

Bringing the stories to the people: online sound at the National Library of Australia

Marie-Louise Ayres
Associate Curator of Manuscripts
National Library of Australia
mayres@nla.gov.au

Judith Pearce
Director, Business Analysis
National Library of Australia
jpearce@nla.gov.au

Dianne Dahlitz
Senior Librarian, Oral History
National Library of Australia
ddahlitz@nla.gov.au

This paper describes the National Library's project to make web delivery of its sound collections possible. The paper outlines the key infrastructure components required for routine web delivery of streamed sound files and XML encoded transcripts and summaries from a very large collection, and the standards adopted and/or adapted to facilitate delivery. It also describes the content conversion strategies required to convert analogue recordings to digital format, and to convert printed summaries and transcripts to encoded documents. The paper will conclude with a demonstration of the Library's online sound delivery system, and a brief view of the multiple discovery pathways to these remarkable resources.

Introduction

Over the last 10 years, the National Library of Australia has progressively digitised more than 100 000 collection items – in a number of different formats – and made them available online. Working on the principle of ‘easiest to hardest’, the Library has developed digitisation standards and workflows, persistent identification schemes, digital collections management procedures and applications, and digital delivery systems for pictures, printed music, maps, manuscripts and books. Each collection type has brought its own unique digitisation, management and delivery issues. Printed music and books, for example, require administrative metadata sufficient to allow users to ‘turn’ pages. Maps require zoom and pan functionality. Images of manuscripts need to ‘hang’ off Encoded Archival Description finding aids.

The Library holds a 38 000 hour original sound recording collection, consisting principally of oral histories and folklore recordings (many with timed summaries and up to 30% with transcripts). These resources have hitherto only been available to onsite users, through the Inter Library Loan service, or to users purchasing physical copies of sound recordings and associated documentation. These 38 000 hours contain many of Australia’s most important stories, ranging from interviews with senior politicians, leading arts figures and professionals to the stories and recollections of ordinary Australians from many walks of life, to archival performance recordings of music in the Library’s collection. Making these wonderful stories and sounds available to a much wider audience is a high priority for the Library: considerable intellectual and financial investment has therefore been dedicated to this end.

This paper describes the Library’s project to make web delivery of its sound collections possible. The paper outlines the key infrastructure components required for routine web delivery of streamed sound files and XML encoded transcripts and summaries from a very large collection, and the standards adopted and/or adapted to facilitate delivery. It also describes the content conversion strategies required to convert recordings to BWF (Broadcast Wave Format), and to convert printed summaries and transcripts to encoded documents. It outlines the digital rights management infrastructure to be implemented in a further stage of the project, supporting mediated online access to items with restricted access conditions.

In addition, the paper discusses the cultural and workflow changes required to update existing acquisition, technical and descriptive procedures to recognise the new reality of online sound delivery. Problems encountered during the project – including the need to move from a preservation-focused management strategy to one focused on both preservation and access, a forty-year legacy of less than transparent access conditions, more than two thousand hard copy only transcripts, the challenges of using proprietary streaming formats, and the difficulties of prioritising conversion work – are outlined. The paper will conclude with a demonstration of the Library’s online sound delivery system, and a brief view of the multiple discovery pathways to these remarkable resources.

The collection

The Library acquires approximately 800 hours of recorded material each year, much of which is commissioned work. Employing specialist interviewers, it aims to record 100 interviews for the Eminent Australians program each year. This includes interviews with Australians prominent in fields such as the arts, publishing, science, law, politics, government, music, technology, the environment, Aboriginal history, sport, education, religion, immigration, industry, commerce and journalism.

Other programs are the Social History and Folklore collections. Through the Social History program the Library records interviews documenting significant social trends and conditions. Over the past few years several large projects have been conducted in collaboration with other institutions: these include the Cultural Context of Unemployment collection, and the Australian Response to AIDS collection. In recent years, the Library has focused on multicultural collecting, including projects such as the Maltese Australians, Polish Australians and Chinese Australians. The high profile project Bringing Them Home, the stories of removed Aboriginal children, contains over 330 interviews. Current social history projects include the Tradition Bearers and the Drivers.

Through collecting for the Folklore program, an extensive and important collection of field recordings has developed. Traditional Australian folk song and dance music is included in the early collecting of John Meredith, Norm O'Connor and Alan Scott, and in more recent years that of Chris Sullivan and Rob Willis.

The project

The National Library of Australia has been planning web delivery of its sound collection for several years, partly because it wants to provide easy access to all its original formats, and partly to support our specialist online services, including MusicAustralia and Australia Dancing. In line with the Library's policies and practices for other collection formats, we knew that we needed a robust, sustainable, standards-based and 'whole of collection' solution, which would result in routine digitisation and delivery of this content in the future. These policies – and the sheer size of the collection – mean that the hand-crafted solutions for online sound delivery in place in some Australian and a number of international institutions were not suitable for our purposes. There were no 'best practice' examples of online delivery of a substantial oral history and folklore collection, and in fact it seemed that the international community was waiting for the Library to develop solutions that others could learn from. We knew from the outset that this would be a major project, presenting new challenges to the Library's digital services team, and requiring several different components.

Our early assessment of requirements was that in order to deliver a single oral history or folklore recording via the web, the following conditions must be met:

- The sound recording must be preserved digitally – as a high quality, uncompressed audio file – on the National Library of Australia's Digital Object Storage System (DOSS), a mass storage system used for all the Library's digital collections.
- Summaries and transcripts must be encoded as XML documents to support a wide range of user activities, and to support data-sharing. These documents must also be stored on the DOSS.
- The recording, summary and transcript must be managed by the Library's Digital Collections Manager (DCM) application, built in-house over a five year period because no commercial solutions support all our requirements. The DCM supports a wide variety of digital management tasks: preservation; creation and management of 'derivative' copies suitable for delivering online; storage of administrative and technical information; and rights management.
- A web delivery system to provide end user access to the sound file, and to any associated text summaries and transcripts must be in place. The system must include: the ability to play sound over slow network connections; the ability to read a summary and/or transcript while listening; the ability to 'jump' to points of interest in the

summary/transcript and simultaneously move to the correct point in the sound recording; the ability to switch between summary and transcript ‘views’; and the ability to navigate within and search the summary and/or transcript or a wider set of digital resources.

- The web delivery system must also be able to deal with complex content models – a fragment from a larger recording, a performance with multiple performers, a conference held over several days, a collection of field recordings, a group of interviews conducted as part of the same project. The system must include the ability to navigate from fragment to whole recording to collection and back again.
- There must be no restrictions on public use of the recording or the transcript for ‘open’ delivery. Restricted items should be made available to users with appropriate permissions via a secure extension to the delivery system.

This assessment led us to plan our project in three stages:

- Managing the digital collection;
- Delivering the digital collection; and
- Preparing content for delivery.

Managing the digital collection

Digital Object Storage

The National Library of Australia has long been an international leader in the preservation of sound recordings. The Library has a dedicated and internationally recognised team devoted to this part of our preservation responsibilities. As a result, approximately 60% of the National Library of Australia’s 38 000 hour audio collection has been preserved: 30% has been digitally preserved on the DOSS; 9% has been digitally preserved onto CD-R; and 21% is preserved in analogue format. The Library’s oral history and folklore items were collected on analogue tape until 1992, when collection on Digital Audio Tape (DAT) commenced. From 2005 collecting direct to hard drives commenced, and the WAV files generated are then transferred to the DOSS.

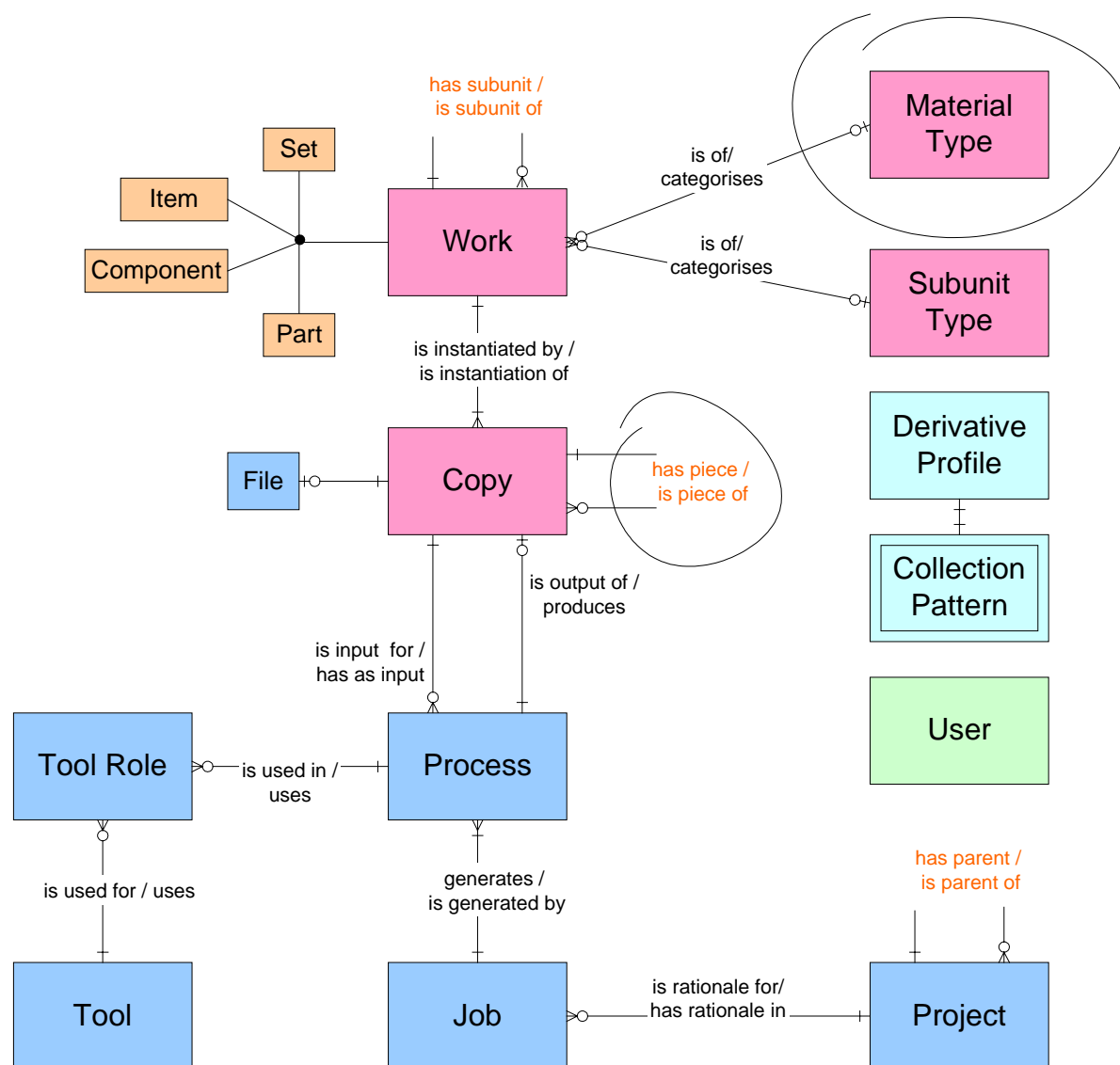
The Library uses the Quadriga Digital Audio Workstation system which manages conversion of analogue and digital formats to BWF (Broadcast Wave Format), and ‘ingestion’ into the DOSS. Conversion processes occur in ‘real time’ for analogue and DAT original material.

The Library has a 15 year plan to migrate the entire sound collection (including annual intake) to the DOSS, and is currently more than 2 years into this plan. The entire collection should therefore be on the DOSS by 2018. The Library currently has a storage area network 14 terabytes in size as well as a tape library able to store a maximum of 60 terabytes, of which 30 terabytes are currently in use. The tape library enables files that are not frequently used to be stored with a ‘nearline’ status and recalled at need. By 2018 it is anticipated that the sound collection alone will occupy around 100 terabytes of storage.

Digital Collections Management

In 2005, the Library's Digital Collection Manager – already managing still images (pictures, maps, printed music, manuscripts) and text items (finding aids for manuscript collections) – was enhanced to meet the special management and delivery requirements of our sound collection material. The DCM information model is based on an object-oriented approach centred around works, copies, jobs, processes and projects. In practice, very little needed to be done to extend this data model to support the management of audio material. A new category was added to enable separate processes to be invoked based on material type (still image, audio and text). In addition, it was found necessary to allow a copy to have pieces or be a piece of another copy (see circled areas of Figure 1 for areas that had to be amended)

Figure 1: Digital Collections Manager information Model



This was needed to deal with the fact that a single recording may have been copied over time to different carriers. A record of each piece and its technical characteristics had been maintained in an Access Database – the Stocktake Database. Much more than a ‘stocktake’ aid, this database recorded essential descriptive, administrative and technical metadata about each sound recording. As part of the project, the data from the Stocktake database was

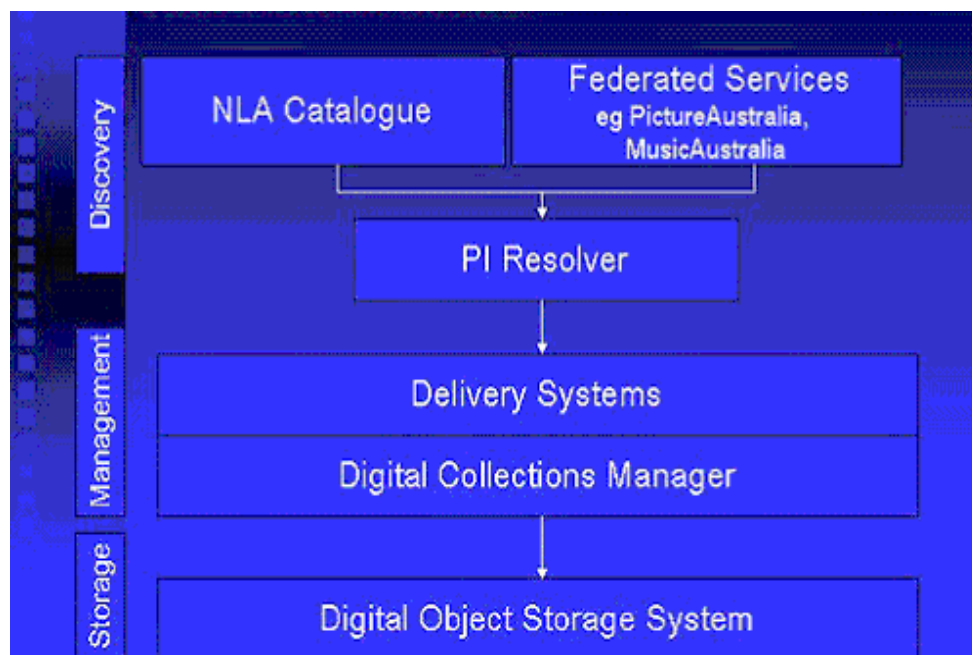
migrated to DCM. This meant that, for the first time, DCM was being asked to manage not just the products of digitisation, but also the entire analogue sound collection.

The real work lay in creating an audio upload utility and upload workflows that could interoperate with the Quadriga System. The BWF standard supported by Quadriga requires a significant amount of preservation metadata to be incorporated in the file header, including the full provenance of each file, which needed to be derived from the data in DCM. An extension was also required to the Library's persistent identifier scheme to accommodate analogue and digital copies as well as transcripts and summaries and the files required for streaming delivery. In addition, DCM needed to be enhanced to manage summaries and transcripts as well as the sound files themselves.

Delivering the collection

Alongside the enhancements to the Digital Collection Manager, the project team developed the online delivery system for this material. The existing delivery systems draw on structural and descriptive metadata in a mirror copy of DCM to deliver the Library's digital objects. Their place in the Library's digital services architecture is shown in simplified form below:

Figure 2 Access Architecture



Persistent identifiers are used to link from resource descriptions in discovery services such as the Library's catalogue, Libraries Australia and other federated services to a contextual display for the digital object. The Persistent Identifier Resolver Service directs the requester to the appropriate delivery system.

The audio delivery system brought new challenges (described below), and the development of the system was also used as an opportunity to go back and think about a more 'generic' solution for our other collection formats. We had begun to realise from our lengthening enhancement register that collections other than manuscripts need finding aids, items other than printed music and books may have multiple parts, pictures as well as maps may require

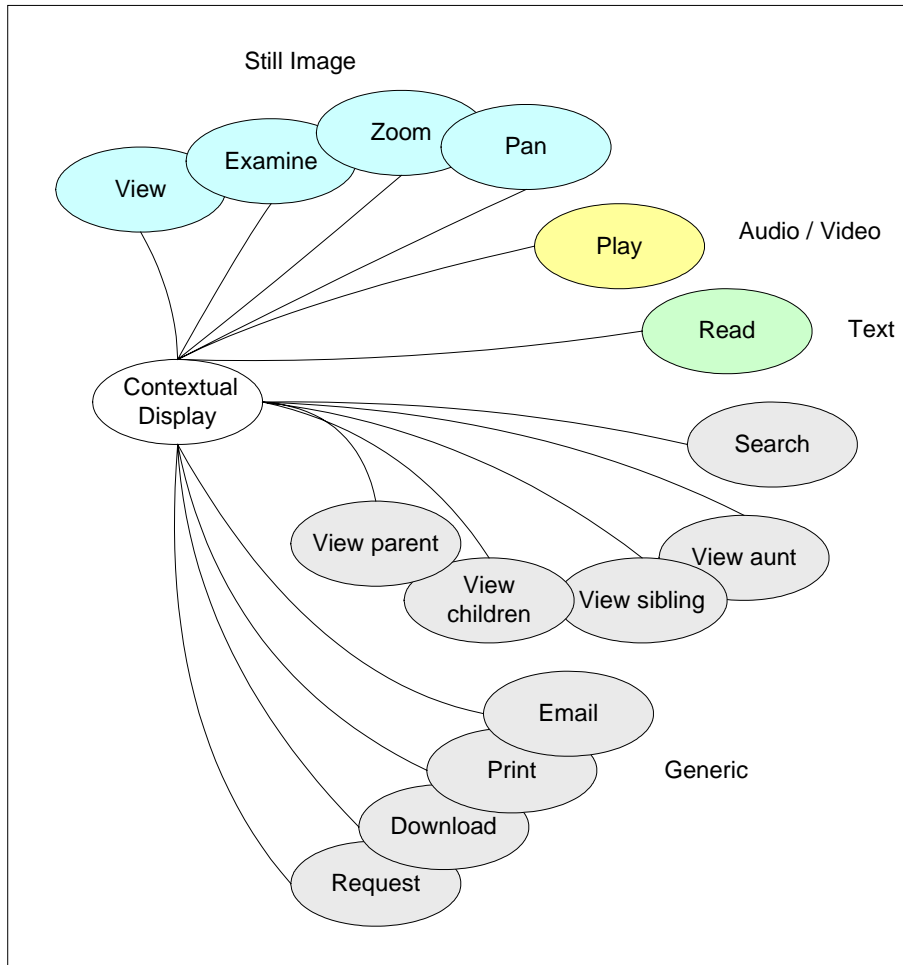
zoom and pan functionality and all formats need consistent brief and full record displays, navigation, print, download and order functions. Our aim was to make the Audio Delivery System as generic as possible while being responsive to the special delivery needs of sound.

Delivery system development for audio was certainly challenging. Major issues included:

- The complexity of the material itself and discrepancies in granularity between how it was originally recorded, how it has been described in the catalogue, how it was registered in the Stocktake Database and how it has been digitised. Examples include: a single interview split across multiple recordings; multiple interviews on a single recording; and the need to deliver fragments such as individual songs performed during a folklore interview.
- Developing a user interface that supports the delivery of both sound and text for the same item. Although we are making no attempt to synchronise sound and text (that is a whole other problem), the user interface needed to be designed to support both listening and reading workflows. We also decided to treat an item as deliverable if it had a summary or a transcript or a sound recording, with the need for appropriate messages and ordering options when the item itself had not yet been digitised or there was no online transcript.
- The decision to design and implement additional print and search functionality. This includes the ability to print individual pages, parts of transcripts or entire transcripts, and the ability to search within the text of individual transcripts, across the transcript corpus and across the Library's collections (the latter will be implemented at a later date).
- The lack of any international models for simultaneous sound and transcript delivery, or of a suitable XML schema for encoding oral histories.

The Library utilises Use Case modelling² to develop requirements specifications. As we learnt more about how complex our content and its requirements actually were, we had to develop ever more use cases. Our specifications modelling was an iterative process over several months, requiring careful management of document versions.

Figure 3 Use cases



We took a similarly systematic view to deciding on an encoding standard for our transcripts and summaries. Converting these documents and delivering as HTML would have allowed our users to read the transcript as they listened online. It would not, however, support the more advanced services we wished to offer. These include the ability to allow users to simultaneously jump to chosen time-points in related sound recordings and transcripts, the ability to deliver ‘fragments’ such as individual songs in an oral history recording, and the ability to encode names, titles and subjects in transcripts to support advanced searching. We also wanted to ensure that we used an encoding standard that would maximise opportunities to contribute oral history and folklore content to federated services relying on shared standards.

We knew at the outset that TEI³ offered the best set of possible encodings, but also knew that we did not need the full range of TEI options. TEI has a base tag set for transcriptions of spoken language, but the purpose of this tag set is to support linguistic or acoustic research and it is too specialised for oral history purposes. After analysis of our own content, we used the TEI pizza chef application⁴ to put together a combination of TEI Lite plus some of the available ‘toppings’ suitable for oral history material. Even so, we needed to create our own profile to accommodate additional needs, such as the need to encode the status of a transcript (draft or complete), or to generate appropriate cultural messages, for example to indicate that the names of deceased indigenous people might be included in a transcript.

Preparing content for delivery

While working on delivery system development, the project team also analysed requirements for converting and delivering complex oral history and folklore content which can consist of one or more sound files, summary and transcript. A pilot – consisting of a small number of oral histories and folklore recordings – helped us to test the technical, workflow and resourcing needs required to deliver audio content on the web. The pilot also helped us to assess what we needed to do to create copies of audio recordings suitable for online delivery (streaming derivatives and/or downloadable files), to convert existing electronic and hard-copy finding aids to a format suitable for online delivery (XML delivered as searchable HTML text), and to deliver these digital objects through our delivery system.

These standards and guidelines are now being applied to newly digitised material and to new acquisitions and this will slowly create a body of deliverable audio content over time. The real challenge, however, will be to make a significant proportion of the freely accessible material available online in the immediate future. With the needs understood, we specified the resources that would be required to do this. As at December 2005, the Library is still considering options for resourcing the project. A strong driver for finding this funding is that a significant proportion of the Library's sound collections has already been digitised for preservation purposes and that this long term preservation investment can now be leveraged to provide online access.

Metadata alignment

We were very fortunate that this collection content was already well described. Every oral history and folklore item has been catalogued at item level in our local catalogue. Our collection is therefore discoverable through our own OPAC, Libraries Australia and other discovery services such as MusicAustralia. The sound component of the collection was also described in the Stocktake database and this data has been migrated to DCM.

Our digital delivery systems all rely on the import of a catalogue record to provide descriptive metadata. It has been a point of faith for the Library that the catalogue be the first point of entry to the collections and the catalogue record to be the master resource description for its digitised collection items⁵. The Library's catalogue currently includes a catalogue record for each bibliographic item in the oral history and folklore collection. The Digital Collections Manager was originally populated with brief metadata from the Oral History Stocktake database, with a separate DCM record for each piece (that is, each tape, cassette, CD or digital file). A single bibliographic item can be associated with one or many pieces. A single piece can be associated with one or many bibliographic records. There is therefore a need to align these record sources, ensuring that the delivery system can correctly deliver all files associated with a single bibliographic item and vice versa.

A simple overlay of catalogue records over the descriptive metadata in the existing DCM records will unfortunately not be possible. This is because only a minority of records have a one-to-one relationship; and the catalogue and DCM records both contain data which must be preserved (e.g. the catalogue records provide richer description, but DCM records include coded play time data suitable for inclusion in MARC 306).

We therefore need to:

- Inspect each catalogue record and associated DCM record/s for each bibliographic item to determine which data from the DCM record should be added to the catalogue record before uploading to DCM.
- Create parent/child relationships between DCM records associated with a single bibliographic item.
- Update DCM with sub-unit details (e.g. Session 1, Session 2) for complex items associated with more than one sound file.
- Use the standardised forms of access conditions in the catalogue records to set access conditions in DCM.
- Add URLs to Voyager records for items on open access.
- Seed these URLs to Google and other search engines to ensure that our users find our collection regardless of whether they start their search from one of our specialist services.⁶
- Make the resource descriptions accessible through the Library's OAI Gateway for use in other discovery services.

With more than 9000 bibliographic items and 18 000 DCM records, we obviously hope to develop some automated approaches to this work, but will commence on a record by record basis to establish patterns and understand the task more fully.

Creating streaming derivative and downloadable files

Unlike some other sound archives – which consist primarily of song length or sound bite length files – our sound files are very large. Individual recordings are typically of 60 minutes duration, and a single oral history interview can be many hours in length. With such large files, download over even a fast internet connection is lengthy, and impractical for users without broadband access. The Library's aim is to make it easy for Australians, anywhere, to access the content that they want: we therefore decided that we needed to stream our sound files, with an additional download option with appropriate warnings about file size.

This decision means that we rely on proprietary streaming formats, and that our users therefore need to install proprietary players in order to listen to our content. From our point of view, this is an undesirable but necessary position – at least for the time being. We undertook extensive research⁷ on streaming file formats, encoders and players, and decided to offer our users the option of RealAudio and QuickTime streaming. This offers us some redundancy in the event of one of these proprietary solutions disappearing from the online landscape. There are, of course, risks in this approach, and some of these became apparent during our pilot. For example, we had to use the QuickTime plug-in rather than the QuickTime native reader to support time-pointed delivery in earlier operating systems. This adds to the complexity of the application and reduces usability for QuickTime users. Nevertheless, this approach – of offering more than one proprietary solution – is the approach taken by a number of major archives, including the ABC and BBC, and appears to be the best option for the time being.

To create QuickTime and RealAudio streaming derivative files, we must pass our master sound files through the relevant encoders. Both encoders require the existence of a 16-bit derivative master from which the streaming files can be created. 'Born digital' collection items and files created by digitising analogue tape are created as 24-bit files. To create

streaming derivatives from these files, we need to create a 16-bit derivative master from the 24-bit preservation master. This has raised questions about when streaming derivatives should be made and at what point we should do this for digital content uploaded to the DOSS before our streaming encoders were commissioned in mid 2005. Hitherto, preservation requirements have driven the digitisation plan, with only a small number of items digitised 'out of order' in response to orders or specific requirements for Library web publications. This has been sound practice, given the compelling need to save our collections to stable preservation formats.

However, the advent of the delivery system inevitably adds access imperatives into the digitisation priority arena. It is almost inevitable that offering online access to portions of our sound collection will raise new expectations among our users and stakeholders. We will need to carefully balance these needs and clearly articulate what our priorities are as we make parts of our collection available online.

There is also the question of whether we should be making downloadable versions of audio digital content that is eligible for online delivery. We have not yet set up an encoder for the generation of MP3 files but the emergence of podcasting as a mainstream technology may make us review the priority for doing this. Podcasting itself is not necessarily a service that the Library would consider for the delivery of its digital content. Our audio files are very large and it is most likely that users will prefer to be alerted of the existence of relevant material through resource descriptions before choosing to download the file. However, a new generation of users will have MP3 players and may wish to listen to this material on their own terms rather than through streaming technologies.

Converting summaries and transcripts

The Oral History collection includes approximately 9000 items, of which approximately 4000 have associated timed summaries and/or transcripts. Of these, more than 2000 transcripts are available in hard copy only. The remaining transcripts – like the timed summaries – are available as Word documents in a variety of formats and layouts. This content needs to be marked up in TEI Lite for delivery through the system.

Converting existing hard copy and electronic summaries and transcripts to the Library's TEI dtd is a major task, requiring extensive data analysis, XSLT programming resources, and significant post conversion quality control and editing, in addition to strong project management skills. As any librarians faced with legacy data will know, close analysis of our transcripts and summaries revealed that while a fair degree of automation can be used, variation in document formats means that a substantial human task will still be required. This is, inevitably, a costly part of the project.

To achieve mark-up of our summaries and transcripts, we will need to:

- Identify, retrieve and get a firmer estimate of the number of hard copy only transcript pages. There are an estimated 50 000 pages of such content.
- Contract with a commercial provider to scan and OCR (or double key) hard copy content.
- Undertake some pre-processing of our existing Word files.
- Convert all documents to TEI using XSLT.
- Load and process all files on completion, including extensive quality assurance work.

A considerable amount of work has already been done in this area, including specification of the TEI Lite set to be used, hand-encoding for testing purposes, and a thorough evaluation and estimation of the potential to automate conversion of summary and transcript data. As is usual in such conversion projects, the variety of content and format of the transcripts will require careful conversion specification and post-processing quality assurance work.

Of course, a major outcome of this project must be development of new workflows and processes, to ensure that all new oral history and folklore content available for public access can be delivered online immediately after collection, and with a minimum of Library staff time. This project will therefore specify new templates for interviewers creating time pointed summaries, and a new statement of requirements against which commercial transcription agencies can be tested. Staff will require skills and training in creating XML documents to oversee and manage these new requirements.

Rights management

The Library has collected oral histories over a fifty year period, during which there have been a number of access forms and different wording of conditions. It remains the case that some interviewees' intentions are unclear, especially as many interviewees signed forms with no concept that online delivery might one day be possible. However, the Oral History section has worked hard over the last two years to clarify and interpret the various historical access conditions and to record these in standard formats in catalogue records. Access forms for all new oral histories determine public use of the material as "disclosure in any way, by any media". A policy decision that historic recordings with no restrictions on use should be deemed suitable for delivery over the internet has also been made. As catalogue records are uploaded to DCM, staff will switch delivery access to 'available' if this is appropriate.

In the first instance, the Library's delivery system will only deliver content available for public access. Many oral history recordings are restricted for designated periods of time; users must gain explicit permission from the interviewees to use the oral history recording or transcripts. In the near future, we hope to enhance the DCM's rights management capacity, and to enhance the delivery system so that users who have gained such permissions can use the collection material online, via secure authenticated access.

Conclusion

As our librarian colleagues will know, there is much truth in the adage that if we knew how much work would be involved in a project at the outset, we might never embark on it at all. We knew that this project would be a big, hairy and audacious one, but it turned out to be even bigger and more complex and more resource-intensive than we had anticipated. This has been a major investment for the Library, at a time when there have been many other calls on the Library's development resources – not least the redevelopment of Kinetica into Libraries Australia.

Nevertheless, the benefits of these projects will be enormous. The Library's web audio delivery system resulting from this work is highly innovative. We know that our digital collection items are used many more times than those same items in their original formats. We know that all sorts of people – including those who would never have come to the National Library in Canberra – are now using those objects for their own purposes. At this stage in our project, we can only imagine the new audiences for this wonderful collection,

and the impacts that providing this online access will have for those audiences. We look forward to seeing these wonderful stories informing Australians' understanding of ourselves and our culture as they are made available via the Internet to anyone, anywhere.

Endnotes

¹ Kevin Bradley, the Library's Manager of Digital and Audio Resources Preservation, is the President of the Australasian Sound Recordings Association, and the Vice-Chair of the Technical Committee of the International Association of Sound and Audio Visual Archives. Kevin is currently seconded to the Australian Partnership for Sustainable Repositories project as Sustainability Officer.

² Use Case modelling is a tool used to model required user tasks for a system. A single Use Case typically models basic and alternative courses or workflows to achieve these tasks. Many documents on Use Case modelling are available on the Internet.

³ Text Encoding Initiative 2005, <<http://www.tei-c.org/>>. TEI was originally developed to support sophisticated markup of historical texts.

⁴ Text Encoding Initiative 2005, <<http://www.tei-c.org/pizza.html>>. This application allows users to select a TEI Lite 'base' and to add 'toppings' according to local business needs.

⁵ The Library's commitment to the catalogue is explained in Pearce, J and Berko, M 2004, 'The hybrid library revisited', a paper presented at the VALA 2004 conference, viewed 15 December 2005, <<http://www.nla.gov.au/staffpaper/2004/pearce1.html>>.

⁶ See Boston, T 2005, 'Exposing the deep web to increase access to library collections', a paper presented at the Ausweb 2005 conference, viewed 15 December 2005, <<http://www.nla.gov.au/staffpaper/2005/boston2.html>>, for a description of this process.

⁷ The Library is happy to provide its research to other institutions on request. Please contact the authors.