. Preservation Metadata Extraction

Current Situation

Metadata is created and entered manually in the majority of the repository systems. This is unsustainable in the long term and leads to non-standard implementations. There is a need to integrate existing metadata extraction tools with repository software so that the systems ingest data and create metadata in a way that is largely transparent to the user. This requires the development of an interface between the extraction tools and the repository to match Fedora's SIP Creator service or DSpace's Simple Archive Format. The existing metadata extraction tools are listed below:

Applications & Tools

National Library of New Zealand Metadata Extractor: The NLNZ tool is a Java-based tool that extracts preservation metadata from digital objects and outputs that metadata in a standard format (XML). The metadata is extracted from the headers of a range of file formats, which currently include MS Word 2, MS Word 6, Word Perfect, Open Office, MS Works, MS Excel, MS PowerPoint, TIFF, JPEG, WAV, MP3, HTML, PDF, GIF, and BMP (Zealand 2007). The output schema appears to be customisable.

Introduction: http://www.natlib.govt.nz/en/whatsnew/4initiatives.html#extraction
Installation: http://www.natlib.govt.nz/files/Preservation/NLNZSetup.exe
Document: http://www.natlib.govt.nz/files/Preservation/docs.zip

The main constraint with the NLNZ tool is that it does not map directly to PREMIS fields. However, the NLNZ schema was developed at the same time as the PREMIS set and there are many parallels. The output from the NLNZ tool could still be used to populate a PREMIS set.

2.1.1.2. Object Validation

2.1.1.2.1. Current Situation

The file format of a digital object should be validated during ingest. Non standard implementation of particular formats can cause an object to become inaccessible. A digital object with valid format should comply with syntactic requirements for its format and meets additional semantic-level requirements (College 2006). Fedora does not currently support object format validation, but it does list this as one of the candidate capability services. DSpace is integrating JHOVE into its workflow code.

2.1.1.2.2. Applications & Tools

JHOVE is a java tool developed by Harvard University for automatic identification, validation and metadata extraction of digital object. It currently supports AIFF, ASCII, BYTESTREAM, GIF, HTML, JPEG, JPEG2000, PDF, TIFF, UTF8, WAVE, XML format.

Website: http://hul.harvard.edu/jhove/
Document : http://hul.harvard.edu/jhove/documentation.html
Software: http://hul.harvard.edu/jhove/distribution.html

On the DSpace Wiki, JHOVE's format identification functionality is described as unreliable. DSpace currently uses DSpace's identification protocol based on file extensions, which itself is unreliable and could lead to misidentification and possible failure of valid files. DSpace developers expect that other developing tools will provide more reliable format identification (DSpace Wiki). However, it is worth noting that the JHOVE tool will be further developed under the Global Digital Format Registry work being undertaken by Harvard University. For the purposes of this simple system, misidentification does not pose a major risk, and the ability to connect with GDFR could be valuable.

DROID (Digital Record Object Identification) is a software tool developed by The National Archives (UK) to perform automated batch identification of file formats. (Archives 2007)

Website: http://droid.sourceforge.net/

Document: http://droid.sourceforge.net/wiki/index.php/Documentation
Software: http://droid.sourceforge.net/wiki/index.php/Download

DROID's main limitation is that it is primarily a batch identification tool rather than an ingest validation tool. It does, however, meet many of the requirements of such a tool.

The object validation tools above are all used for object format identification and validation. Object validation should also include content model validation. But few tools have been developed for object content validation.

2.1.1.3. Implementation issues

The main issue of concern is how to integrate the metadata extractor tools with available repository system, such as DSpace and Fedora or if these systems are going to develop similar tools or add the plug-ins to support extractor functions themselves.

An initial requirement could be to develop XSLT mapping, and a user interface for using metadata extractor tools in these repositories. One set of tools which could support this process are the VTLS Open Source Components (OSC) for Fedora. OSC provides a suite of useful tools and services that integrate with Fedora repositories. Metadata Extraction Services are listed as one of its features.

VTLS Open Source Components (OSC):
Website: http://www.vtls.com/Products/osc.shtml
Software: http://www.vtls.com/Products/osc-download-form.shtml

Adobe community is developing XMP (eXtensible Metadata Platform), a mix of RDF (Resource Description Framework) and XML, used to create and store metadata within resource files such as JPEG, PDF. (Adobe)

2.1.2. Archival Storage

2.1.2.1. Archival Storage

Archival Storage is the services and function necessary for the storage of the Archival Information Package. Archival storage encompasses the data management module and includes a suite of sub-processes such as storage media selection, transfer of AIP to storage system, data security and validity, backup and data restoration, and reproduction of AIP to new media.

For the purposes of this analysis, the system was defined as consisting of a pc/server, hard drive, probably in RAID and LTO2 or 3 tape drive for storage and backup (LTO was the preferred approach because of available error testing regimes, other tapes could be substituted). The system is expected to manage between 1TB and 20TB of data.

The implementation of the module will include the possibility of error measurement and also consider other approaches to the verification of media and content. An implementation would also determine the necessary extent of backup; however, for the purposes of this test, hard disc with two tape copies is considered an adequate approach.

The OAIS Functions of Archival Storage embeds the notion of Hierarchical Storage Management (HSM) in the conceptual model. At the time OAIS was written the situation where large amounts of data could be affordably managed in other ways was not envisaged. The practical issue that underpins the need for HSM is the differing cost of storage media, e.g. where disc storage is expensive, but tape storage is much cheaper. In this situation HSM provides a virtual single store of information, while in reality the copies can be spread across a number of different carrier types according to use and access speeds.

However, the cost of disc has fallen at a greater rate than the cost of tape, to the point where there is an equivalency in price. Consequently the use of HSM becomes an implementation choice. Under these circumstances a storage system which contains all of the data on a hard disc array, all of which is also stored on a number of tapes, is a very affordable proposition, especially for the envisaged size of this test system of up to 20 terabytes. For this type of system a fully functional HSM is unnecessary and instead what is required is a much simpler system which manages and maintains copy location information, media age and versions.

2.1.2.2. Storage and Backup

2.1.2.2.1. Strategies and Technologies

Storage management processes begin with the arrival of a storage request from the Ingest module. The Archival Storage components may need to choose appropriate storage media for the AIP based on the retrieval frequency or storage management policies (CCSDS 2002). Storage management policies or strategies are critical in these processes. However none of the analysed repository systems seem to have proper storage strategies in position. This highlights not only a significant gap in the available open source systems, but also a failure in strategic thinking, where the technical system is divorced from the repository system not only in fact, but also in planning. This is a critical issue for the recommendations in this report.

Storage technologies and products can be split into three main types: direct-attached storage (DAS), network-attached storage (NAS) and the storage area network (SAN). NAS has better performance and scalability than DAS and it is cheaper and simpler to configure than SAN. NAS technology is, from cost benefit view, the most appropriate scalable technology for system of the size under discussion.

For full protection of servers, NAS appliances and workstations we will need to choose backup software such as Amanda and the backup approach that data is first copied to a disk storage system and then copied again to a tape backup system. This type of backup approach has the benefits that can speed up backup and restore, reduce network bandwidth, provide consolidated storage and a cost effective solution for long term archive (Exabyte 2005). Some of the NAS applications, such as FreeNAS, also provide backup functions.

2.1.2.2.2. Applications & Tools

There are many open source and commercial NAS software systems available. For the purposes of this investigation the report concentrates on the three most popular open source NAS applications, FreeNAS, Openfiler and NASLite.

FreeNAS

FreeNAS is NAS software that can turn a PC to a NAS server. It can be installed on, and booted from Hard Drive, Floppy Disk or USB key with a 16MB system image. It supports RAID Array, SCSI, PATA (IDE), SATA, CF and USB drives and can transfer data from FAT, NTFS and EXT2,3 format systems to attached disk, iSCSI target or initiator through CIFS (Samba), FTP, NFS, RSYNCD, SSHD. A user friendly Web interface is available for system configuration and administration. RSYNCD, a multiplatform incremental backup utility, can be configured as service for data backup from FreeNAS server to client.

The limitations of FreeNAS include: A high dependence on its native file system format, UFS. Data on FAT, NTFS and EXT2, 3 format systems can only be transferred to UFS formatted drives. Writing to, or accessing a file on a FAT32 drive which implements CIFS protocol will corrupt the file due to a bug on FreeBSD, the underlying operating system (Cochard-Labbé and Jaggard 2007). Additionally FreeNAS does not appear to support the backup or transfer of data to tape drive, which is a requirement of test system.

Website: http://www.freenas.org/
Download:
http://www.freenas.org/index.php?option=com_content&task=view&id=5&Itemid=27

Openfiler is a Network Storage Operating System. It can be installed on *x86* Intel-based machine that will act as NAS server. Storage can be shared on IP network through CIFS, NFS, HTTP/DAV, FTP, iSCSI. One unique feature of Openfiler is that it can be used to build a Network Attached Storage (NAS) and/or Storage Area Network (SAN) appliance and manage them in a single console. Other features of Openfiler include the support of multiple authentication types, such as NIS, LDAP, Hesiod, Active Directory and NT4 domain controller, and user and group quota on volume. Openfiler is also fronted by a powerful web-based management interface. (Openfiler)
Website: http://www.openfiler.com/
Download: http://sourceforge.net/project/showfiles.php?group_id=90725

Openfiler does not provide native backup functions. Additional backup software has to be run with Openfiler to enable backup to attached devices or over the network. Another limitation is that it must have a central database server to authenticate users and groups.

NASLite is a early Linux distribution for 486 and above x86 based computers providing a way of using the computer as network-attached storage. It supports serving files to clients running Windows, Linux or Mac OS X. Other proprietary versions are available which support different networking protocols, or booting the operating system from a USB Mass Storage device. (NASLite Wiki)
Website: http://www.serverelements.com/naslite.php

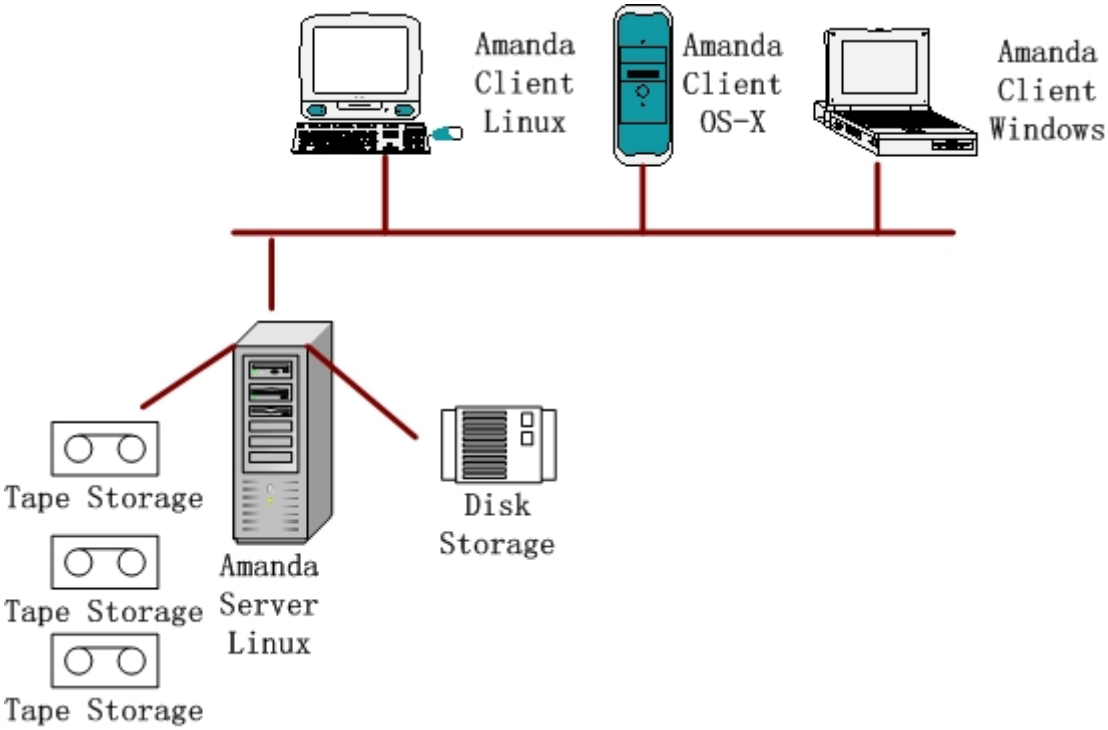
NASLite-1 is free for use, but NASLite-2 is not. Like Openfiler NASLite-1 does not provide backup functions. NASLite-2 is intended for use in any low-security environment. It does not support user management.

Apart from using a NAS system with an additional backup application, there is some powerful backup software that can accomplish the storage and backup requirements locally or across network.

AMANDA
According to it web page, “Advanced Maryland Automatic Network Disk Archiver (AMANDA) is a backup system that allows the administrator to set up a single master backup server. AMANDA allows the back up of multiple hosts over the network to tape drives/changers, disks or optical media. AMANDA uses native dump and/or GNU tar facilities and can back up a large number of workstations running multiple versions of Unix. AMANDA uses Samba or Cygwin to back up Microsoft Windows desktops and servers” (Amanda Home Page http://www.amanda.org/).

It is claimed to be the world's most popular Open Source Backup and Archiving software. It received Linux Journal Readers' Choice Award for "Favorite Backup System" in 2005. Website: http://www.amanda.org/

As illustrated in the figure below AMANDA can backup data from various platform and client systems to an AMANDA server. It has been designed to manage many hundreds of client servers and can also use one single machine as both client and server. Standard operating system utilities such as dump and GNUtar are used for backup. The underlying media does not need to be reformatted. AMANDA can support various tape storage devices naturally or by using tape changer script. It can be used to allocate data from different client to store in different tapes and can be configured not to reuse certain tapes. Each tape can be labelled with the command amlabel. A Perl script is also provided to print labels for AMANDA backup tapes:
http://wiki.zmanda.com/images/5/5d/Amandatape.txt



AMANDA Network (Drawn after Preston (Preston 2007))

The backup process starts by storing the backup data from client to the holding disk attached to AMANDA server. According to the definition from Preston, “A holding disk is one or several directories on any file system that is accessible from the Amanda server. It could be as

small as a single 10 GB directory on the Amanda server drive, or as large as 5 to 10 TB on a Fibre-attached RAID array" (Preston 2007). The benefits of using holding disk as cache are it can avoid shoe-shining (moving the tape back and forward across the heads), improve backup performance and provide additional safety. There is a very sophisticated full and incremental backup schedule function in AMANDA to optimize the backup times and performance. In order to protect data on the backup media AMANDA can also use symmetric or asymmetric encryption algorithms to encrypt backup data (Preston 2007).

All the features listed above show that AMANDA software could be used on one pc/server, hard drive and LTO2 or 3 tape drive for storage and backup requirements for the test system.

Bacula is an open source network based backup software. It supports remote backup of many operating systems, including Linux, Solaris, FreeBSD, NetBSD, Windows, Mac OS X. The backup data can be stored on various media like Tape, DVD or On-line Disk. Website: http://www.bacula.org/

2.1.2.3. Error Checking and Verification

In some commercial software, such as Symantec Backup Exec, tape read/write error can be report automatically during the data backup and verification process. This function is normally implemented with cyclic redundancy check, a technology using checksum against data to detect errors for transmission or storage. It is recommended that an error checking function should be implemented in the future version of open source backup systems. That error checking function should be incorporated into software like Amanda or similar. It would label and check tapes and roster analysis.

Error checking is difficult to implement in open source because that capability is linked to specific hardware. This is not insurmountable, but does present some issues in negotiation between hardware manufacturers and open source software developers. A commercially available stand-alone LTO Cartridge Memory Reader is the "Veritape" from MPTapes, Inc. and recently, Fuji Magnetics announced a Chip Reader Diagnostics System for LTO-Cassettes, bundled with software. The fact that third party suppliers can provide commercial systems suggests that an open source solution is possible, at worst; it may be possible to incorporate such systems into a software solution.

Nonetheless, a low-tech possible alternative to proper error testing is described in the following paragraph.

The data management software has a catalogue (with a printer attached). The hard disc (in RAID) contains a complete set of data. All data is copied onto identical tape copies. There are at least two copies. As data is copied onto a tape, a unique identifier is printed onto a label (human readable) which is attached to the tape. The same identifier can be recorded onto the header of the tape. The data management system can be scripted to prompt the user to find and insert the tape identified by the system. Rather than checking the tape for errors the system will verify the content of the tape against the hard disc. The hard disc can check the veracity of its own data content and is aware of any failings itself. If the verification of the tape fails, the system can produce a new tape from the hard disc. Assuming the system verifies 2 tapes a day, every tape and its duplicate can be verified three times per year. In the event of a disc failure requiring the data tapes to replace it, there will be two tapes which have

been checked within the previous 4 months. The risk that both tapes and the hard disc would fail is very low.

2.1.2.4. Hardware and the Petabox

PetaBox is the cheap hardware based storage system that the Internet Archive developed to manage and store the large amounts of data created as a result of their collection activities. According to its website Petabox (http://www.archive.org/web/petabox.php) was originally created to safely store and process one petabyte (a million gigabytes) of information. A commercial offshoot of the Internet Archive sells the Petabox as a business-related concern. It is said that the Internet Archive releases the technologies to open source community and everything about the PetaBox can be downloaded from its official website: http://www.petabox.org.

The major selling points of PetaBox is low power (as low as 27W per TB), high density and low cost (currently under US$1.50/GB including storage, racks, software, networking, management tools and other components). Though not a complete solution of the type envisaged in this repost, the principles established to develop the PetaBox had strong resonances with the proposed project and it is likely that the open source component would prove useful to this project.

2.1.3. Data Management

Data Management, in the OAIS, is the services and functions for populating, maintaining, and accessing both descriptive information which identifies and documents archive holdings and administrative data used to manage the archive, in other words the catalogue of content and the statistical record of data content.

The requirement for the test system is that it will allow the administration of the descriptive information, including maintaining the schema and loading and updating the records. There will also be a need to be able to interrogate the system to produce result sets of holdings, access usage statistics, contents summaries including sizes and other necessary technical and management information, and descriptive information for AIP file types. For many types of materials some level of control over access to content, or security control, is necessary.

2.1.3.1. Technical and Management Information

Much of this capability is built into the existing repository software. In DSpace the management module use index tools such as Lucene to extract indices or caches from AIP and then these data are maintained in the asset store. Module keep indices or caches up-to-date by periodically polling asset store API, which is similar to ‘incremental harvesting’ in OAI-PMH.

In Fedora the description information is stored as Dublin Core metadata wrapped with Fedora object data in the FOXML file. Currently Fedora supports full-text search and RDF triple store search. There is concern about the performance and usable issues of these search methods.

The problem of how to obtain the cost and benefit tradeoff between assuring AIP and description or discovery information consistency and rapid access to objects needs to be considered.

2.1.3.2. Report Generation

DSpace has quite comprehensive statistical report generation functions including customisable general summary of items, collection, community visits, OAI requests, archive content, user login and popular search.

Fedora does not provide statistical report generation functions but relies on front end systems. Fez, a University based Fedora front-end systems, has the features of generating statistical report by author, community, collection and subject.

2.1.3.3. Security Management

Repositories should have security management functions to control access to its preservation content through a set of sub-module, such as customer management, authentication and authorization. DSpace provides customer profiles management, non-hierarchical authorization, authentication such as X509 certificates. (Tansley 2003) The policies in DSpace could be defined to authorize access to items according to users' or groups' identity, membership or other permissions. (Michael J. Bass 2002) Fedora could be configured to use LDAP and Tomcat JAAS to manage authentication and user information. Fedora also provides XACML-based policy enforcement module for authorization or access control purposes.

2.1.4. Administration

Administration, in the OAIS, is the services and functions for managing system configuration, monitoring operation, providing customer service and updating archival information. It is also responsible for management process such as negotiating submission agreement with producer, auditing submission, control physical access, establishing and maintaining archive standards.

The requirement for the test system is that it will provide friendly interface and restore mechanism for system configuration, monitor and report system operations, performance and usage, provide certain level of customer service and control access for archival information update.

2.1.4.1. Manage System Configuration

The configuration and customization of the existing repository systems are rather complicated or require users to have programming or development experience. For example, there are more than ten steps to edit the configuration file for adding image preview functions for DSpace. To configuration Fedora security policy users have to learn how to write XACML. One of the express recommendations of this report is that a complete system be packaged and made available with a simplified wizard type installation protocol. Additionally it would be desirable if the system can monitor and control the change of configuration, provide the mechanism to restore the default or previous configuration.

2.1.4.2. Monitor Operation

Monitoring the system operations, performance and usage can help establish and improve repository standards and policies and provide information to Preservation Planning. As discussed in 2.1.3.2.Report Generation section, DSpace and Fez provide statistical reports related to the system usage and archival holding. But other monitoring functions such as performance, function and operation monitoring are not included in the analyzed repository systems.

2.1.4.3. Customer Service

The customer service section would provide functions for administers and users to manage customer accounts, collect and respond to feedback, manage bill and payment if system offers pay service. If deemed necessary, all these functions could be built on top of existing repository systems as an extension.

2.1.4.4. Archival Information Update

The system should control who can update and what can be updated and provide audit mechanism for the update process. The policies in DSpace can be defined to authorize access to items according to users' or groups' identity, membership or other permissions (Michael J. Bass 2002). Fedora can restrict access to the repository, to a group of objects, to certain kinds of data streams or to disseminations according repository-wide and object-specific policies. (Fedora 2007)

2.1.4.5. Preservation Planning

OAIS states "this entity provides the services and functions for monitoring the environment of the OAIS and providing recommendations to ensure that the information stored in the OAIS remains accessible to the Designated User Community over the long term, even if the original computing environment becomes obsolete."

Preservation planning for a small scale institution which is solely interested in the preservation of its own simple digital objects is a readily identifiable task. While a well funded national institution must be involved in research and development of the major digital preservation problems, a smaller institution requires much less. It need only have a system with an architecture that will allow it to take advantage of the solutions that the world's great cultural, educational and technical institutions are developing, and must have acquired the necessary information for that purpose.

The metadata requirements are defined in PREMIS, and adherence to this approach will support future preservation actions. Most of PREMIS can be readily mapped to schemas already used and supported, and if the repository is only responsible for simple digital objects as defined in the introduction, the burden of acquiring and managing that data is quite small. The major problem in implementing PREMIS in virtually all current open source systems is the inability to record the changes that might occur to an object as a result of some preservation action. In PREMIS these are referred to as events. "The Event entity aggregates metadata about actions. A preservation repository will record event for many reasons. Documentation of actions that modify (that is, create a new version of) a digital object is critical to maintaining digital provenance, a key element of authenticity." (PREMIS Data Dictionary pg. 1-5).

The other critical need for a small scale institution is the knowledge of when a preservation action is necessary. In OAIS this is described as the monitor technology function. The system must be able to be notified when events occur which might cause the obsolescence of a given format or object and must be able to implement a preservation action (most likely a format migration), if that is deemed necessary. Other than that, the system must only maintain the content until the next event occurs.

2.1.4.6. Monitor Technology

OAIS states that a system should track the emerging technology, new information standards, and format and computing platforms to identify the technologies which could cause obsolescence. This module provides reporting, technology alert and data standards for preservation strategy development, migration planning and information package design. (CCSDS 2002)

2.1.4.6.1. Applications & Tools

The National Library of Australian and Australian Partnership for Sustainable Repositories are working on the AONS Generic Extension project to develop platform-independent software to monitor the file formats in digital repositories. The software can send notifications or recommend preservation actions to repository managers if the monitoring results show that file formats are obsolete or at risk of become obsolete. AONS software uses many third party tools to support its services. It determines the file formats of repositories content through DROID and JHOVE, retrieves file format information from PRONOM, Library of Congress Sustainable Digital Formats website or format registries such as the GDFR. (NLA 2007)
Web page:
http://wiki.nla.gov.au/display/APSR/AONS+II
http://www.apsr.edu.au/aons/index.htm

AONS builds on the work carried by Jane Hunter at University of Queensland. Known as PANIC (preservation webservice architecture for newmedia & interactive collections), the system argues for a semi-automated preservation systems based on semantic web services http://metadata.net/panic/publications.htm.

Format Registry website and system:
Global Digital Format Registry (GDFR): http://hul.harvard.edu/formatregistry/
PRONOM: http://www.nationalarchives.gov.uk/aboutapps/pronom/default.htm
http://www.nationalarchives.gov.uk/aboutapps/pronom/tools.htm
Library of Congress Sustainable Digital Formats website
http://www.digitalpreservation.gov/formats/

VersionTracker, a software version registry website, provides tools to ensure software and drivers up to date. But it mainly focuses on commercial software.
Website: http://www.versiontracker.com

2.1.4.6.2. Applications Limitations & Potential Needs

The services to monitor technology function as specified in OAIS is an emergent field and is under development by GDFR, APSR, and many others. Significant progress has been made in format tracking. More components may need to be added for other parts of the

technology monitoring and there needs to be some development in third party services. Nonetheless, this is a critical area in the preservation of digital content.

2.1.4.7. Provenance Information

According to CCSDS, the information that documents the history of the Content Information in OAIS is Provenance Information. “This information tells the origin or source of the Content Information, any changes that may have taken place since it was originated, and who has had custody of it since it was originated” (CCSDS 2002). The records stored in repository systems over a long period of time may require changes to be made, and these changes should be recorded and tracked for backup, restoration, statistic and provenance tracking purposes.

According to DSpace versioning proposal, versioning should control change of format, “reformatting”, and change of content, “revision”. DSpace does not truly support versioning at this moment. It can only create additional bitstreams, bundles and items. “There is not currently any way to pick out versions… There is also no current way to persistently refer to "the latest version of this document", which appears to be commonly desired by users.” (Ockerbloom 2006) Fedora versioning focuses on “revision” whilst does not deal with format migration. The versioning functions could include getting access to the past and latest versions, versions suppressing or purging, metadata shifting with versions, versions access control. (Ockerbloom 2006)

2.1.4.7.1. Applications & Tools

There is a version management module provided by JSR 170, a Java Content Repository API that defines a standard to access content repositories from Java code. **JSR170** supports a hierarchy versioning model that could provide a clean and easy-to-use programming model for JSR170 compliance repositories. (Sommers 2005)
JSR170 website: http://jcp.org/en/jsr/detail?id=170

Apache Jackrabbit, the Open Source Content Repository for Java, is a fully conforming implementation of the Content Repository for Java Technology API (JCR), which including JSR 170 versioning model.
Apache Jackrabbit: http://jackrabbit.apache.org/

2.1.4.7.2. Applications Limitations & Potential Needs

None of the widely adopted repository systems provides standard, out-of-the-box versioning management function. Although the Fedora repository system includes an infrastructure to support versioning of digital objects and their components, it only provides API to access the versions and does not seem to be easily approached. It would be desirable for repository systems to provide standard and easy to use versioning management components for the long term preservation.

Additionally if one representation of the record is transferred from another representation, for example, transferring JPEG to TIFF, derivation information and technical data should be recorded in the system in a way that retransferring can be perform when it is needed.

2.1.5. Access

The OAIS Reference Model defines “access” as the entity that “provides the services and functions that support Consumers in determining the existence, description, location and availability of information stored in the OAIS, and allowing Consumers to request and receive information products.”

The model goes onto state that access functions include;

- “communicating with Consumers to receive requests;
- applying controls to limit access to specially protected information;
- coordinating the execution of requests to successful completion;
- generating responses (Dissemination Information Packages, result sets, reports); and
- delivering the responses to Consumers.”

2.1.5.1 In House Access

The primary access functions available in the leading repository platforms provide end-users with the means to request information about collection, and items within collections, in the repository. In most cases, this simply means providing end users with ‘browse’ facilities to locate items using predefined branching structures, or controlled vocabularies. Additionally, all repository platforms support a ‘search’ interface for end-users to locate items, or to search through the full text of items via user-defined queries.

2.1.5.2 Delivering and Generating Access

In addition to browse and search functions, the OAIS Reference Model specifies two other access mechanisms.

The first is “delivering responses” to end users. This, for example, could include a wide range of repository management notifications that might be useful to repository managers. It might also include notifications to end-users about new content in the repository, in the form of RSS feeds for instance.

The second access mechanism is “generating responses”, which covers the creation and dissemination of DIPs (Dissemination Information Packages). DIP’s can be programmatically created within the repository environment for a number of purposes. In this report, for example, we envisaged a scenario where the complete contents of a repository is written to a DIP consisting of the zip file containing the digital items and metadata (encoded using METS), which are then transferred mass data storage media for transfer to a third-party repository for archiving.

2.1.5.3 Access Services

The access functions of repositories are typically defined by the repository software itself. There are two basic design approaches adopted by repository developers here.

The first is to crate a 'monolithic' repository application that has the access functions built in from the start. DSpace exemplifies this design approach by providing a generic user interface that provides browse, search and notifications for end-users and repository managers. The advantage of this approach is that DSpace is relatively easy to install and maintain and has a consistent user- interface. The disadvantage is that it is difficult to extend DSpace to cater for the needs of specific users and communities.

The second approach is exemplified by Fedora, which is essentially designed as repository 'engine' with a set of API (application programming interfaces) that enable communities to build and customize their own user-interfaces and access functions. The advantage of this approach is that it provides the maximum flexibility to build and customize increasingly sophisticated access services. The disadvantage is that building new access services is a time consuming and expensive task requiring wide community support.

2.1.5.3.1 Applications & Tools

One solution to the problem of creating new access services has been the development of the Manakin framework (http://di.tamu.edu/projects/ xmlui/manakin/). Manakin provides a repository-independent method of developing new access services and repository 'front-ends' that encourages the development of common software components and code- reuse. It has been initially developed for DSpace, however, it could easily be adapted for Fedora.

2.1.5.3.2 Applications Limitations & Potential Needs

More broadly, repository developers await the development of common, standards-based, APIs for repositories that would make the development of new access services easier.

3 Other Issues and Systems

3.1 Shared Content and Exchange Formats

3.1.1 Preservation Distribution Information Packages

As stated in Section 2.1, there are three roles for a Dissemination Information Package (DIP): access, exchange for the purpose of distributing risk, and copies for archives of last resort. The second and third uses described here are effectively Preservation Distribution Information Packages (P-DIP) and require a certain level of agreement with regard to their content and form. A P-DIP would contain all the relevant information necessary to continue the archival role undertaken by an archive. This would include the data that makes up the object itself in its archival form, the technical metadata, descriptive metadata, the structural metadata, rights metadata, and the metadata created to record provenance and change history. It would need to be packaged in a standard form so that it could be used to recreate the archive if data was lost, or so that another archive could take up the role of managing content if that was deemed necessary.

Whether this is used to support remote content replication or to support federation of cooperating archives, the agreement about standard form and exchange is necessary. A network of repositories which can create and accept such DIPs would be a most effective

preservation strategy, spreading the risk of failure due to natural or man made disaster, or just lack of resources at a critical time in the life-cycle of the digital object.

3.1.2 Standards Development Work

Significant work is currently underway to establish standards for creating DIP and/or SIP to exchange data between repositories, and other third-party applications. For example, the Australian Partnership for Sustainable Repositories is collaborating with the National Library of Australia to create a series of METS (Metadata Encoding and Transmission Standard) Profiles to define DIPs for a range of collection types, such as, electronic journals, images, music and researcher information etc.<http://pilot.apsr.edu.au/wiki/ index.php/Main_Page

3.1.2.1 Limitations & Potential Needs

Despite this progress, consensus and community-based activity is required to develop a METS profile for Preservation Distribution Information Packages.

3.2 Training

This report provides a potential pathway to the preservation of simple digital objects. A critical part of its success in any given situation is the interaction between those who use and implement the system, and the technology itself. Careful socio-technical consideration at in the design is one aspect, but the other is training. UNESCO supports various training in audio visual archiving, however, in order to make a system such as the one described, viable, careful attention would have to be paid to practical, hands on guidance in the use of the system and a clear and practical explanation of the necessity of the various parts. This would have to include downloadable instructions from the distribution site.

3.3 Repository Software Comparison

The OAIS Reference Model has undoubtedly been very influential in the design and development of repository platforms, however, of the small number of candidate repository platforms that could be recommended to UNESCO, none implement OAIS fully, nor do they do so in ways that make them entirely suitable for the preservation of online cultural heritage. It is important to note that repository software addresses only some of the aspects of digital preservation, and generally assumes that the technical aspects are taken care of “elsewhere”. Henry Gladney observes this when he states that most approaches to digital preservation “fail to distinguish between digital repository and digital preservation. The former topic is well developed, with software offerings that have been refined for about a decade.” (Gladney 2004).

In this section we briefly describe three repository platforms that might form the basis of a recommendation: DSpace, Fedora and Greenstone. Before describing the strengths and weakness of these repositories, we should point out that a number of groups have formally evaluated, or are evaluating, repository software. The National Library of New Zealand, for example, recently published an authoritative report titled, Institutional Repositories for the Research Sector (http://wiki.tertiary.govt.nz/tertiary/wikifarm/InstitutionalRepositories/uploads/Main/IR_report.pdf). Such repository evaluations should be consulted before a final decision about the choice of repository software is made.

3.3.1 DSpace

The DSpace repository platform is a very popular and widely adopted repository within the higher education and research sectors, although knowledge its use within the museums and cultural heritage sectors is limited. One of the reasons for the popularity of DSpace is that it is relatively easy to install and maintain, and has a ready made user-interface that integrates data management and access functions within the system's architecture. DSpace is also focused on one aspect of the problem of long-term preservation through the inbuilt support of such features as the Handles System (http://www.handle.net/software) for persistent identifiers. Moreover, a strong international developer community has evolved to support DSpace and new features are being added constantly.

One of the strengths of DSpace is its integrated feature set enabling institutional users to quickly establish a repository and then start adding new items to the collection. This strength, however, is also so one of its major weaknesses, in that DSpace has evolved into a monolithic software application, and complex code base, that introduces potential scaling and capacity constraints for some large institutional users. In order to overcome these constraints, the DSpace software would need to be essentially redesigned and recoded. Though the DSpace people are aware of this issue, there are no clear plans to achieve this goal.

Overall, the short term benefits of adopting DSpace have to be contrasted to the uncertainties surrounding its long-term future. This could be a particular issue for the UNESCO users recommended repository in so far as new preservation services identified in this report would need integrated into the current DSpace code-base.

3.3.2 Fedora

Fedora (Flexible Extensible Digital Object and Repository Architecture) is an increasingly popular repository system that is designed as a base software architecture upon which a wide range of repository services can be built, including preservation services. Compared to the speedy adoption of DSpace, Fedora has been slower to gain adopters because it lacks a dedicated user-interface and access service out-of-the-box. This limitation has been solved with the development of commercial and open-source web-based front-ends for Fedora. A particularly capable open source front-end is named Fez (http://sourceforge.net/projects/fez). Fez has been developed at the University of Queensland with support of the Australian Partnership for Sustainable Repositories (APSR) and uses PHP code in combination with the Postgres database for its user interface elements and other valued-added repository services.

The main strengths of Fedora are its flexible and reportedly scalable architecture. Though some analysts claim limitations to the extent that Fedora can be scaled up, the experiences of institutional adopters indicate that Fedora can scale to cope with large collections, yet is sufficiently flexible to store multiple types of digital items and their complex relationships. Despite the lack of an out-of-the-box user interface, the architectural separation between the 'backend' and 'front-end' of Fedora means that there are few limitations to the features that can be added to it, whilst still remaining interoperable with other software applications and systems. Because of these strengths, Fedora is attracting an international community of developers who are contributing to its development.

The main disadvantage of Fedora is the high level of software engineering expertise required to contribute to its core development.

3.3.3 Greenstone

The Greenstone (http://www.greenstone.org/cgi-bin/library) repository is included in this brief survey because it has been developed and distributed in cooperation with UNESCO and the Human Info NGO. Historically, Greenstone was one of the first repository systems widely adopted, however, it has since been overshadowed by DSpace and Fedora. Greenstone version 2 had several limitations due to its now outdated Perl code base, however, Greenstone version 3 was released in February 2007 and it has been completely redesigned and recoded in Java to take advantage of the latest technologies (http://greenstone.sourceforge.net/wiki/index.php/Greenstone3).

It is too early to assess the strengths and weaknesses of Greenstone version 3, nevertheless, the Greenstone development group based at the University of Waikato, New Zealand, has formed long-term relationships with UNESCO's partner communities and institutions in the fields of education, science and culture around the world, and particularly in developing countries, and thus is in a good position to adapt the new software to provide a sustainable preservation environment for cultural heritage collections.

While comparisons between DSpace, Fedora and Greenstone understandably focus on their technical capabilities, it is important to take into account their levels of institutional uptake and community support they enjoy because these are critical to maintaining the sustainability of repository systems, particularly open-source initiatives. Here institutions provide direct and indirect resources that fund open-source repository initiatives and the user community assumes responsibility for many key tasks: updating and maintaining the underlying software code, writing training resources and supporting community forums to share knowledge and experience. DSpace and Fedora are currently the best supported repository systems, thus Greenstone 3 would need to significantly to improve its international profile in order to for it to be considered a sustainable platform, nonetheless, the new version 3 offers advantages well worth considering.

3.4 Packaging and Distro[1]

3.4.1 Packaging

Currently users have to undergo a complicated manual process to install and configure the majority of the repository systems. The installation process might include code compiling, database and web server applications download and setup, customizing configuration files and so on. The installation time largely depends on the experience of users. This is untenable if we wish to make the use of such technology more usable. There are some software packaging or distributing tools can help to shorten these processes, such as Dpkg, RPM-Build for Linux, WiX toolkit, InnoSetup for Windows and PackageMaker for MacOS. Linux system have more diverse ways and tools for software packaging than the commercial systems, probably as a result of the huge open source communities and the lack of commercially driven standard approaches.

[1] When applied to Linux or other open source systems, distro is a collection of software, tied together with a few tools or is used to describe the process of doing so, such as Ubuntu

The most accepted package formats in Linux are RPM and DEB, which are distribution dependent package formats. They may not be compatible with all Linux distributions as different distributions could have varied files system directories and package dependencies. The tools for creating these formats include:

Dpkg (Debian PacKaGe)

Dpkg is a build-in low level tool for package management in Debian. “dpkg-build” package command can be used to create Debian .deb files, an application package that has been used by various Linux distribution including Ubuntu, Knoppix, SimplyMEPIS, Linspire, Xandros and Debian Linux.
Instruction: http://www.linux.com/article.pl?sid=07/02/21/1546215

rpm-build

The rpm-build package contains the scripts and executable programs that are used to build packages using the RPM Package Manager.
Instruction: https://pmc.ucsc.edu/~dmk/notes/RPMs/Creating_RPMs.html

Many other tools for package and installations in Linux are emerging with innovative approaches, such as autopackage, kilk and zero-install.

Autopackage can create package that could be compatible with any Linux distributions. It resolves package dependencies based on file system scan instead of database search. (Byfield 2005) However it would cause package failure if distributor changes the related package dependencies. (Licquia 2005)
Website: http://autopackage.org/

Kilk provides the self-contained application image (.cmg file) for installation. The image contains all the libraries required, therefore user just needs to copy the file to hard disk and run it directly. No other system changing or files placing need to be done. (klik 2007)
Website: http://klik.atekon.de/

3.4.1 Distro

There are a many opinions regarding these software systems and the traditional RPM, DEB package tools. Some argue that the distro and its dependent packages are updated and released frequently and communities are already used to this package system and so other package formats will have high barrier to enter the “market” (Byfield 2007). Others state that Autopackage and the like are a useful complement to exisiting system for the distribution of non-core package.

In addition to the above mentioned ways of creating a software package, developers can also choose distributing the package with Linux Distro. One example we could look at is the Fez and Fedora installation on Kubuntu. They are currently making the Live Kubuntu CD which contains Fez and Fedora packages.
http://dev-repo.library.uq.edu.au/wiki/index.php/Installation

LiveDistro or Linux Live CD

LiveDistro is the linux Distribution stored in bootable media such as CD that can be executed from the media without installing to hard disk. (wikipedia 2007) Some of the Live CD can be used as installation disk to set up the Linux system, such as Ubuntu/ Kubuntu Linux Live CD. The advantage of using Live CD is that you can try the Linux system and its packages without any installations. There are tools to create Live CD and customize the software packages in the Distro.

UCK - Ubuntu Customization Kit
It can be used to customize the language, packages, documents and booted modules of original Ubuntu/Kubuntu live CD.
http://uck.sourceforge.net/

"**Linux From Scratch**" website provides detail procedure from customizing Linux system to building Linux Live CD:
http://www.linuxfromscratch.org/lfs/
http://www.linuxfromscratch.org/blfs/
http://www.linuxfromscratch.org/livecd/

Linux Live Script
The script tools for creating Linux Live CD from installed Linux system.
http://www.linux-live.org/

The List and links of official Live CDs for Linux Distributions:
http://www.frozentech.com/content/livecd.php

With the emergence of Virtual Machine technologies, building platform independent package would be less important. Package can be installed and run on any operating systems with the supports of Virtual Machine software. Using VM software even has the advantages of securer environments and faster booting speed. (wikipedia 2007)

The detailed comparison of Virtual Machine package:
http://en.wikipedia.org/wiki/Comparison_of_virtual_machines

3.4.1.1 Ubuntu

http://www.ubuntu.com/
There are a number of software distributions, or distros, and each offer particular advantages according to the requirements of the users. DistroWatch http://distrowatch.com/dwres.php?resource=about is a site which monitors the distros and provides statistical comparisons regarding their use. Ubuntu consistently rates highest on that site.

Ubuntu is a model for the type distribution and support process envisaged for the simple open source repository described in this report. Ubuntu provides a free, open source, gnome licensed set of software for desktops or servers. It contains the required desktop software included in OpenOffice.org, the Mozilla Firefox web browser and the GIMP image editor, as well as a number of other functionalities combined with their user interface. These are packaged using the distro software and supplied as an ISO image for a CD which can be downloaded or posted.

Ubuntu’s motto is “Linux for Human Beings” and they state on their site they have a clear focus on the user and usability (it should "Just Work"). The open source community supports development.

The aim for this archival repository project should be to build an easy to use, low cost system that the community of users supports and to which it contributes. The initial project should be guided by best practice, but informed by a pragmatic approach. The hardware should be available as a set of options selected from affordable solutions, and the software written and distributed in an open, supported way. Eventually the community of users should guide its direction and manage its development so that it becomes a truly open and responsive system.

References:

Bekaert, J., X. Liu, et al. (2005). Using Standards in Digital Library Design & Development (Tutorial), Joint Conference on Digital Libraries.

Byfield, B. (2005). "Autopackage: Toward a universal package manager for the desktop." Retrieved March 2007, from http://www.linux.com/print.pl?sid=05/11/22/2021228.

Byfield, B. (2007). "Autopackage struggling to gain acceptance." Retrieved March 2007, from http://specialreports.linux.com/print.pl?sid=07/02/09/0854226.

Bradley, Kevin 2004 “Sustainability Issues”, APSR Report http://www.apsr.edu.au/documents/APSR_Sustainability_Issues_Paper.pdf

Carpenter, G. (2004). "Choosing a Tool for Object Identification, Validation, and Metadata Extraction." Retrieved Feb 2007, from http://wiki.dspace.org/index.php?title=JhoveLNZComp.

Carpenter, G. (2006). "TechMDExtractor." Retrieved Feb 2007, from http://wiki.dspace.org/index.php/TechMDExtractor.

CCSDS, C. C. f. S. D. S. (2002, January). "Reference Model for an Open Archival Information System (OAIS)." Retrieved January, 2007, from http://public.ccsds.org/publications/archive/650x0b1.pdf.

CCSDS, C. C. f. S. D. S. (2002, January). "Reference Model for an Open Archival Information System (OAIS)." Retrieved January, 2007, from http://public.ccsds.org/publications/archive/650x0b1.pdf.

Cochard-Labbé, O. and B. Jaggard (2007). "FreeNAS Setup and User guide (Version 0.684b)."

College, J. a. t. P. a. F. o. H. (2006). "JHOVE - JSTOR/Harvard Object Validation Environment."

Day, M. (2006). Seminar: OAIS Model application in digital preservation projects, Digital Curation Centre, UKOLN, University of Bath.

Exabyte (2005). Disk-to-Disk-to-Tape (D2D2T) An Exabyte White Paper.

Fedora (2007). Fedora System Documentation.

Gilbert, M. W. (2003). Digital Media Life Expectancy and Care

Gladney, Henry 2004 Open Letter to Ms. Martha Anderson,
NDIIPP, Library of Congress,
http://home.pacbell.net/hgladney/NDIIPPcritique2004.pdf

Glick, Kevin, E. W., Robert Dockins (2006). Fedora and the Preservation of

University Records Project, Digital Collections and Archives Tufts University, Manuscripts and Archives Yale University.

Hague, T. (2001). Publication of Digital Preservation Testbed White Paper Migration: Context and Current Status.

Kiehne, T. P. (2005). Digital Preservation Plan for the Texas Legacy Project.

klik. (2007). "What is klik." Retrieved March 2007, from http://klik.atekon.de/wiki/index.php/What_is_klik.
Lavoie, B. and R. Gartner (2005). Technology Watch Report Preservation Metadata.

Lavoie, B. F. (2004). Technology Watch Report: The Open Archival Information System Reference Model: Introductory Guide. Office of Research OCLC Online Computer Library Center, Inc.

Licquia, J. (2005). "Autopackage Considered Harmful." Retrieved March 2007, from http://www.licquia.org/archives/2005/03/27/autopackage-considered-harmful/.
Michael J. Bass, D. S., Robert Tansley, Margret Branschofsky, Peter Breton, Peter Carmichael, Bill Cattey, Dan Chudnov, Joyce Ng (2002). DSpace –A Sustainable Solution for Institutional Digital Asset Services –Spanning the Information Asset Value Chain: Ingest, Manage, Preserve, Disseminate Internal Reference Specification Technology & Architecture.

National Library of New Zealand: Initiative. (2007). Retrieved Feb 2007, from http://www.natlib.govt.nz/en/whatsnew/4initiatives.html#extraction.

Ockerbloom, J. M. (2006). Proposal for support of versioning in the new DSpace architecture.

OCLC/RLG PREMIS Working Group (2004). Implementing preservation repositories for digital materials: current practice and emerging trends In the cultural heritage community. Report by the joint OCLC/RLG Working Group Preservation Metadata: Implementation Strategies (PREMIS). Dublin, Ohio, OCLC Online Computer Library Center.

PADI. "Digital preservation strategies." Retrieved Feb 2007, from http://www.nla.gov.au/padi/topics/18.html.

PADI. "Preserving Access to Digital Information." from http://www.nla.gov.au/padi/index.html.

Preston, W. C. (2007). "Amanda chapter in Backup and Recovery." Backup & Recovery Retrieved Feb 2007, from http://wiki.zmanda.com/index.php/Amanda_chapter_in_Backup_and_Recovery.

Rooney, D. (1997). A Contextualising, Socio-Technical Definition of Technology: Learning from Ancient Greece and Foucault. *Prometheus*, 15: 399-407. http://www.acro.edu.au/rooneySOCIOTECH.pdf

Sommers, F. (2005). Catch Jackrabbit and the Java Content Repository API.

Tansley, R. (2003). DSpace as an Open Archival Information System: Current Status and Future Directions, Hewlett Packard Laboratories.

Tansley, R. (2004). DSpace 2.x Architecture Roadmap, DSpace Technical Lead, HP.

Tansley, R., M. Bass, et al. (2006). DSpace System Documentation.

TASI "Metadata Standards and Interoperability."

TASI. "Digital Preservation and Storage." from http://www.tasi.ac.uk/advice/delivering/digital.html.

Team, F. D. (2005). Fedora Open Source Repository Software: White Paper.

The National Archives. (2007). "The technical registry: PRONOM." Retrieved Jan 2007, from http://www.nationalarchives.gov.uk/aboutapps/pronom/tools.htm.

Wikipedia. (2007). "LiveDistro." Retrieved March 2007, from http://en.wikipedia.org/wiki/LiveCD.

Wikipedia. (2007). "Virtual machine." Retrieved March 2007, from http://en.wikipedia.org/wiki/Virtual_machine.