IJDC / *Conference Paper*

Mad with the Writing: 100 Years of Collecting – 20 Years of Digitising – 3% Completed

Robyn Alison van Dyk
Australian War Memorial

Abstract

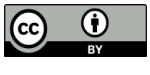
This is a story about the Australian War Memorial's historically significant, preeminent archive, which represents 100 years of collecting, 20 years of digitising and to date is three per cent digitised! Whether the format of the original record is created in paper or digital, we are all still "mad for the writing", and there is only one thing for certain: that for the next 20 years, we will still be digitising! We all want to read and access collections online. It is hoped that developments in technology will continue to make that desired access faster, easier, safer and more efficient. Artificial Intelligence will most certainly play a role in ensuring that we will not be trying to "drink from the fire hose" when accessing digital collections into the future. Transcription technologies will continue to deliver research data that is relevant today in a variety of contexts.

Submitted 31 January 2025 ~ *Accepted* 20 February 2025

Correspondence should be addressed to Robyn Alison Van Dyk, Email: robyn.van-dyk@awm.gov.au

This paper was presented at the International Digital Curation Conference IDCC25, 17-19 February 2025

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/



Introduction

This is a story about the Australian War Memorial's historically significant archive, which represents 100 years of collecting, 20 years of digitising and to date is three per cent digitised!

The Australian War Memorial is one of the most recognised cultural institutions in Australia. The vision for the Memorial came from the battlefields of the First World War as a way for Australians to remember and understand the experiences of war. The Memorial was created to be a place for the living to mourn, remember and understand and as a tomb for the fallen, fitting of their sacrifice and for their spirit to reside. The original vision for the Memorial came from Charles Bean. Bean was Australia's sole official correspondent during the First World War and also was appointed as the Official Historian of this war. Bean developed an Australian War Records Section in 1917 to collect records for the future Memorial's archive, and his vision for the Memorial building, collection and exhibits was to be a reminder of the terrible human cost of war, and an "object lesson" to "rid the world of the dreadful system which settles disputes by warfare." (Bean, 1923)

Yesterday to Twenty Years Ago—Digitisation and Records of War

In 1940, writing from "at sea" during the Second World War, Major Andrew Handley titled his letter home "Mad with the writing". He wrote: "At times, especially at night, I sit out on deck by myself & look into the water & wish that I was back home again... a man has to be an orphan to go to war... I have named the state as being 'mad with the morbids'". Handley's collection of letters home from the Middle East detailing his daily life and military service abroad in 1940 and 1941 forms one of over 15,000 collections of personal letters and diaries archived by the Australian War Memorial and is on our digitisation to-do list!

Today, digitisation is no longer the new shiny thing for museums, archives and libraries. For most cultural institutions in Australia, especially those that are state- or nationally-based, the practice of digitally preserving paper-based collections is 20 years old and now a well-trodden path. For the Memorial, digitisation is business as usual. The Memorial allocates annual funding in its budget for digitisation, and there are permanent staff employed in digitisation roles. Digitisation is one of the key established fundamental preventative conservation procedures for high-risk collections in the Memorial's archive.

The Memorial's archive is accessible by the public and, if a digitised copy is not available, the collection is made available in the original format. The Memorial implemented digitisation as a means of protecting its archive from degradation through handling and use. The Memorial's preservation principle was that once a collection is digitised, it is made available on the website, and the original is only then accessed for special purposes.

Twenty years of digitising has improved speed and processes for digitisation and built up an institutional confidence in requirements and outputs. Digital preservation standards are currently settled upon, and the infrastructure is established and in place for sustainable digitisation programmes to continue. The Memorial is in a position to scale up quantities when we receive sponsorships or a government grant. In 2019, the Memorial received $8.3 million from government to digitise its at-risk collections. The Memorial was able to scale up its business and processes and, although very interrupted by the Covid pandemic and shutdowns in Australia, delivered these collections to the web.

Twenty years ago, the Memorial was an early adopter of digitisation. The Memorial began publishing official records online in 1999, creating and publishing a Roll of Honour database along with the original source documents, the 'roll of honour circulars'. Twenty years ago, the Memorial's first foray into digitisation was to target heavily used official records. People were seeking data around their ancestors' service. The public wanted records that gave them fundamental information including dates, military units, places and operations. Not technologically capable of digitising private records, the Memorial tackled digitising its extensive official record collection with the intent to make it available on the website. Collections targeted for digitisation 20 years ago gave people the

data about military service that they required, and digitisation and publishing online prevented original First World War nominal rolls from being thumbed through and used on a daily basis. Copyright for publication was easier to obtain for official records, and the pages were relatively uniform and able to be captured using the old flatbed scanners that we used back then. There are still more than 500,000 grayscale images captured between the years 2000 and 2010 displayed on the Memorial's website from these years. The Memorial has not gone back and recaptured these in colour as we have too many other collections needing digitisation, and so we keep forging forward.

The digitisation of official records was a relatively easy start to the Memorial's digitisation journey. Originally, the large series of images for official records were unable to be controlled in the Memorial's Collection Management System (CMS) and were stored on a server and controlled via a Microsoft Access database. All the digitised archival collections existed outside of the Memorial's CMS. Images were bundled into PDF files reflecting the original files and published to the web in the collection's original archival hierarchy, all controlled by the Access database. When the Memorial implemented a Digital Asset Management System (DAMS) in 2009, and integrated the digitised collections into the DAMS, the PDFs were treated as digital assets, and then of course, every time there was an error or an image needed to be corrected, we were messing with an object that we had deemed as an asset.

By 2014, it was realised that the Memorial's digitisation processes were completely unsustainable. We needed to ditch creating PDFs as a means of bundling collections for publishing. We needed a more robust platform than the Access database, and we needed new procedures for controlling and managing the digitisation of hundreds of tiny private records collections and their images that the public wanted to access online. We needed to automate the process.

Luckily, in 2014, the Memorial received from government some additional funding to deliver to the website collections that reflected the centenary of the First World War. It enabled the Memorial to change our processes and build a sustainable platform. We started to work with the owners of the existing CMS to enable collections to be managed in hierarchies. This work enabled a parent record and the child records to reflect the nature of the original files and allowed the Memorial to publish records in a logical way that reflected the analogue paper arrangements. We split the collection management in the database under a parent record—almost like a tree with two different branches. One side of the tree would reflect the paper records. The other side would manage the digitised collections. We utilised Apache Solr (an open-source platform) to dynamically bundle and publish to the website images that were indexed and stored in the DAMS. Rather than manually creating PDFs, we now had a more dynamic means of bundling images on the fly that would be responsive to any changes or rescans required. Our CMS in 2015 had 500,000 catalogue records, mainly related to objects and photographs, and that year, a project was developed to successfully integrate the archival collections—2 million catalogue records—from the Access database into the Memorial's CMS/DAMS. In 2015, the Memorial achieved having all its digital collections in the one database.

In the lead up to the centenary of the First World War, the Memorial also began digitising its collections of personal manuscripts, letters and diaries. The public were seeking more online context to reflect the experience of war; what it was like in the trenches of Gallipoli; or fighting on the Western Front. The key deadline for the online release of these records was 25 April 2015. This date was the centenary of the Gallipoli campaign of the First World War and a highly significant date for Australians to mark and commemorate. Copyright was a huge issue for web publishing at this time. We had built the platform and capability but were slowed to a snail's pace by the necessity of dealing with permissions. Australia since 2015 has undergone some beneficial reform to the Copyright Act 1968, including, in 2017, the removal of perpetual copyright on unpublished works.

The Memorial's systems established during this time were robust and low-maintenance, and they served the Memorial effectively until recently—when the old CMS was no longer supported by the company, and we had to get a new one. The Memorial's new CMS looks great and has a good web interface, as well as enhanced search and usability. However, due to the proprietary nature of today's CMS, there was no more adjusting the back end of the CMS for our bespoke requirements. The CMS was out of the box. The Memorial, being a shrine, museum and archive,

has struggled to make all our business fit neatly into out of the box CMSs. Rather than being able to tinker and adapt the CMS to meet our needs, we now have to pay for changes and use prescribed web scripts to bulk upload metadata. Today, we are less agile than we were yesterday. Today, we are busy wrangling with our new DAMS and CMS.

The Memorial's current permanent digitisation team involves a Senior Curator and two curators who manage the digitisation team and deal with collection policy and legal frameworks, selection of material for digitisation, managing vendors and sponsorships. In addition, there are four permanent Assistant Curators who manage and curate the data within the systems, including doing copyright research and managing permissions; and there are four image production staff members to scan the collections. All images are ingested, managed and preserved in the DAMS and published to the web from the DAMS as bundled images. Scanning is relatively quick and easy. We do a 100 per cent check for accuracy, but the time and resources expended are mainly in the creation of the metadata for the management and curation of the images in the systems.

Indexing the Archive, Curating and Publishing Data

The by-product of digitisation is improved data for search and discovery, but this also comes at a cost in resources and labour. The Memorial's recent project to digitise the 3-million-page Second World War Commander's Diaries, for example, had an 18-month lead-in time for deeper cataloguing. This was required to ingest the images into the CMS and DAMS and display the images online. The need to consider the lead-in time to digitise a collection has been a lesson learned from previous projects and is now carefully factored into the budget and timeline of any project.

Once digitised, a collection offers many opportunities for the generation of beneficial data. The rolls data on the Memorial's website amounts to approximately a million names and is enriched through related and linked data. This data has been generated through the help of many hundreds of volunteers. The Memorial has been digitising and deep-indexing its online collections since 1999. The data includes conflicts, names, places, military units and dates and is curated and displayed in searchable format on the Memorial's website. The indexed data reflects the original archival records and usually displays alongside the digitised page for historic evidence. This data is deeply valued by the Memorial and includes the Roll of Honour, which details members of the Australian armed forces who have died during or as a result of warlike service. The Memorial preserves all its indexed data in its CMS, and the data is published via the CMS and DAMS to the web.

Last year, the Memorial launched a transcription tool. Since public launch on 14 February 2024, the response to the Transcribe Project has been truly overwhelming. In just over 48 hours, more than 1 million words were transcribed by more than 20,000 volunteer transcribers. To date, this has now increased to over 2.5 million words from over 32,000 contributions. Transcribe has been used throughout 2024 to promote the richness of the collection, engage with supporters of the Memorial and position the Memorial as a leader in embracing innovation. The transcriptions improve readability and public accessibility and most importantly create machine readable text that can be used for a variety of socially beneficial purposes. The content is discoverable, searchable and usable. The Memorial views the transcribed collection as another form of digital preservation for these precious words written on the battlefields more than 100 years ago. The data generated will have research application.

The Memorial has much more to do in the area of indexing historical data and crowd sourcing transcription. In 2025, the Memorial will explore using the transcription tool to create research data sets related to the some of the Memorial's important historic collections. A collection that consists of lists of and administrative information about Australian army vehicles from the Second World War through to the end of the Vietnam War will be the first that we will trial. The Memorial also holds detailed records of the spread of malaria during the Second World War and unique Influenza research data collated in 1918–19. The influenza data was captured in hand-drawn spreadsheets and represents hundreds of pages of historic data that is known and still accessed today by virologists but really needs to be digitised and indexed. These collections form part of the Memorial's future plans for the transcription tool.

A unique set of over 2000 aerial photographs of Palestine captured during the First World War has had considerable international interest. Parts of this collection are grid-referenced to an existing

set of First World War topographical maps. The collection also includes in-depth diagrams, depicting the way the images can be placed together to create mosaic images. The Hebrew University in 1978 sponsored the indexing of the collection, and most recently, the Australian War Memorial working with volunteers re-indexed and enhanced the indexing of this collection. The collection is currently being partially digitised thanks to a university sponsorship from the UK for archaeological research, and the data and images will be published on the Memorial's website in 2025 and made freely available for all researchers. Digitisation is the start, but there is much that can be done with the images, data and geo-referencing.

Tomorrow and the Next 20 Years: Access to Recent and Digital-Born Material

While three per cent doesn't sound like very much, it represents over 4 million pages from the archive that have been scanned and published on the Memorial's website.

Digitisation will continue into the next two decades. Given the trends of the last 20 years, the process will hopefully be faster and more efficient, with better machines able to capture a higher standard of image. TIFF files are likely not to be considered a preservation standard, and we will be relying on our DAMS to identify obsolete files and convert them into a readable current standard. The Memorial will continue to prioritise paper collections that are in high circulation to the public through the reading room for digital preservation and access online.

Today and in the coming years, people expect to see more recent collections accessible online. To date, the Memorial has focussed mainly on delivering collections related to the world wars, but in the last 20 years, Australians have been involved in many conflicts and peacekeeping operations, and veterans and families are in need of access to material related to these recent conflicts. Records created within the last 20 years come with a whole spectrum of issues that older records do not have. The younger the records, the higher the sensitivities. Privacy is now a huge issue in relation to this material. Email and other personal information needs redacting, but also these are records of war and may inadvertently reveal matters of operational security. Australia's Archives Act 1983 provides easy answers to these questions, but web access for private records and manuscripts that are under 20 years old is not covered by the Archives Act 1983, and these issues require a new policy. Our current solution is to closely read these collections, consider their suitability for online access, discuss areas that require redaction and also, if documents are not suitable for the web, make decisions on the level of access we can give to this material. This is a huge labour soak but needed if we want to make more recent collections available, and we are chipping away at it slowly.

The near future will see the Memorial managing the growing cost of storage and ensuring our systems are robust in managing the long-term access to digital collections. There is a requirement under the Archives Act 1983 for the Memorial to make available for access records that are over 20 years old, and this includes born-digital records. There is also a requirement under the Australian War Memorial Act 1980 for the Memorial to develop and maintain "a national collection of historical material". Since its inception in May 1917, the Memorial has collected and managed the operational records of the defence department as well as its own corporate records. The Australian Department of Defence has two petabytes of digital records to transfer to the Memorial, which will grow to over ten petabytes over the next ten years. These are operational records related to conflicts that Australia has been involved in over the last 20 years. Given nascent developments in archival research and big data, it is likely that the solution for management of, control of and access to these records will need to have capacity to facilitate an artificial intelligence (AI) interface. There is already a growing focus on AI within the Australian government. New technologies will transform the way we work and also mean widespread change at a scale and speed that is relatively unprecedented. The use of AI in the Memorial's Transcribe project, search results and face-matching in photos is only the beginning. It is likely that AI will play a role in controlling the Memorial's digital-born records into the future. Cataloguing text and titles and dynamically enhanced search across digital collections is not too far away.

To date, the Memorial has relied upon facilitating public access to digital collections by publishing them on our website. If the Memorial can't publish a digitised or digital-born collection to the web for reasons of sensitivities or copyright, then how are we to give the public access to this material? Print it out?—Not likely! And so currently the Memorial is working on offering a form

of supervised digital access that would be available only in our public reading room. At a very basic level, the records could be loaded onto a stand-alone computer, which is clunky and not a solution. Alternatively, there could be exploration of IP-based and local access in the reading room; or code-and-password access could be provided to documents stored within the DAMS. This third option is looking most likely for 2026.

The Memorial's main lesson learnt from the last 20 years is that we are thankful that we got started early and learned along the way. We live with imperfections of the past, including dirty data, grayscale images and catalogue data that may be minimal. We have learned that we have an army of volunteers who just need the necessary web tools and platforms to help us bring the collections and data to the web. The rolls and collection data are highly significant and useful for the public who access them, but in recent times, we have learned that our data management, preservation and web display are bespoke, and any transfer of data to new systems is complex and resource-intensive. It was 20 years since we last transferred and changed CMS, and due to this, it was bound to be complex. How we code and manage the web interface between our CMS and DAMS is now more modernised and better-documented for any future troubleshooting.

Whether the format of the original record is created in paper or digital, we are all still "mad for the writing", and there is only one thing certain for the next 20 years: we will want to read it all online. It is hoped that developments in technology will make that access faster, easier, safer and more efficient.

References

Australian War Memorial registry file, first series] Australian War Memorial (Its nature and method of raising funds) and Provision of site at Federal Capital AWM 93 12/12/6; Bean, 1953.

Australian War Memorial People data. https://www.awm.gov.au/advanced-search/people

Australian War Memorial Collection data. https://www.awm.gov.au/advanced-search

Australian War Memorial Transcribe. https://transcribe.awm.gov.au/