Date : 01/07/2008

Media Matters: developing processes for preserving digital objects on physical carriers at the National Library of Australia

Authors: Douglas Elford, Nicholas Del Pozo, Snezana Mihajlovic, David Pearson, Gerard Clifton, Colin Webb (presenter), National Library of Australia, Canberra, Australia

Meeting: **84. Preservation and Conservation, (PAC), Information Technology, IFLA-CDNL Alliance for Bibliographic Standards (ICABS) and Law Libraries**

Simultaneous Interpretation: Not available

WORLD LIBRARY AND INFORMATION CONGRESS: 74TH IFLA GENERAL CONFERENCE AND COUNCIL
10-14 August 2008, Québec, Canada
http://www.ifla.org/IV/ifla74/index.htm

Abstract:

The National Library of Australia has a relatively small but important collection of digital materials on physical carriers, including both published materials and unpublished manuscripts in digital form. To date, preservation of the Library's physical format digital collections has been largely hand-crafted, but this approach is insufficient to deal effectively with the volume of material requiring preservation. The Digital Preservation Workflow Project aims to produce a semi-automated, scalable process for transferring data from physical carriers to preservation digital mass storage, helping to mitigate the major risks associated with the physical carriers: deterioration of the media and obsolescence of the technology required to access them. The workflow system, expected to be available to Library staff from June 2008, also aims to minimise the time required for acquisition staff to process relatively standard physical media, while remaining flexible to accommodate special cases when required. The system incorporates a range of primarily open source tools, to undertake processes including media imaging, file identification and metadata extraction. The tools are deployed as services within a service-oriented architecture, with workflow processes that use these services being coordinated within a customised system architecture utilising Java based web services. This approach provides flexibility to add or substitute tools and services as they become available and to simplify interactions with other Library systems.

This paper describes the Library's development of these systems and its expectations for future work.

Introduction

This paper describes the work of the National Library of Australia in designing a workflow solution for the highest priority preservation risks confronting its collection of digital materials on physical carriers such as handheld optical and magnetic disks. The solution discussed is neither a perfect nor a complete one, but it does seek to

address the NLA's identified time-critical risks for these materials, and takes the Library forward in an area previously characterised by unsustainable and inadequate hand-crafted preservation actions carried out with inadequate resources.

Background – The 'problem space'

The National Library of Australia's main statutory role is to build, maintain and provide access to a national collection of information resources about Australia and the Australian people. Since the early 1980s the national collection has included digital materials of various kinds. While the main focus of outside interest is usually the Library's program of collecting and managing Australian online resources through its PANDORA archive (http://pandora.nla.gov.au/), the Library also holds a comparatively small but important collection of digital materials on physical carriers. This collection includes both published materials and unpublished 'manuscripts' in digital formats, many pre-dating the development of the Web and any attempts to preserve materials from it.

It is in some ways misleading to talk of these materials as a 'collection'. In fact, the materials are held in various general and special collections throughout the Library. Significantly for the present workflow project, they come into the Library through various deposit, acquisition and processing streams, on a variety of carriers types. Most items fall into one or other of four carrier categories: CD-ROMs, DVD-ROMs, 3½ inch floppy disks and 5¼ inch floppy disks. All of these carriers are considered to be temporary storage media only, and accessing data on them depends on the availability of appropriate hardware and software. In addition, the ongoing accessibility of logical file formats held on the media depends on operating systems and software applications, which may or may not be available.

Over the years, the balance of physical carriers in the Library's digital collections has changed, reflecting changing technologies and the preferences of publishers, creators and users. There has been a steady increase in acquisition rates which is predicted to continue for some years to come.

Some brief history

By the mid-1990s, the Library recognised that it held and was actively collecting a lot of physical format digital materials, but had virtually no arrangements in place or planned for keeping them accessible. Over the next few years we undertook a range of initiatives aimed at improving this situation. These initiatives included:

- Surveying what we held, where they were, what we knew about them, and how accurately or usefully they had been described in cataloguing
- Improving cataloguing processes to more accurately describe these materials
- Negotiating criteria for prioritising the materials for preservation attention
- Developing a procedure for copying content from floppy disks to CD-ROM media as a first step in improving the longevity of their content data
- Developing a procedure for documenting the technical dependencies of incoming materials

- Developing a framework procedure to guide attempts at recovering data where access was considered to have already been lost (usually because we simply didn't know what file formats had been used)
- Negotiating general procedures for collection curators and processing staff finding digital materials in collections being acquired or processed
- Establishing procedures for materials to be routed to Digital Preservation for assessment, documentation and/or copying.

These procedures were followed with varying levels of resourcing over a period of seven to eight years, during which time we realised that:

- No matter how enthusiastically we followed this approach we could not keep up with the level of intake
- We were hand-crafting processes that could be done much more efficiently with some automation
- The processes involved double-handling because the processing staff found them too time consuming to fit into their existing workflows
- We were only addressing floppy disks and no other kinds of materials, even though the intake of other materials was far outstripping the intake of floppy disks
- Even for floppy disks, we were only buying a little time by moving their contents to CD-ROMs.

By 2004 these manual procedures had been largely put aside under the pressure of other priorities, leaving a growing backlog of highly exposed materials.

Drivers for finding a better approach

In 2003, the Library undertook a detailed risk assessment of all its digital collections – web archives, physical format materials, digitisation copies, digital audio and digital image files, and so on (Dack: 2004). The likely loss of data on physical carriers was identified as the most pressing risk facing the Library, because of the expected short-term deterioration of the carriers, possible corruption of the data on them and the likely loss of availability of hardware required to access often 'one of a kind' content.

The physical carriers in the collections have a high vulnerability to loss. The chemical and physical composition of the carriers themselves, and the way data is recorded on them, mean that their viable life expectancy can be as short as a few years. Experience has shown that while there is a typical deterioration curve, the position of any specific item on that curve at any particular time cannot be reliably predicted. Given that many of the items held by the Library dated back more than a decade, our risk assessment identified these materials as being at high risk of early loss.

The availability of software dependencies is obviously also a critical issue, but one with a longer lead time. If access is lost, software obsolescence becomes a moot point. Steps must be taken to preserve the bit stream that encodes the data before media deterioration or media obsolescence occurs.

Improving storage conditions for physical media offered some promise of slowing down the natural rate of deterioration, but it could only be seen as a stop gap measure. Getting the data off its short-term carrier and into more reliable managed storage was seen as the most pressing priority.

However, identifying and securing the chain of multiple hardware and software components required to access the data on the carrier at the physical level poses a challenge even before basic copying or further analysis can take place. There was a recognised need to maintain the technical environments in which these digital objects worked, or at least to record enough information about those environments to recreate them when needed, or to transform the objects so they would work in different technical environments. This need implied two imperatives: the importance of understanding the technical environment and the value of holding on to components of those environments so the data can be processed, even if only to retrieve the data for transformation to another format or storage environment.

The Library's existing cataloguing and documentation systems were acknowledged to be inadequate for capture of the required level of metadata that could maintain useful linkages between objects, components of the technical access environment, and information about which components need to be maintained, for what purpose and for how long.

The Library's existing procedures and workflows were unable to address these risks in the required volume, with the necessary speed, or across the required range of materials. Therefore we had to move away from hand-crafting to a more industrial way of processing this material.

The context of the problem

While aspects of the problem will be common to many collections holding physical format digital objects, understanding the context of the risks in the National Library of Australia may help in deciding the extent to which the Library's approach is potentially applicable elsewhere.

The NLA collects digital materials through multiple acquisition streams, and generally has had little control over the carrier formats on which the materials arrive. While most items fall into a small number of widely used carrier formats, the full range of carriers in the collections is much wider. Any long-term solution for the NLA has to make provision for almost any kind of carrier, while seeking efficiencies in processing the predominant formats.

The NLA's collections include both published and unpublished materials, the latter usually from personal collections and likely to be idiosyncratic in a number of ways. Frequently, unpublished digital materials are transferred to the Library without detailed manifests of what is included and often with very limited (if any) descriptions of content, file types, applications required for access, and so on. While the information content can usually be assumed to be unique, its importance is sometimes ambiguous until the data can be viewed.

On the other hand, the published items tend to be more easily dealt with because they usually carry some information on their packaging; they also often come with bundled software. In many cases copies are held by other libraries, and some items remain "in print" and replaceable for some time. However, published materials are more likely to be problematic to preserve because of copy protection measures.

Published or unpublished, it is not possible to know just by looking at the carrier whether the data will be readable or corrupted, and it may not be immediately apparent whether the Library holds software or even the hardware needed to access and read the data.

In considering the need for a digital preservation workflow solution, an analysis of a sample of physical media log file records, recorded to March 2006, produced the following statistics about media types (Elford: 2008).

Physical Media *(with one or more per collection record)*	**No of Collection Records**	**% of total sample records (2458)**
3 1/2" floppy disk media	287	11%
5 1/4" floppy disk media	15	0.6%
CD-ROM media	2103	86%
DVD-ROM media	43	2%
other media	10	0.4%

The variety of operating systems, software applications and file formats noted in the sampled log files, including releases spanning a 20 year period, also indicates that there is considerable diversity that needs to be managed at the logical level.

It is estimated that the current total number of physical media items selected for preservation is approximately 9,000. Current growth is estimated to be approximately 1,200 -1,500 items per year (around 120 selections per month).

There are factors likely to lead to significant increases in the current rate of intake, as well as the diversity of carrier types and access dependencies. The NLA operates in a context of having no legal deposit provisions for electronic materials. The Australian Government has recently conducted a review to consider changing this situation. If legal deposit provisions are extended to electronic materials as a result of the review, the Library would expect an increase in the physical format digital publications coming into its collections.

At the same time, there is increasing interest among the Library's manuscripts and music curators in collecting unpublished materials in digital forms created by writers,

musicians, public figures, and so on. There is also an increased interest amongst researchers in having access to such materials. This is expected to lead to significantly increased acquisition of unpublished digital materials. Therefore, the Library is confronted by both a large backlog of material to be processed, and by rapid growth in collection intake.

The Digital Preservation Workflow – The 'solution space'

The underlying aim of the Library's Digital Preservation Workflow Project was to provide a semi-automated and scalable means of moving data from physical carriers to a digital mass storage system for ongoing preservation management, offering at least the same guarantees of bit stream security as the Library's other digital collections.

The system is intended to improve on the previous largely manual processes in a number of ways, only adding marginally to the work processes of existing staff who are processing the same materials for other purposes, while minimising the additional time and skills required by those staff, and retaining flexibility for Digital Preservation specialists to intervene if special attention is needed.

The system allows staff in collection processing areas to:

- Accurately capture data from physical format digital carriers, or from other digital sources put through the workflow
- Generate and capture related metadata into a database
- Automatically invoke a range of modular services to analyse, validate and process the data objects, depending on the type of material and current policies on how material will be processed
- Upload the captured data and related metadata to the Library's managed mass storage system
- Provide limited access to materials, subject to permissions and the availability of suitable software
- If necessary, refer materials that cannot be successfully processed by the system to Digital Preservation for special hand-crafted attention.

The system allows future integration of services that will enable transformation of older file formats, or the association of alternative rendering paths that would allow older file formats to be reliably rendered.

Although initially envisaged as working through a centralised processing point (much like many digitisation projects), using a large multi-drive jukebox, the system design evolved to a decentralised model in which the most common physical format media would be efficiently processed and ingested during standard cataloguing processes via a number of 'mini-jukeboxes'. These small portable jukeboxes can be attached to a user's computer. They can accommodate a flexible array of physical carrier drives and hence can be tailored for specific collection areas. The less common and more difficult carriers are diverted to Digital Preservation where a wider range of drives is available.

In its initial version, the system is primarily intended as a route to preservation storage, while capturing supporting metadata and undertaking some processing that will facilitate later preservation action. The current system is not intended to provide full collection management and preservation capability as such capability was beyond the scope of the resources available.

At this stage, the system is intended to deal with the most pressing preservation risks, but it is also designed to contribute significantly to long-term preservation, and to interoperate with a future full preservation management system. Once the at-risk data has been stored safely, future work can focus on longer-term accessibility strategies and systems, using the metadata harvested during the automated ingest process, and using further format level preservation assessment and planning to deal with file format obsolescence.

The initial version of the Digital Preservation Workflow System will be available to Library staff after June 2008. After completion of the system rollout, a more in depth analysis and description of the project along with its software and associated documentation will be made available via the project's SourceForge site (www.sourceforge.net) under an open source license. Some of the more interesting aspects of these workflow services are briefly outlined below. A high level illustration of the Digital Preservation Workflow System is provided in APPENDIX 1 Figure 4.

A high level view of the general process is illustrated below. Many of the activities are performed automatically by the system, with few user tasks required.

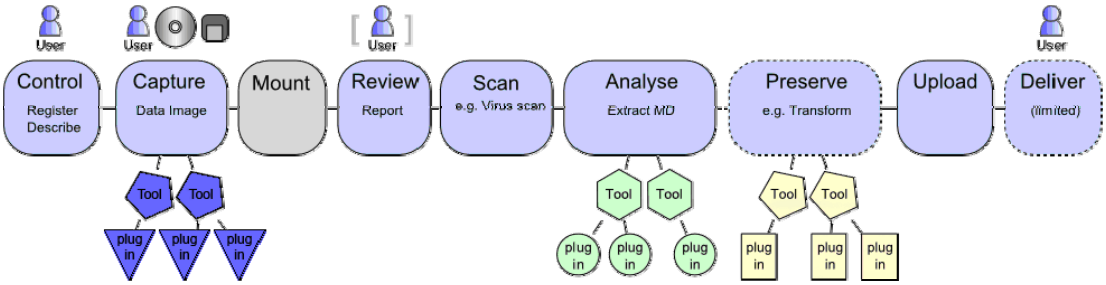

Figure 1: General Process View

Workflow Implementation

Overall we have taken some distinctive approaches to build services.

- a RESTful [Representational State Transfer (REST) style services] (Fielding, 2000) approach is used to build and expose most of the digital preservation workflow services such as mounting and unpacking, file analysis, and metadata extraction services
- the client software is exposed through a listening port. This software is installed on the user's Windows-based computer
- a Queuing service is utilised to ensure reliable messaging
- all services talk to each other through XML

Informal assessment was made of a number of integrated workflow systems and Application Programming Interface (APIs) and although one was chosen for a more detailed investigation, we concluded that none adequately fulfilled our specific requirements. We therefore elected to build a customised system architecture using Java based web services. This enables a flexible, extensible and adaptable solution wherein various plug-in tools can be updated or replaced by new alternatives with minimal impact on the system as a whole. These are discussed further below.

The system user interacts with the workflow system via a web interface from a standard desktop environment which a Java-based client software installed along with a multi-lane SATA card, to enable connection of a mini-jukebox. The client software allows the workflow system to execute programs on the user's computer.

The web based user interface leads the user through the process of initiating a job, associating a bibliographic record with the job via the Library's pre-existing Catalogue service call, adding a physical media part to the digital holdings record, selecting the type of physical carrier from a pop-up menu and inserting the carrier into the appropriate drive (the system analyses the hardware set-up and, for optical media, automatically opens the appropriate drive).

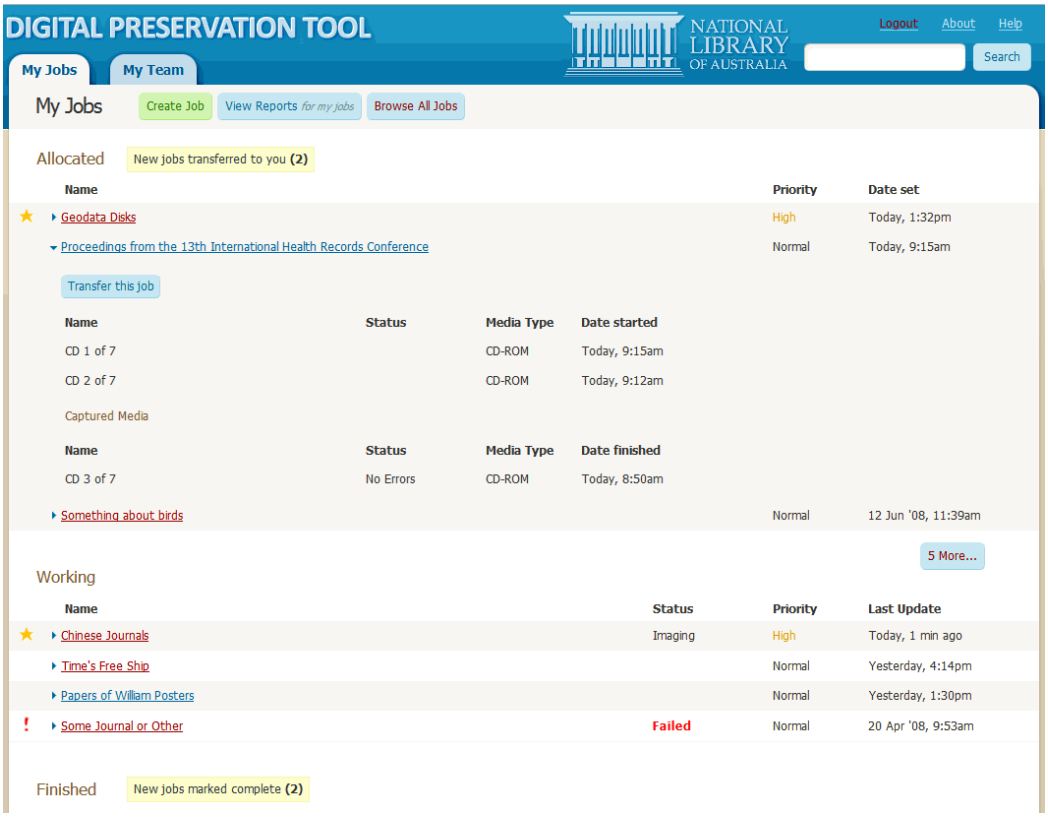

Figure 2: Pre-release User Interface screenshot for the Digital Preservation Workflow System.

Finally, the user initiates the media imaging process. We have found that obtaining a bit-level image from various source media requires a number of different open source tools. Based on the user's selection, the system automatically ascertains the appropriate tool to image the media. Due to the modular nature of the architecture, the initial tool set can be upgraded or replaced over time as required. The current implementation uses the following tools:

Media Type	**File System**	**Tools**
CD-ROM	e.g. ISO9660, HFS,	CDRDAO (version 1.2.2)
VideoCD	e.g. ISO9660 + video file	CDRDAO (version 1.2.2) + VCDXRIP (version 0.7.23)
DVD-ROM	UDF, UDF Bridge	dd (version 0.5) [will not work on copy protected discs]
Block Device (e.g. USB Flash, Hard Disk)	any (e.g. NTSF, FAT32, FAT64, HFS, HFS+, EXT2, ZFS)	dd (version 0.5)
Floppy Disk (3 1/2, 5 1/4)	any (e.g. NTSF, FAT32, FAT64, HFS, EXT2, ZFS)	dd (version 0.5)

These tools are automatically invoked by the client software on the user's computer. The appropriate software analyses the media and creates a bit level image, a sector-by-sector copy of raw data as it is recorded on the media. This captures all the data available on the media including deleted and hidden files, as well as maintaining as closely as possible the original structure of the data.

The user can image multiple media simultaneously or have them running in the background while performing other tasks. The user is shown a progress bar and process status of the operation and upon completion is prompted to remove the media. After the disk image is completed, a checksum is calculated and the disk image is transferred to a mass storage working area, whereupon the transfer is verified by recalculating the checksum.

The first step in processing the image is to mount it. Disk images generally could contain multiple file system types such as ISO9660, NTFS and HFS. Therefore we have chosen to mount the images on a Linux server, as Linux already supports known major file system types and can also be adapted to new file systems. Once the disk image is mounted on a Linux server it can be read and accessed as though it were a physical disk, allowing further analysis to be carried out without compromising the original media. The structural arrangement of the files within the image is analysed and recorded as a manifest in the form of a METS StructMap complying with the Australian METS profile. The manifest can be displayed to the user as a directory tree view for reviewing and for cross checking against any other available manifest.

The system continues to process the disk image automatically, calling a Virus Scanning service and an Analysis service. The Analysis service utilises a combination of several sub processes which implement various tools such as DROID, JHOVE and the National Library of New Zealand's Metadata Extractor to extract information

about each individual file in the mounted disk image, storing the output using the appropriate PREMIS-related metadata fields.

Information about each file, such as its persistent identifier and location, is passed to the Analysis Service. We are currently using the following software in the analysis services.

- Checksum for each file (SHA-1 via Jacksum version 1.7.0)
- Virus check using ClamAv (version 0.93)
- File identification using DROID (version 2.0)
- Metadata extraction services using JHOVE (version 1.1) and the National Library of New Zealand's Metadata Extractor (version 3.4GA) tools.

The arrangement of the files and directories, file names, sizes, and dates, can be determined by the operating system. Recording this information is an important part of maintaining the relationships between files that may be needed for the material to function; for example, to allow linked spreadsheet documents to draw on data across a number of files. It can also provide context for further analysis, as well as representing the way in which the material was organised and related by the creator, maintaining respect for the original order. To enable monitoring of file integrity, the system calculates file hash signatures or checksums that can be later recalculated to verify that files have not been altered in the interim.

The metadata that is extracted throughout the process is stored in a supporting database developed for the project. The data model created for the workflow system was designed to be capable of exporting to METS as well as other formats as required. For example, data could be exported using the recently developed Australian METS profile (Pearce: 2008).

Physical Carrier Ingest via Portable Mini-Jukebox

To extend the input capacity and reliability of the existing Standard Operating Environment (SOE) computers used by staff so that they can best utilise the Digital Preservation Workflow software, the system uses portable, multi-drive 'mini-jukeboxes'. This approach enables a more flexible and distributed method of ingest compared with using either standard limited specialist workstations or a centralised jukebox.

The mini-jukebox accommodates a flexible arrangement of between 1 to 4 drive bays that can be customised to specific collection area requirements, to deal with 3 1/2" floppy disks, CD-ROMs and DVD-ROMs. It also enables the use of better quality and more reliable drives than are currently deployed in the SOE workstations (Elford: 2008).

An expected increase in the number of flash thumb drives and materials collected for upload from external hard drives can be accommodated by using a 'temporary scratch disc' on an optional internal hard drive inside the mini-jukebox, before transferring the imaged data into the system. Alternatively, the workflow system software could be loaded on to such an internal hard drive, so that the existing mini-jukebox units could then be made more transportable and more easily taken to the data source.

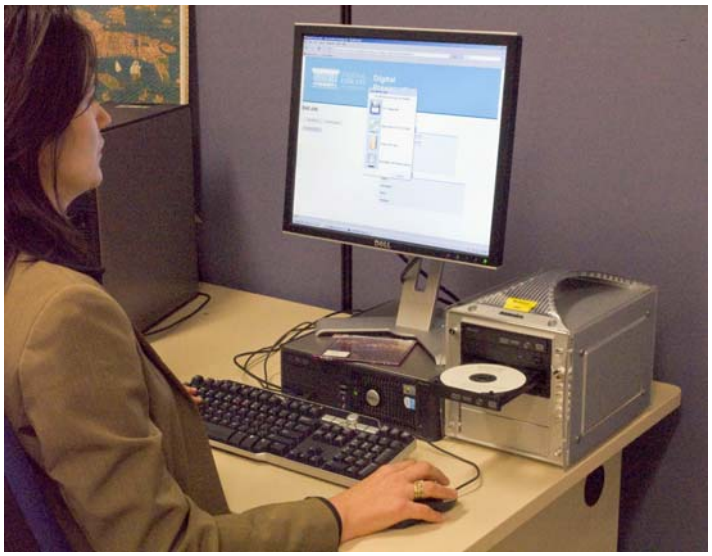

Figure 3: User selecting media for imaging via a mini-jukebox attached to a SOE workstation

One of the most difficult aspects of proceeding down this route was finding the right hardware/software combination given we were using a Windows-based environment. Although we initially looked at a USB 2.0 based solution, the current generation Digital Preservation mini-jukeboxes uses SATA as a transfer protocol, delivered via a single multi-lane SATA cable. This decision was based on a number of considerations. Firstly, the transfer rate on USB is quite slow, especially if multiple devices are sharing the same connection. Secondly, there are a number of issues that we encountered integrating our system within the Windows environment. The Library's Windows-based SOE computers would not map USB devices consistently enough for the workflow software to detect and connect to, especially if other devices were already connected such as USB thumb drives, card readers, MP3 players or network drives. For example, we found that CD-ROM devices connected via USB would lose their drive mapping over time. We were able to overcome this by utilising a multi-lane SATA connection to the jukebox, for which we can hard-code the drive mappings.

Although this restricts the portability of the solution somewhat, the mini-jukeboxes can still be shared amongst those computers that have had a multi-lane card installed. While Windows' method for assigning drive letters requires each drive in the mini-jukebox to be manually assigned a mapping, this only needs to be done once per drive, per computer when the application software service is initially being installed. An ideal solution would allow us to connect drives to any computer, and to have those drives automatically accessible by our software. However, we did not find a feasible way of doing this within our timeframe.

Spin-offs from the Physical Carrier Workflows Project

The Workflows Project has required the creation of services, tools and methodologies that are expected to have a much wider application.

Project Management

The manner in which the project was run was conducive to the outcomes of the project; as a cooperative initiative, led by the Digital Preservation section, with both dedicated and consultative resources drawn from the IT Division. Recurring reference to other Library stakeholders and intended users helped guide project goals. This organisational approach provided a deep level of engagement between sections and provided some hedge against the severe difficulty in recruiting specialised staff for the project.

The project team included a project manager, two digital preservation specialists, two programmers and a business analyst, all co-located to facilitate coordination and exchange of ideas. Additional systems architects, standards and metadata specialists were coopted as required. This approach was innovative within the Library's history of developing digital preservation solutions.

The project aimed at agile prototyping with a series of iterative releases to generate feedback and refine both requirements analysis and system design. This approach has worked for us in this instance and has also provided a prototype for future projects.

Mediapedia Service

A well-received outcome of the project has been the initial planning of the Mediapedia database service for physical carrier media. Mediapedia (www.nla.gov.au/mediapedia/) This service is intended to aid in the discovery and prioritisation of risks associated with various physical carrier media types by providing basic technical details on a wide range of media formats, their hardware and software dependencies for access, and associated conservation information. Using a classification system that covers a multitude of physical and analogue media, the Mediapedia system may be used to categorise and capture previously undocumented business knowledge from a variety of sources including IT, Digital Preservation and external industry sources. Access to this information assists Library staff in identifying media and the associated handling requirements and risks in material offered for collection assessment or already being processed. Plans include making this service available both internally and externally by implementing a community driven web based information resource.

Environment Service

Another service to be developed, to support the workflow system, is an Environmental database, which will provide file format access information. This database will store information about various software and hardware components, and create relational links between these items to describe the dependencies required to access specific file formats. These dependencies can be based on various

'characteristics' for each file, such as the languages or fonts used in a text document, as well as the file format itself.

As with the Mediapedia Service, this service will be designed as an open and collaborative community registry of information, where users can not only contribute and retrieve their own sets of access dependencies, but also information about dependency sets that have been contributed by other institutions or individuals. The data will be expressed in an implementation neutral fashion and designed to easily integrate with existing systems.

Conclusion

Although our ideal would be to retrieve all information from all media types, this may not be a realistic expectation. Factors such as the physical degradation of media over time, or use of Technological Protection Mechanisms make the retrieval of all information impossible for some items. Even within the Library we are dealing with a process of 'differential survival', calling for management decisions about whether to focus on dealing with the majority of our digital materials, or rather on the small subset of the most difficult materials. The NLA's Digital Preservation Workflow System is intended to reconcile that dilemma by automating processes for the majority of physical format digital materials, allowing specialised Digital Preservation assets to be used on more problematic material as required.

Residual issues

What is needed to support the system

Any system for dealing with the preservation of digital materials on physical carriers requires a framework of policy and organisational procedures to guide the way it is used and to resolve dilemmas as they arise. The NLA still has to finalise or refine the following strategies and documentation to support its Digital Preservation Workflow System:

- Guidelines and training for new work practices and workflow processes
- Guidelines on identifying simple and complex materials
- Guidelines on discarding digital materials
- Guidelines for acquisition staff on recognising and dealing with offered materials which will be difficult to preserve
- Policy on the retention of previous preservation versions when digital content is transformed to new preservation versions
- Policy on appropriate provisions for copyright compliance and on dealing with materials published with Technological Protection Mechanisms (TPMs)
- Policy on the preservation of non-Australian materials
- Policy and planning related to future legal deposit scenarios

- Design of a manifest template for depositors and acquisition staff
- Research, planning and implementation of strategies to recognise and overcome file format obsolescence risk and to ensure content remains deliverable, and
- Resource strategies to deal with the backlog of unprocessed materials.

Proposed future extensions of the system

As well as making modifications in response to user and administrator feedback and system review, the following extensions are expected to need attention to maximise the usefulness of the new system:

- Addition of new tools or versions of tools and services, including tools for undertaking preservation actions such as normalisation
- Acceptance of a wider range of source materials, including embedded media files, archival data sets, and email attachments

Interoperation with currently non-connected or yet to be developed Library systems including a universal upload system, delivery systems, and the records management

Bibliography

Dack D. (unpublished) 2004. **An Assessment of the Risks to the National Library of Australia's Digital Collection**, National Library of Australia, Canberra.

Elford D. (unpublished) 2008. **Legacy Physical Media Review**, National Library of Australia, Canberra.

Elford D. (unpublished) 2008. **Mini Jukebox Requirements Report**, National Library of Australia, Canberra.

Fielding R. 2000. **Architectural Styles and the Design of Network-based Software Architectures**, http://www.ics.uci.edu/~fielding/pubs/dissertation/

Pearce J., Pearson D., Williams M., and Yeadon S. 2008. **'The Australian METS Profile – A Journey about Metadata' in D-Lib Magazine, Volume 14 Number 3/4 March/April 2008**, http://www.dlib.org/dlib/march08/pearce/03pearce.html

References

Australian METS Profile: http://www.loc.gov/standards/mets/profiles/00000019.html

ClamAv: http://www.clamav.net/

DROID: http://droid.sourceforge.net/wiki/index.php/Introduction

JHOVE: http://hul.harvard.edu/jhove/

Mediapedia: www.nla.gov.au/mediapedia/

National Library of New Zealand Metadata Extractor: http://meta-extractor.sourceforge.net/

PANDORA archive: http://pandora.nla.gov.au/

SHA-1 via Jacksum: http://www.jonelo.de/java/jacksum/

SourceForge: www.sourceforge.net/

The system's deployed capabilities

The system:

- Accepts CD, DVD, USB thumb drives, hard disks, and a range of floppy disk sizes
- Accepts complex, old or damaged materials, via the Digital Preservation section
- Accepts items from a range of operating system platforms – e.g. DOS, Windows, Macintosh & Linux
- Accepts items in multiple natural languages: e.g. English, Asian languages
- Extracts metadata from a range of file formats
- Documents, but may not be able to deal with items carrying copy protection measures.
- Aims to achieve processing times of no more than 10 minutes per item for staff working in existing accessioning or cataloguing workflows.
- Allows items to be processed individually or in a batch process.
- Allows staff to process most items in one pass
- Is usable from distributed locations within the Library
- Is scalable, capable of handling a high volume of items in a multi-user environment
- Is based on a pluggable, configurable, flexible and extensible architecture to allow a range of tools and process modules to be inserted
- Allows workflows to be configured for different material types
- Is secure
- Provides an authenticity trail of events related to processing
- Uses a standards-based metadata schema for representing item structure and metadata
- Operates in the Library's developing Service Oriented Architecture
- Interoperates with other existing Library systems
- Uses desktop jukeboxes configured to accept a range of media types and deployed to collection areas
- Uses a web-based user interface.

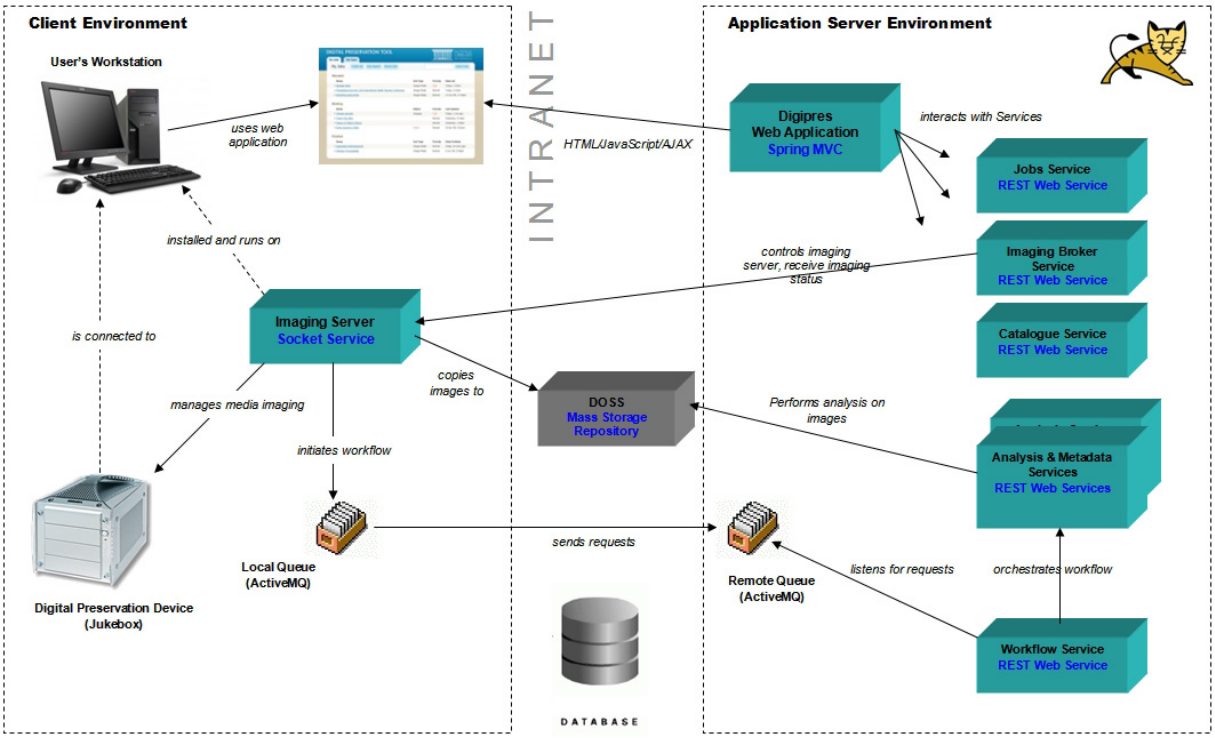

Figure 4: High Level Visualisation of the System Architecture