5-2015

Digitising and Accommodating Archival Collections in a Museum Collection Management System: The ANZAC Connections Experience

Theresa Cronk
Research Centre at the Australian War Memorial

Digitising and Accommodating Archival Collections in a Museum Collection Management System: The ANZAC Connections Experience

Description

[Excerpt] The Australian War Memorial (AWM) is an archive, library, museum and memorial located in Canberra, Australia. Its diverse collections are primarily managed in the museum collection management system, *Mimsy XG*. The Research Centre is one section within the Memorial and is tasked with digitising, managing and providing access to both archival and library collections.

Anzac Connections is the Australian War Memorial's major digitisation and web development project to mark the centenary of the First World War. Involving the delivery of digitised collections of ephemera, letters and diaries from the Memorial's Research Centre, this project marked the dawn of a new phase of digitisation. This project also encompassed improvements to the accessibility of collections on the Memorial's website.

This article will discuss the challenges and solutions met by the Memorial in digitising and providing public access to archival collections via a museum collection database. It will focus on the issues experienced at the outset of the *Anzac Connections* project with digitisation workflows, multiple collection databases and the online display of archival collections. It will also cover the solutions that were adopted to address each of these issues and conclude with the project outcomes that will be beneficial for any future digitisation projects undertaken.

Keywords
digital archives, library, Australian War Memorial, collection management

Digitising and Accommodating Archival Collections in a Museum Collection Management System: The ANZAC Connections Experience

Theresa Cronk

Research Centre at the Australian War Memorial

The Australian War Memorial (AWM) is an archive, library, museum and memorial located in Canberra, Australia. Its diverse collections are primarily managed in the museum collection management system, *Mimsy XG*. The Research Centre is one section within the Memorial and is tasked with digitising, managing and providing access to both archival and library collections.

Anzac Connections is the Australian War Memorial's major digitisation and web development project to mark the centenary of the First World War. Involving the delivery of digitised collections of ephemera, letters and diaries from the Memorial's Research Centre, this project marked the dawn of a new phase of digitisation. This project also encompassed improvements to the accessibility of collections on the Memorial's website.

This article will discuss the challenges and solutions met by the Memorial in digitising and providing public access to archival collections via a museum collection database. It will focus on the issues experienced at the outset of the *Anzac Connections* project with digitisation workflows, multiple collection databases and the online display of archival collections. It will also cover the solutions that were adopted to address each of these issues and conclude with the project outcomes that will be beneficial for any future digitisation projects undertaken.

The issues and challenges

When *Anzac Connections* was first scoped, the Research Centre had been digitising official government archival records for almost 10 years and had accumulated 1.5 million pages in the form of digital files. These were primarily archival files that were organised and listed on a collection management database at item level, making them well suited to digitisation. These files, with the classification of Official Records, are controlled on the National Archives of Australia's in-house collection management system, *RecordSearch*.

1. Collection management systems

The Memorial used two collection management systems when *Anzac Connections* commenced and there were reasons supporting the maintenance of both these systems. These two databases, Research Centre Database and Mimsy XG, are described in detail below.

The use of these two collection databases limited the ability to unify the deployment of data to the web and aggregate the data in more appropriate ways. It also hampered the ability to link people records to collections in the one system and present that information in one place on the web.

The Research Centre Database

The *Research Centre Database* (RCDB) was an Access database that had been developed in-house by the Memorial. It performed two major functions. The first was to allow the images associated with the archival records described on *RecordSearch* to be managed within the archival hierarchies of series, sub-series, file and item level descriptions. The second function was to manage the data contained in the 645 486 biographical roll records that were created from a proportion of the digitised images and to maintain the links between the two.

1.1 Interface for RCDB Index Editor showing the series that had been digitised and their publication status online

1.2 Interface for an excerpt from the First World War Embarkation Roll in RCDB Biographicals

The RCDB was supportive of translating archival hierarchies to display in a logical way on the Memorial's website. The inherent flexibility in how this database could be manipulated for web display contributed to the logical and orderly fashion in which researchers could navigate through online collections. Digitisation procedures revolving around this database were also well established.

The RCDB, however, was not suitable for managing large quantities of small collections. This database was designed for individual record series, each containing a large number of files. It was not intended to manage the hundreds of individual archival collections that were expected to be captured as part of *Anzac Connections*. These

collections also already had a database presence in another Memorial system and replicating these solely for digitisation was not the solution.

The absence of standardisation for processes in the RCDB was also an issue. Later projects, much more than some of the earlier ones, were reliant on developing archival hierarchies in the database based upon the archival control symbols of files. The use of titles and descriptions were largely based upon the online display requirements of each individual record series.

The biographical rolls component of the RCDB was also affected by the absence of standardisation. There was no one-to-many relationship between each instance of a person's name, with each occurrence entered separately. Knowledge of each biographical roll was required to understand what the data in each field was and how it was to display, if at all, online.

The bulk of the biographical information in the RCDB was historical data gleaned from archival records themselves. This resulted in lots of abbreviations, variations and references to entities that reflected the way in which these were known at a particular point in time.

Mimsy XG

Mimsy XG also had the ability to create archival hierarchies. However, it could only cope with two levels – collection and item – and the inherent relationships of archival hierarchies were not visible when collections from this database were published online. Mimsy XG also contains a people authority that had the ability to link to objects in the collection that were strongly related to an individual service person. Finally, Mimsy XG was also heavily reliant on a subject thesaurus for a variety of terms including units and places.

2. Collection records and digitisation processes

The archival collections to be digitally captured by *Anzac Connections* were only described at the collection level on Mimsy XG. There were few file level or item level records that could easily be adopted for the purpose of attaching digital images as had previously occurred with Official Records. Individual pages/items within these collections had often been digitised on demand or request, independent of arranging and describing the entire collection. This was also true of the ephemera collections that were to be included in the project.

Additionally, collections were continually being used by researchers. This meant that the order of items or pages within both these archival and ephemera collections may have been disturbed. As only a very basic understanding was held as to what was within each of these collections, a lot of work was required to prepare individual collections for scanning. Fundamental to this was the realisation that digital images have to be managed differently to the approach applied for physical collection items.

3. Online display

Online access requirements dictate that a file should be easy to download by researchers. This was one of the reasons why Research Centre items were always displayed as bundled pdfs, if they had multiple pages. The treatment of items with less than 4 pages differed between RCDB and Mimsy XG. From the latter database, images in this category were displayed as low resolution jpegs. This meant that researchers were not able to zoom in and were limited as to the ability to read textual documents. As a result of some technical issues with Adobe, which was used to display the online images, some of these access pdfs also displayed in black and white rather than full colour.

Consistency of web display for all digital images was another issue, as it depended on the collection database being used. Researchers had to navigate to three different areas of the Memorial's website in order to find and search collections. The challenge was to unite all digital content in a way that would allow researchers to go on

a journey of discovery to find more information and content related to their field of enquiry from wherever they landed on a collection related page.

Addressing the issues and overcoming the challenges

1. Digitised Collections as a collection classification

New collection classifications can be added to Mimsy XG, and it is these classifications that are ultimately used to determine how particular types of records display on the web. For example, the classification of Published Collections may choose to display a collection number and an accession number, whereas other areas may choose not to do so. This ability grants flexibility in how the database can be used to meet the needs of a diverse and wide-ranging collection.

Although the Digitised Collections team had existed in the Research Centre for almost 10 years, it was project focused and not widely considered a curatorial area. This changed in August 2012, when the concept of Digitised Collections as a curatorial area was introduced, and a new classification appeared in Mimsy XG soon afterwards.

This decision came as a result of the discovery that several external institutions were making distinctions between physical collections and the associated digital versions. In a document outlining file naming conventions, The University of Colorado Boulder mentions that collection documentation should record the "physical collection from which the digital collection was derived." [i] The New York Public Library, Library of Congress and Yale University Manuscripts and Archives were also observed allocating a digital identification number to their digital collections. [ii] The National Library of New Zealand incorporates a digital identification number assigned to their digital collection items into the associated url link. [iii]

With this approach in mind, the concept of a "Digitised Collection" developed and was defined as:

items which are the digital surrogates of Research Centre material created for preservation and access. Items with this classification are managed by the Digitised Collection curatorial team with the original item (or physical item) still being managed by the relevant curatorial area

This means original items are not managed by the Digitised Collections team but if a surrogate is needed for some reason, such as format change from paper to digital, responsibility is taken for the surrogate created.

2. Developing a conceptual model

Developing a visual representation of the concept of physical and digital versions of archival and ephemera items as records in their own right was the next task. The "hierarchical tree" was enthusiastically endorsed and adopted to model how this concept would work in Mimsy XG.
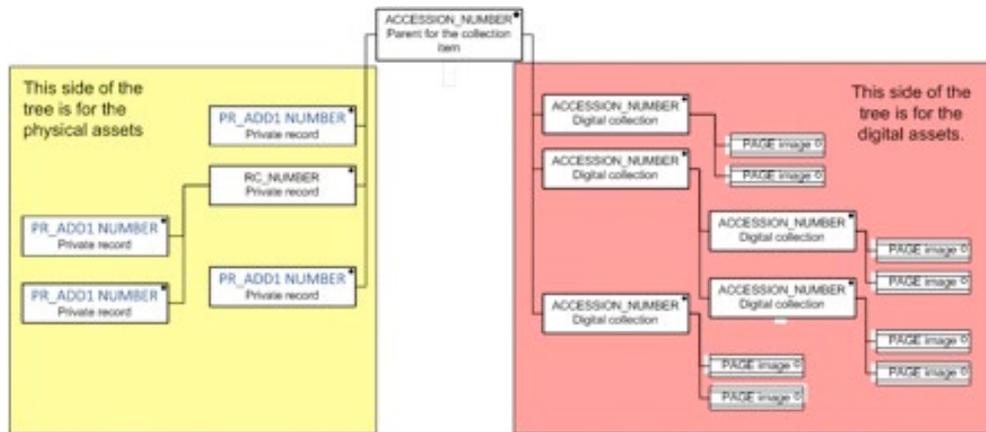
1.3 The hierarchical tree concept distinguishing records of physical and digital collections

At the top of this tree, there is an accession number for the parent collection level record in Mimsy XG.

The left hand side of the tree is where the curatorial areas manage new acquisitions and items going on display or loan. These records are rarely released to the web. The right hand side of the tree is where digital assets are managed. These are

created if an item needs to be scanned because it is going on loan or exhibition, featuring in a publication, or is just part of a digitisation project.

The Research Centre now has a flexible system for digitising collections on demand. It is possible to insert new records into the hierarchical structure at every level and assign an appropriate relationship with other existing records in that structure. It is not necessary to arrange and describe the entire collection to create this structure. Previously different systems of numbers existed to achieve the same result.

3. New accession numbers prefixed with RCDIG

To be able to achieve the scalable system described above, the concept of prefixing the accession numbers for all Digitised Collection records with RCDIG was developed. There were a number of reasons for this.

Firstly, the Memorial has a history of using prefixes for accession numbers. This also assists with keeping numbers unique as it is known that entirely numerical accession numbers were previously used sometimes.

Secondly, the method of sub-dividing collections and the associated meaning of the accession numbers used during this process have changed over time.

Thirdly, the instillation of meaning into accession numbers had resulted in some very long numbers during an early digitisation project. To indicate Item 14 within Series 2, Sub-series 2, File 1 of the First World War concert and theatre programs collection [Published Collections Souvenirs 2], the accession number PUBS002/002/002/001/014 was applied. These numbers were discontinued due to the difficulties associated with their use for citations, reading and also remembering.

In comparison, numbers prefixed with RCDIG are much shorter and simply indicate that an item/record is part of the Research Centre digitised collection. That is all this number is required to do. An example is RCDIG0000139. These records also

incorporate a control symbol or collection number to identify the physical collection from which the digital version originated.

4. Translating these ideas into the collection management database

The hierarchical tree previously mentioned was the concept used to translate all of the above ideas for digitised collections into the collection database. The functionality required to do this was inherent in the 'whole/part' field. It is here that catalogue records can be linked together and their position in the hierarchy described using a set of pre-defined archival terms, eg. collection, series, sub-series or file.
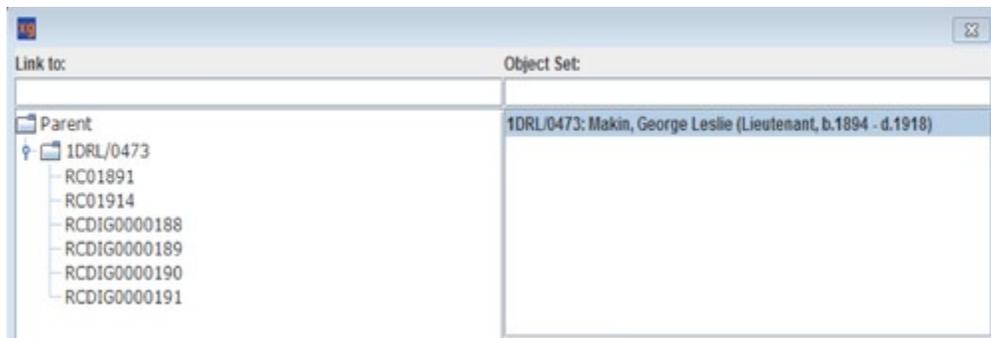
1.4 The whole/part field of Mimsy XG where 1DRL/0473 is the parent and the RC/RCDIG numbers are children of that collection

It was the new field called 'Broader Text' in Mimsy XG that provided the database answer. Drawing on the links that had been made in the whole/part field, the data was now displayed in the broader text field in a hierarchical way. Collection records, and any subsequent level records, are now displayed as a hierarchy that can be clicked down through to arrive at the images attached to file level records.
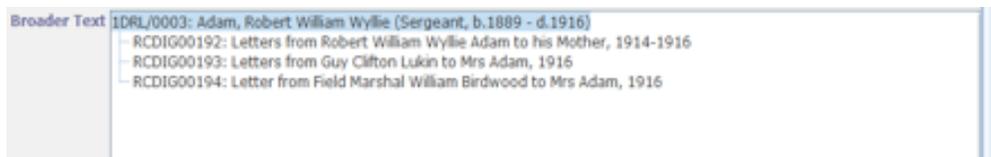
1.5 The broader text field in Mimsy XG showing the relationship between the parent collection and children

To get this data to display in the same manner on the web, it was necessary to find a way that the records could sort themselves into the hierarchy. This was only a small problem that was overcome with the adoption of the 'Hierarchy Level Sort Order' field. The web display now relies on the numerical value entered in this field to indicate where a record appears in the hierarchy that is displayed on every catalogue record.

Using 'Related Objects', links can be made between versions of a physical item. These are recorded in the Linked Objects flexi-field and displayed in the link ledger. This also allows an explicit distinction to be made between the physical or original item record and its associated digital derivative. The 'physical descriptors' field was also developed to allow the number of digital images associated with each catalogue record to be captured.

1.6 and 1.7 The Related Objects field in Mimsy XG and how it is used to distinguish between the physical and digital records

5. The migration and cessation of RCDB

The development work in Mimsy XG for *Anzac Connections* revealed that there was strong potential for all archival records and images to be managed in one system. This led to the successful migration and consequent demise of the RCDB.

The migration of the component of RCDB that managed images was relatively easy due to the pre-existing archival hierarchies. However the flexibility of RCDB in this

regard, as outlined above, also presented challenges. These were reinforced through the exercise of mapping fields in RCDB across to Mimsy XG. There was a lot of data clean-up required where data was missing in the hierarchies and, in some cases, the re-mapping of hierarchies using control symbols was required before an individual series could be migrated.

It was the migration of the biographical roll records that proved to be the most difficult, as this was not akin to any other data set that had been managed in Mimsy XG previously. This data followed strict procedures and associated permissions for updates. Upon discovery of a module within Mimsy XG that was not used, work commenced to customise it so that it could meet the unique requirements of the biographical rolls.

Documentation became crucial to understanding the data and the way that it had been stored in the RCDB before it could be migrated. Mapping a particular field in the RCDB to a corresponding field in Mimsy XG was not as easy as it sounds. Even if data was in the same field of the RCDB it could mean something different for every roll. This meant that the data had to be unpacked and understood as to its original purpose for each of the data sets, and only then mapped to a field in Mimsy XG. Rules also had to be established around these in order to determine how and what data was to display on the web.

There was a tight turnaround for this to occur and it was not feasible to create an authoritative person record for the names that existed in the RCDB. Instead, a new person record was created during the migration process for each and every record that existed in the RCDB biographical roll component, even though many of these were multiple instances of the same name. This resulted in an additional 645 486 people records in Mimsy XG, on top of those which already existed. The data clean-up required to create an authoritative people record and develop a one-to-many relationship between collection items and roll records is now an ongoing process.

There was also the problem of moving free-text historical data into a database where it had to conform to an established thesaurus. A short-term solution so that the

migration could proceed was to create new directories in the thesaurus for the data impacted upon in this regard that was appended with the label 'RCDB – DO NOT USE'. Since then, the challenge has been to find a solution to merge these terms into an authoritative list of terms that both provide the historical context but also standardise the terms for searching. Places were particularly problematic as town boundaries change, and towns – and even countries – change their names.

The solution adopted was to enter place names multiple times in the thesaurus, using scope notes to distinguish between them. Each place was entered into the thesaurus as a hierarchy and each level in the hierarchy was assigned a specific type. For example, the hierarchy Sydney, New South Wales, Australia were given the types city, state and country respectively. The web rules defined which of these terms displayed on the web so that the data mirrored the historical record.

The solution adopted was to enter place names multiple times in the thesaurus and using scope notes to distinguish between them. There were also difficulties in the web display of places as strings containing terms that were not on the original record. To overcome this, each level in the places hierarchy was given a name and web rules were written around this that determines the terms which are required for web display.

Bringing it all together: The outcomes of *ANZAC Connections*

1. Improved digitisation processes

The digitisation workflow model below shows nine different stages in the digitisation process and is the model that was developed and successfully implemented for the *Anzac Connections* project.

*1.8 Workflow model adopted for digitisation projects*

This model is based on the concept of "batches" – a group of archival collections that collectively totals between 5000 and 6000 images. Batches were a workable solution to the issue of streamlining digitisation processes as it was realised that the same amount of work goes into one collection as goes into many. Pooling resources and working through several collections at once was an efficiency that proved to be feasible. We also adopted "projects" which are defined as a single collection that produces more than 10 000 images. An example of a project as defined in this way is 3DRL/2316 Papers of Sir John Monash KCMG KCB.

One of the most important stages is the preparation stage, which is when the metadata that will become crucial in managing the images and creating database records is captured. This approach can be resource intensive to begin with but the information it produces is later recycled in multiple ways.

At the beginning of each batch/project, a box list is captured detailing the arrangement of every included collection immediately prior to commencement of scanning. This also includes basic titles, dates, authors and descriptions of items for later identification.

This information becomes the basis for the Digitisation Checklist. The Digitisation Checklist is colour coded to track each step of the digitisation stage and is also where any re-arrangement of collections into archival levels of description is documented. This

information, together with an itemised listing of the contents of every file, is later used during the creation of database records.
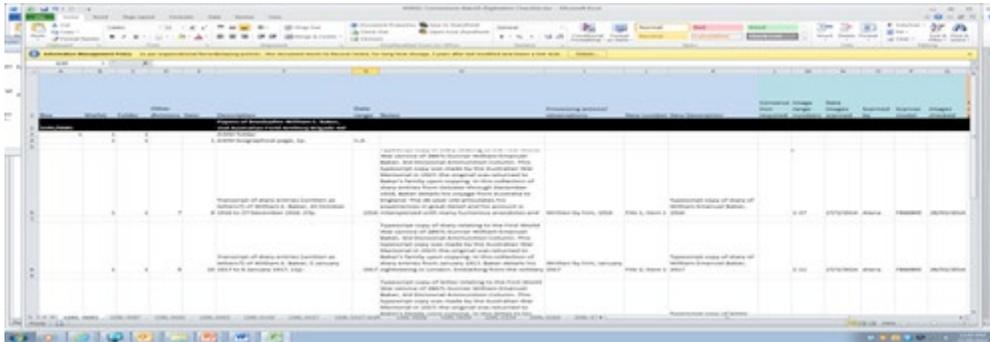
1.9 The Digitisation Checklist – used to manage all activities for digitisation projects

The importance of the Box List and Digitisation Checklist cannot be overemphasised. As the old and new arrangements are kept side by side, we are able to see the old and new positions of items within collections for future work on these collections. This is very important as we do not physically re-arrange the collection, but rather are concerned with how it is presented digitally. Scanning according to the box list does not cause issues when images are processed into new arrangements.

2. Improvements to search functionality across collections

The focus of *Anzac Connections* was always intended to be people and their stories. The aim of the project was to allow people to read a biography, view all of the biographical information relating to a serviceperson's military service and browse through related collection items. This has been achieved with the new People profiles and Biographies page. These pages do not currently exist for all service personnel but more are gradually being added to the website over time.

This above link shows that this has also allowed more faceting and browsing options for researchers. Someone who is interested in a particular type of person can now navigate directly to a listing of particular groups of people using the "Filter by"

option, eg. people who received the George Cross. It is also possible to filter names by related conflict or place of birth.

Subject terms, units, places and people mentioned in these diaries and letters are also now provided on the catalogue records. These provide links to an automated search for other collections on particular topics. For example, if you were interested in finding records related to other people who were in the 8th Australian Infantry Battalion, you can click on this term under unit on the right hand side and be taken to the search results for other related records and books.

3. Improvements to the web display of archival collections

Fundamental to this new display of digitised images was implementing archival hierarchies to describe the relationships between records in a way that would make sense to researchers. Archival hierarchies rely on grouping records together to maintain information about how they were created, stored and used. This concept had to be maintained in the digital environment as well as providing a seamless method of allowing researchers to quickly click through to the records that they were interested in. This was achieved using a combination of the labels "Record type", "Navigation" and "Collection number" on all item/file level catalogue records.

The standard previously used to display images captured from archival files on the AWM's website was as a static pdf. Whilst the option to download a bundled pdf is still provided, images are now also available as individual full colour, high resolution jpegs. These are displayed one page at a time in an image viewer at the top of the page. An example of this new functionality can be viewed here.

In terms of catalogue records, we have created hundreds of file or item level records to attach these images to for web display. This is in addition to the collection level records that already existed to broadly describe the contents of what could extend from one lone wallet through to one or many more boxes of archival records.

The advantages of item or file level records is that we can now provide more detailed description of the contents of an individual diary or list the entire contents of a file of letters. Although the onus is still on researchers to find, for example, the three paged letter that they are after in a file of 100 other letters, there is now an index available that can assist in this process.

Conclusion

Anzac Connections delivered a sustainable platform that enables collections and data to be published online directly from the Memorial's collection management system, MimsyXG. The online publication of images and data sets in the past required the creation of web pages and search functionality that was unique to each data set or series of archival records. Biographical data sets and digitised collections, following authorisation from curators, now flow automatically to the web and are seamlessly integrated into collection search results. This has improved control and conformity over the online display of collections as well as simplifying the process of releasing collections online.

The requirement to document and understand the data and archival arrangements used in legacy digitisation projects, as part of the migration of RCDB into MimsyXG, was crucial to the establishment of standards, linkages between records and web publishing rules for digitised collections and data sets. These standards, linkages and rules have improved the quality of data and have also increased the number of access points by which researchers can find information on the website. Having the data associated with digital images maintained in one collection management system is a major step towards undertaking the data clean-up required to improve these access points even further.

The ability to create archival hierarchies and links between records was one of the biggest challenges faced at the outset of the project. The concept of approaching

physical and digital collections as separate entities was the first step towards achieving this. This model provided a flexible system that can be used to digitise any item in an archival collection, regardless of the level of processing that has been applied. Not only did this subsequently lead to the migration of an entire database but has also provided the answer to a large backlog of collections that were digitised on demand over several years.

The focus of the *Anzac Connections* project was always intended to provide easier access to people and their stories. The new links between people, collections and biographical data has simplified the process of researching the war service and experiences of men and women during the First World War and are also being continually added to as more data clean-up is completed.

The online delivery of 215 individual archival collections, approximately 57 000 pages, would not have been possible without the introduction of box lists and the new digitisation checklists to track the progress of each collection through the digitisation process. Although these new processes are initially resource intensive, the documentation that they produce has proved to be invaluable in later stages of the project and for any future processing work on the associated physical collection.

The *Anzac Connections* digitisation project, as outlined above, involved approximately five years of research and experimentation into improving digitisation processes and the web display of the Memorial's online collections. The success of this project relied heavily on the collaborative efforts of a team of archivists, curators, database business owners, database administrators and web developers. The workflows, documentation and processes that were established as part of this project have now been incorporated back into the Memorial's digitisation program for archival records.

References

Digital Projects Advisory Group University Libraries, 2008. Guidelines on file naming conventions for digital collections. Retrieved
from: http://ucblibraries.colorado.edu/systems/digitalinitiatives/docs/filenameguidelines.pdf

Selago Design, 2011. Mimsy XG User Guide 1.5. Retrieved
from: http://support.selagodesign.com/Library/MIMSY%20XG/MIMSY%20XG%20Manual%20Version%20I.pdf

Sitts, Maxine K (ed), 2000. Handbook for Digital Projects: a management tool for preservation and access. Northeast Document Conservation Centre. Retrieved
from: https://www.nedcc.org/assets/media/documents/dman.pdf

Wisser, Katherine M (ed), 2007. North Carolina ECHO exploring cultural heritage online: guidelinesfor digitisation. Retrieved
from: http://digital.ncdcr.gov/cdm/ref/collection/p16062coll9/id/1838

About the Author

Theresa Cronk joined the Research Centre at the Australian War Memorial in 2003 and is currently Acting Head of Published and Digitised Collections. She has an Honours degree in History and a Graduate Diploma in Archives and Records Management. Theresa has been involved in all aspects of digitisation since 2006. As part of her work, she has researched and developed solutions for issues surrounding the online display of archival collections and preserving these images in the Memorial's digital asset management system, MediaBin. More recently, Theresa has been working on a cross-institutional research project on the role of music during the First World War using digitised archival records.

Notes:

---

i http://ucblibraries.colorado.edu/systems/digitalinitiatives/docs/filenameguidelines.pdf

ii http://digitalgallery.nypl.org/nypldigital/id?1563432; http://memory.loc.gov/cgi-bin/query/h?ammem/scsmbib:@field(DOCID+@lit(scsm001058)); http://images.library.yale.edu/madid/oneItem.aspx?id=2294179&q=&q1=0299&q2=&qc1=contains&qc2=&qf1=subject1&qf2=&qn=&qo=&qm=&qs=&sid=&qx=1004.2

iii http://mp.natlib.govt.nz/detail/?id=9460&l=en