Anzacs Online:
starting a new major digitisation project

Robyn van Dyk
Senior Curator Published and Digitised Records
Australian War Memorial
robyn.van-dyk@awm.gov.au

Theresa Cronk
Curator Digitised Records
Australian War Memorial
theresa.cronk@awm.gov.au

Abstract
Anzacs Online is the Memorial's ambitious 1915 Centenary project, which will make the Memorial's most important collections available for all Australians online. The project will involve the digitisation of hundreds of collections to be added to the website, creating a comprehensive digital archive on the ANZACS and their deeds, and on the wider Australian experience of war. This represents a new era of digitisation for the Memorial; we will be embarking on digitising collections that involve different processes and web innovation to make them available online.

Introduction

Anzacs Online is the Australian War Memorial's current major digital preservation/access undertaking. It is a project to digitise the Memorial's most significant First World War archival and printed collections and to create a big web presence for those collections to commemorate the 2015 Gallipoli anniversary. The project will have a strong focus on digitally preserving the Memorial's most significant, rare and fragile collections as well as making them more widely available for research. Our selection of records will reflect the ordinary voices of those Australians who served as well as those Australians who were more prominent including commanders, VC recipients and writers.

Anzacs Online aims to not only significantly increase the volume, the range and searchability of digitised collection material on the Memorial's website, but also transform the way that users access and utilise this material. It will provide access online to collections, involving new solutions and platforms for their control, management and display in the digital environment.

The project is also envisaged to produce an engaging, innovative and scalable digital presentation via the web of the existing and new digital collections and their context. Using a new web platform (Drupal) and interface, the Memorial hopes to provide multi-level layers of access to the collection for cultural, educational, research and other purposes related to the Australian experience of war over a century of service.

The focus of this paper will be to outline and discuss the problems and issues faced by the Memorial to deliver a large quantity of private and printed records online. This is a big project with limited resources. With our time frame limited to delivery before 2015 scanning had to commence before we had solved all our technical issues in relation to the project. We are also continuing to select material, curate and develop the project along side scanning and problem solving. Much of the work at present is focused on developing the internal platforms to enable these collections to be managed and displayed online. This paper will focus more on this aspect of the project.

While the scanning process is underway the technical and curatorial issues worked on included:

- Finding solutions within our existing collection management systems and budget for supporting these new digital collections and controlling and displaying archival hierarchies.
- Investigating how the public might access the material via the web.
- Controlling the private records images in the Memorial's new Digital Asset Management System.
- Setting up procedures for selection of material.
- Listing of larger collections in order to be digitised.
- Quality control workflows and procedures.
- More detailed arrangement and description on our collection database and investigating what can be achieved though Encoded Archival Description (EAD), a standard for encoding archival finding aids.

- Copyright clearance and policy development
- Scanning process,
- Training and reporting.

The Memorial's existing digital collections, knowledge and experience

While the Anzacs Online project requires innovation and development to be realised it is also based firmly in the Memorial's existing knowledge and experience. The Memorial has been digitising official records for over a decade and we have established our own business workflows and rules, which we can apply easily to any official record that we add to our digital collections.

Much of what we are producing for Anzacs Online draws upon these established principles and workflows. Anzacs Online is first about preserving records that are irreplaceable. Once the records are digitised and made available online the Memorial no longer issues the original records to the public. Being able to provide appropriate public access via the web is, therefore, always an essential component of the Memorial's digital preservation program.

Our digitisation processes must be completely accurate and our images of high quality. Each individual collection, series or sub-series of records is digitised in its entirety. The digitised records are a faithful reproduction of the original records. The contents have not been altered by the Memorial in the digitisation process. Each image and file is 100 % quality checked one for one against the original.

In December 2010, the Memorial went live with its new Digital Asset Management System (DAMS). The DAMS preserves the images and their technical metadata and ensures that the images are managed and controlled against obsolescence and can easily be accessed and reused.

As a preservation principle, our records are digitised at high resolution for long-term storage and preservation in the DAMS. A lower resolution image is derived from the high-resolution images through an automated process that bundles the images at the file level for online access. Currently the bundled images are also stored in the DAMS, however, we are hoping to change this practice as part of the Anzacs Online process and create bundles deployed from the DAMS 'on the fly' for the web. Images stored in our DAMS are undeletable and the existing bundles are currently creating a multitude of problems especially when a correction has to be made.

For the most part, the digital preservation of the Memorial's paper based records has focused on its large series of official records, such as the unit war diaries and biographical rolls. These collections were chosen because they were heavily used for military and family history research. However, they were also chosen because the Memorial had copyright permission to make them available online.

The bulk of the Memorial's extensive private records collections are works in perpetual copyright where the owner is not able to be located within reasonable

resources. Without permission from the copyright holder to publish online the Memorial was unable, in the past, to digitally preserve these orphan works and display them online. In 2006 a flexible exception to the Copyright Act (Section 200AB) was introduced for libraries and archives for specific non-commercial purposes. The exception has enabled the Memorial to preserve some of our most important private records collections and make them available online.

The Memorial's preservation program has, therefore in the past, been skewed towards what could be made available online under the Copyright Act. For the last 10 years, most of the solutions for access to digitised archival records were developed in the context of official records. The Research Centre designed workflows and the management of digital assets around the official records collection. The Anzacs Online project, with its strong focus on making available significant private and printed records has required new decisions and problem-solving, including how these images will be managed in the DAMS, what collection management system will they be managed in and where and how will these images will be displayed online for public access.

Management of metadata for linking digitised collections

Digitised archival material and books generally need to be displayed as a hierarchy for public access online. Records, for example, need to have their pages displayed consecutively and then grouped as files, class and series. The question of where the metadata related to this linkage between digital objects is saved and managed is a preservation issue.

The Memorial has an in-house created database that stores and manages the relationship information between the images for Official records. The database also deploys the bundled images from the DAMS for their display online. The images are bundled as a PDF (Portable Document Format) at the file level. We are presenting over 2,000,000 paper-based images online using this method. This process has worked well for the Memorial so far, because these individual collections are very large and the number of collections is not great. Members of the public who need to use these collections can be easily directed to our web page where they can be accessed from a list.

One of the first technical problems for Anzacs Online was that the existing in-house created database was not suitable for managing large quantities of digitised small collections. For *Anzacs Online* we will not be digitising one or two big collections but many hundreds of small collections. The established methods of controlling digital collections for internal management as well as for deployment to the web would soon become unwieldy. Ideally the public should be accessing digitised collections through the existing collection management system not a system developed as a work around for a problem. The private records collection is internally managed on a different database to that of the official records collection.

The Memorial uses Mimsy XG (MICA) http://www.awm.gov.au/search/collections/) as its Collection Management System (CMS) for museum related collections - objects, photographs, film and sound. The Private Records collection is also

catalogued into MICA. The Private Records collection consists of over 10,000 records from non-government organisations to the personal papers of individuals from all ranks and services of the Australian armed forces. They include privately donated diaries, letters, notebooks, papers, and cards written during wartime, as well as some reminiscences written after the event. The oldest items date back to the 1860s.

MICA will be the system that the Memorial is developing to deploy private records to the web. We are also planning to make our library management system obsolete and move the printed collection into the MICA database. MICA manages and preserves the metadata requirements of digital objects far more effectively than the Memorial's library system *First*. As part of Anzacs Online, the Memorial has digitised its collection of published unit histories and will display these collections through MICA, not the books database. (http://www.awm.gov.au/collection/books/)

Anzacs Online is above all a digital preservation project and for long term preservation purposes it is important to store and manage metadata about the relationships between digital objects in the collection management systems and in the DAMS.

Research commenced to locate within MICA a space to create levels of arrangement within collections. We needed to look at what fields and modules we had available and were not currently being used and use these fields to describe the relationships between our digital objects. This was a challenge and involved some data clean up as every field we had available was populated with some form of data, often placed incorrectly there. However, by January 2012 we have a clear dataset in the "whole/part field" of our CMS and have inputted and extracted data as a test for online display of hierarchies.

An ephemera collection of over 700 concert and theatre programs of the First World War is being used as the pilot for testing the platforms developed for the project. This collection is small enough to play with but also large enough to test the solutions developed. The collection has a hierarchical arrangement and we have successfully achieved this arrangement within the "whole/part field" in MICA. Using that data we have also prototyped how the records will be displayed on the web (in a test environment). Collection records and subsequent level records are displayed as a hierarchy that can be clicked down though to arrive at a bundled set of images at the file level.

The Memorial has been using PDFs to bundle collections for many years now, however, with the implementation of the DAMS, technical issues with Adobe and requirements for Anzacs Online PDFs have not been serving our purpose. An alternative to the display of access images is close to being selected and implemented. Open internet library book reader is being tested to replace PDFs (http://www.teleread.com/library/new-bookreader-on-the-internet-archive/). This is an open source tool that can support transcription and zoom. It has an attractive page-turning feature.
.

Preparing a medium size collection for digitisation as part of Anzacs Online

The planning and preparation involved for a major digitisation project is longer than most people think. The collection has to be arranged and described and made suitable for the online environment. Personal records are not always presented in a logical or chronological order. The challenge for the Australian War Memorial is to make the digital images of the diaries accessible and usable in the online environment without interfering with the integrity of the original documents.

Case study: Papers of Field Marshal Lord William Birdwood-3DRL/3376

The Birdwood collection occupies two shelf metres and includes diaries, correspondence, drafts of speeches, photographs, scrapbooks, newspaper cuttings, official orders and battle reports. The collection also includes material dating from before and after the First World War. The principal correspondents are Sir Keith Murdoch, George F Pearce, Lord Liverpool, Sir James Allen, Sir Ronald Munro-Ferguson and Lady Birdwood. The Birdwood collection is a fundamental historic source for Australian First World War military history. It is a collection that is always in use at the Memorial.

The Memorial adheres to the archival principle of original order. This means retaining the order in which records were made, stored and used by the original creator or collector In this case, the collector was Birdwood himself. The order of a collection is not altered to make it fit into the digital environment. Rather, the challenge is to find a solution that replicates the physical order of a collection online.

The first stage in getting this collection to the web was to examine the arrangement. The Memorial had an existing EAD guide to this collection (http://www.awm.gov.au/findingaids/private/Birdwood.xml), but this was not detailed enough for managing the images as description stopped at the wallet level. With most digitisation projects, a far more detailed description of the collection is required and extends down to item level. This is not usually needed to manage a physical paper collection. We created a box list of the entire collection and listed every item individually. This was labour intensive and took three people working full-time, three months to complete the task. We listed the date and the title or description of every item, including identifying the writer and recipient of every letter, as well as its location in the collection. It was necessary for us to run the refinement of the arrangement and description of the collection in parallel with the scanning phase. Although the compilation of the box listing was time consuming, it has been of enormous help with other tasks required for digitising collections.

We can now manage the creation of images whilst the physical arrangement is being finalised. We simply converted the box list into a digitisation checklist by adding extra fields such as the date items were scanned and the date they were checked. (See Figure 1)

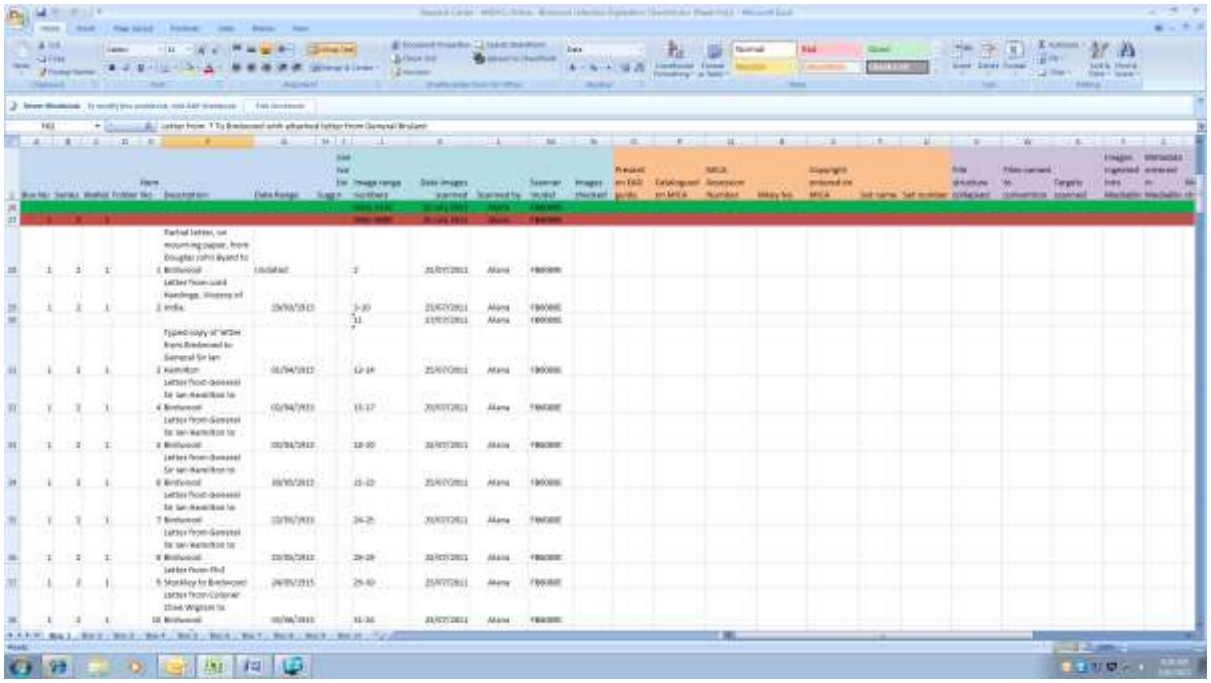

Figure 1: Snapshot of Birdwood Digitisation Checklist

Compiling the box list revealed the need to update and work on the physical arrangement of the collection. Over time, pages had been moved around within folders. Armed with an overview of the entire collection, we are now in a better position to determine where these items belong. During the scanning process, errors are also being discovered between what is scanned and the collection box listing. The digitisation checklist is annotated whenever these problems are discovered and looked into during the checking process. This is one reason why we undertake 100% checking on the collection. Every page that is scanned is being checked against the original item to ensure it has been scanned correctly and in the correct order.

The original box listing is complete and we are now using this spreadsheet as the basis for an “EAD collection listing”. The “EAD collection listing” spreadsheet will be used to create the revised EAD guide to the collection. Any changes that are made to the “Digitisation Checklist” are also replicated in the “Birdwood collection EAD”. This process is working well, as we only have two people involved in the process of making changes to our spreadsheets – these are the person who is scanning the collection and the person responsible for checking the images. The documentation that we are keeping makes it easier to identify any required changes to the directory structure that we are using to store the images prior to their ingestion into the DAMS. Again, we restrict this to two people. It is inevitable that there will be some changes between the initial box listing that we started scanning to and the final finding aid for the collection. Ensuring that both listings are accurate and up-to-date should minimise confusion during this process.

We intend to keep the structure of the collection as it appears in the current EAD guide but we will add information about the contents of every wallet and folder. We also made the decision to include two boxes of official histories in the finding aid to the collection. Although these items are part of the collection, they will not be scanned as part of this project but will physically remain with the rest of the

collection. The detailed notes that we have recorded about problems with the arrangement of the collection will be used to write a processing report when the arrangement and description work is finalised. These problems generally revolve around unidentified individuals, small typing errors, standardising item descriptions and ensuring the original arrangement of the collection is restored with the movement of items into chronological order or another series, when this action is considered appropriate.

The intention of the "EAD collection listing" is be able to transfer information directly into the revised EAD guide. However, the box listing made it clear that we needed to make some changes to the format of our existing EAD guides that would make them more conducive to digitisation projects. We started looking at what we wanted to do with our EAD guides and read widely to see what other institutions were doing with EAD. It soon became evident that there was a lot of scope to improve the appearance of our EAD guides as well as different ways to assist in their compilation. We went right back to the beginning and looked at the EAD tag library of elements and their associated attributes that is maintained by the Library of Congress. We decided to work out what tags we were currently using, which tags we wanted to start using and which tags were mandatory for archival description. We made a reference guide to the element tags. This was a very time-consuming exercise but it gave us a very good understanding of what elements we could use and refined our understanding of the ones that we were already using. We have used this to write the XML (Extensible Markup Language) code for the revised Birdwood EAD guide from scratch.

We are looking at several ways of creating EAD finding aids that are compliant with the International Standard of Archival Description (General). We are currently looking at Archivist's Toolkit (http://www.archiviststoolkit.org/) and the xml exports that this database produces. We are also looking at extending our code for Birdwood into a web based template form and whether it will work with all our collections. Another option we are looking at is whether the xml can be produced directly from our collection database, MICA. Research is also continuing into what other institutions are doing with EAD, how they are using it to present collections online and how these guides are being produced. This research is still in its early days and is based on the references provided in the OCLC research paper, *Over, Under, Around, and Through: Getting Around Barriers to EAD Implementation.* (Combs et al., 2010)

At the beginning of this project, it was estimated that the required storage for the digital images created would be 99.11GB. It was not enough to finalise the arrangement and description problems of this collection in order to be able to process these images. There were other problems that needed addressing – MICA entry, online display of records created in MICA, and ultimately, the ingestion of assets into our DAMS. These issues were all inter-related. Digital assets cannot be ingested into the DAMS unless all information is 100% correct in the related CMS, i.e. MICA. We were still finalising workflows with the new DAMS, but were possibly embarking on a new way of cataloguing our collections in MICA. We had to be mindful of what would work for the present but also in the near future. Reverting items once they are in the DAMS is difficult and this is the other reason why we are insisting on 100% checking of all our images. Although we have made some progress in this area, we are still investigating options. This has also made our

management of digital images during the interim very important and the digitisation checklist has been crucial to this. It has also highlighted the need to document our workflows and procedures and using Birdwood as our pilot, we are currently compiling a training manual for use with all of our digitisation programs. This is expected to cover areas such as overall workflows, descriptions of tasks, technical procedures and checklists.

The Birdwood collection contains items with mixed copyright. The majority of the collection is Commonwealth copyright. Permission has been granted by the Birdwood family to publish the personal records that were created by Birdwood. The orphan works in this collection are very similar in nature to the orphan works in the papers of C.E.W Bean that were published online 2009. The use of Section 200AB of the Copyright Act will be required in order to copy and publish the entire collection online.

It is anticipated that the scanning for this project could be completed sometime in April 2012. This timeline is based on current scanning rates and does not take into account any major problems that could arise in the meantime. There will still be work required following the completion of this stage before these images can be released to the web.

The Anzacs Online project into 2012-2014

The *Anzacs Online* initiative requires three main areas of technical development:

- Cataloguing: developing the internal Collection *Management* System (MICA) to support the private records collection including the more detailed arrangement and description on our collection database and what can be achieved though Encoded Archival Description.

- Web interface and visualisation – establishing what form we want the data to take.

- Linked data – web publishing – discovering what can be achieved with the new web platform Drupal and the ways that we can develop it. Working out how to engineer and present the data. Our aim is to work with tools and standards developed by the W3C (accepted standards) open source.

The Memorial aims to be experimental in its use of Drupal – the Memorial's newly installed web platform. Drupal is an open source web content management system and web publishing platform. Drupal allows us to build and maintain better relationships between our records – this aids discovery of our collection and content. The Memorial employed a web architect who had expertise in developing Drupal as a consultant to work with our in-house web team.

The extent to which the Memorial can develop the third semantic web phase is subject to funding. Releasing a Semantic Web version of our collection data involves enriching the data and its classification through the use of semantic web standards and technologies, creating the necessary cross-referencing linkage

mechanisms through text mining and analytics, putting in place new search capabilities, and setting up workflows and procedures for the life of the project.

The project will give users far greater freedom in the way they utilise our data. The use of semantic search and discovery tools will enable greater customisation and display options for search results. The project will also allow greater community engagement with the web content through social media.

Anzacs Online has been operating for several years as a digitisation project. The real challenge for any digitisation project is solving how these collections will be managed in the digital environment and how they will be displayed and accessed on the web. This paper has covered problems and issues encountered by the Memorial in starting a new project that requires new decisions and solutions. The Memorial now has the internal platforms and the ability to store and manage our preservation images within the appropriate collection management system (MICA), a means of ingesting those images into the DAMS and is testing the solution developed for online display. The Birdwood case study illustrates the very start of a digitisation process for a small to medium size collection. The next phase of Anzacs Online will be on developing the web interface and data visualisation for access online.

References

Combs et al., 2010 *Over, Under, Around, and Through: Getting Around Barriers to EAD Implementation* Combs OCLC Research and the RLG Partnership Dublin, Ohio http://www.oclc.org/research/publications/library/2010/2010-04.pdf

Ellis, J. 1993 *Keeping archives* Port Melbourne, Thorpe

First Database
http://www.awm.gov.au/collection/books/

Guide to the papers of Lord Birdwood
http://www.awm.gov.au/findingaids/private/Birdwood.xml

Mica
http://www.awm.gov.au/search/collections/

MIMSY XG
http://www.selagodesign.com/portfolio/mimsyxg/index.php

Teleread, 2012 *New BookReader on the Internet Archive* Philadelphia, Gadgetell LLC & North American Publishing Company (NAPCO) http://www.teleread.com/library/new-bookreader-on-the-internet-archive/.