PRESENTATION TITLE

ABSTRACT

ANZAC Connections – delivering and connecting real content and data online

Which servicemen from the Traralgon area, fought and did not return? How many Indigenous Australians from Bundaberg Queensland served in the First World War? Which Australian towns was hardest hit by the war? Each month the Australian War Memorial receives thousands of public enquiries such as these. The answers, often yet to be researched, are locked away in analogue pages or in un-integrated or semi-structured data sources within the Memorial's collections and management systems.

The Memorial's extensive online collections have been created from over a decade of digitisation projects and a century of collecting and collection management. Data from cataloguing, indexing and digitising collections has enormous research potential when placed on the right platform.

The Memorial's major digitisation, data and web development project *ANZAC Connections*, launched in December 2013, brings historic documents from the Memorial's archive to all Australians and has delivered an appropriate platform to integrate and make available a substantial collection of rich data that exists from a variety of sources.

In the lead up to the centenary of the First World War the project will progressively deliver new collections to the website, linked open data (LOD), vastly improved search and discovery on the site, and provide ways for people to interact with the collection including tagging and transcribing the collection.

The project blends museum, archival and library material to establish interoperability and discoverability across not only the Memorial's digital historical collections but also linkage with other major cultural institutions including the National Archives of Australia, State Library of NSW and the National Library of Australia.

This paper is not about showcasing a completed project but will demonstrate the Memorial's concept and discuss the issues, difficulties and problems overcome along the way in delivering this major digitisation project.

PAPER

DIGITISATION

The Memorial has a wealth of online digital content giving researchers opportunities to work with primary sources that would have been unimaginable a decade ago. Millions of collection records lead people to film that they can download, photos and art to view, interviews to listen to and books and archival records to read online. The Memorial's latest digitisation project *Anzac connections,* is focused on delivering First World War manuscripts to the web. The project involves the progressive digitisation of thousands of pages of personal letters and diaries from the First World War[1]. These historic collections are original eye-witness accounts of the major events of the war and offer a rich and personal perspective of people's unique experiences.

[1] http://www.awm.gov.au/1914-1918/anzac-connections/

The Memorial's First World War digital collections are freely available online and are heavily used by the public offering valuable resources particularly for all humanities disciplines. These digitised collections are used for all manner of research, however, the data that is created from digitisation and cataloguing has been underutilised.

The Memorial's First World War digital archive includes: Australian Imperial Force unit war diaries, 1914-18 War[2]. Embarkation roll details of approximately 330,000 AIF personnel, recorded as they embarked from Australia for overseas service[3], Recommendation files for honours and awards, AIF, 1914-18[4], First World War Nominal Roll details of approximately 324,000 AIF personnel[5], Red Cross wounded and missing files[6], records of C E W Bean[7], private papers of 150 people who served in the First World War digitised[8]. There are also photographs, film, sound, art, objects and digitsed ephemera series including First World War concert and theatre programs[9].

DATA

Arranging, describing and managing digital collections and providing access to these collections online generates an enormous amount of contextual metadata. Name, place, unit, date and conflict data is generated by the process and is published as part of the catalogue record. Collections that hold biographical data, for example, have been scanned and indexed by the Memorial and published online as searchable rolls.[10] Archival collections are often box listed by the item for tracking during the digitisation process:

Item 1. Farquhar[11]'s works diary from his time at Quinn's Post. Date 13 September - 11 December 1915
Item 2. Sketch of Turkish Mortar Bomb by Farquhar. Date 14 October 1915
Item 3. Farquhar's 2 page description of detonation of a charge at Quinn's Post. Date 11 November 1915

If this data can be made available in the online environment in a way that is more machine processable and linkable, it would not only more efficiently lead people to the Memorial's extensive online digitsed primary sources but would also provide a wealth of research data that can be used in ways beyond its original purpose.

The Memorial is not only looking to make contextual metadata more broadly available but also seeking to bring out and create more data currently locked down in the text of the documents. The Memorial has close to a million archival pages available online as scanned images. The text within the digitised document is not searchable. The only data the search engine is able to pick up is the cataloguing. Transcription of the text within the document to machine readable text will allow

2 While on active service, Army headquarters, formations and units were required to keep a unit war diary recording their daily activities. Series AWM4 comprises the diaries of the Australian Imperial Force (AIF) created during the First World War. http://www.awm.gov.au/collection/awm4/
3 http://www.awm.gov.au/people/roll-search/nominal_rolls/first_world_war_embarkation/
4 http://www.awm.gov.au/people/roll-search/honours_and_awards
5 Nominal Roll recorded to assist with their repatriation to Australia from overseas service following the First World War. http://www.awm.gov.au/people/roll-search/nominal_rolls/first_world_war/
6 These files consist of approximately 32,000 individual case files of Australian personnel reported as wounded or missing during the First World War. http://www.awm.gov.au/people/roll-search/wounded_and_missing/
7 http://www.awm.gov.au/collection/AWM38/
8 The list is at http://www.awm.gov.au/people/biographies/
9 http://www.awm.gov.au/collection/PUBS002/
10 See http://www.awm.gov.au/people/roll-search/all/
11 See, for example, how this data is displayed online. http://www.awm.gov.au/collection/1DRL/0278/

deeper searches into the collection to occur. Through transcription the search engine will be able to identify key pieces of information within the scanned images of a page. This will make the items more accessible and allow for more intelligent searches to be conducted on transcribed collection items.

Given this process is manually intensive in nature; transcription will rely on crowd sourcing the input of interested members of the public. The transcription tools will be made available for the public to access online. The creation of the Memorial's online transcription tools is currently under development and will be delivered in 2014. Transcription will be based on the Wiki model of open and collaborative document management. Where possible, the tool will promote the addition of relevant subject tags which will add to the value of the collection.

The *Anzac Connections* project aims to enable people to have boarder access to data in ways beyond the current static display on the website. A key aspect of enabling data to be more useful for humanities research is to ensure that it can be processed by machines and made interoperable across systems. This needs to occur while still retaining the provenance and complex layers of meaning behind the data. The Memorial's data is published with its authoritative sources and this needs to be understood by people when accessing or linking to the data as well as when making the data more usable and extractable.

A major priority of *Anzac Connections* is to ensure that the digitsed material is being published in a way that it could not be isolated from its metadata as part of a search result. When people are searching we did not want them to find snippets of a document in which a given term is mentioned and then not be able to find the next page or even get a sense that there is a document in its entirety related to that page. The Memorial's collection data is exposed to Google and so it was important to ensure that search engines would not contribute to this problem of isolating a page from its collection and archive of origin. The Memorial's publishing of archival collections exposes the organisational structure of collections.[12] This not only preserves the original context of the documents but also alerts researchers to the types of other materials available.[13]

LINKED OPEN DATA

The Memorial's broader aim for the 2014 delivery of the *Anzac Connections* project is to begin the process of transforming collection cataloguing and description data into linked data. The process towards implementing linked open data (LOD) commenced in 2013 with the integration of the Memorial's two major collection databases to produce a core AWM collection management system that we would use to publish data to the web. This allowed the project to reduce duplication and ambiguity in the data that we were publishing.

After over a decade of creating data the migration, merging and clean-up was a massive processes. The project was working with live data that is accessed by the public on a daily basis. The Roll of Honour database in particular, as a major flagship of the Memorial, could not be open to error and loss of data.

The Memorial's in-house built *Access* Database, containing over two million collection records, was integrated into the Memorial's museum catalogue system (*MIMSY XG*) consisting of 500,000 records. The *Access* Database was originally

12 For example - http://www.awm.gov.au/collection/PUBS002/
13 For example - http://www.awm.gov.au/collection/PUBS002/001/001/001/001/ in above

devised in order to index large archival collections, for example, the *First World War unit war diaries*[14]. *Access* was used because a solution for managing these types of digital collections could not be found in the *MIMSY* CMS, The collection data stored in *Access* was only ever indexed into flat tables and there were no relationships identified between the different tables even though the data was about the same people, events, units and places. Over time *MIMSY XG* has technically developed a greater ability within the system for curators to build hierarchies, relationships and linkages. Although *MIMSY* held less collection data it was chosen because it was stable, had the performance ability to handle large quantities of data and had much greater potential as a data publishing platform. The data was successfully migrated in 2013 and is now published to the web via the *MIMSY CMS.* The process of data cleansing and merging will continue for many years to come.

One of the greatest concerns with the migration of the data was maintaining its integrity when forcing it into a different defined authority structure. We had to ensure that we were not losing information or context. There were many more thesauri terms now available to curators, however, there were also multiples of the same people, places, units and conflicts that required cleaning and merging. Because the data had context and meaning we could not rely on the machine to do the merging without human intellectual evaluation.

The project is focused on publishing to the web a clean set of place, conflict and Australian military unit data in 2014. The successful migration of the data onto the same platform is the beginning of the project's ability to publish and interlink structured data on the Web so it is more extractable and usable. The project made some other significant developments in 2013 to achieve linked data. The has occurred through updating search aspects so that data can answer complex questions through automated methods, developing and improving cataloguing of records by building strong internal links across collection records and improving data deployment processes via SOLR[15].

This year represents 100 years since the beginning of the First World War and many cultural institutions across Australia will also have a project underway to mark this centenary period and will be interested in exposing their war related material. With archival collections related to this war reaching 100 years old many institutions are looking to digitally preserve this material and display it online. If institutions are looking to publish using linked open data (LOD) then an agreed shared ontology vocabulary is important otherwise the descriptions or relationships will exist in a silo with no interaction with outside resources.

The Memorial is seeking to link not only within its own collections but also to relevant outside sources initially focusing on the First World War. LOD if implemented correctly and in coordination with a number of other cultural heritage organisations has the potential to make the digitised First World War collections more accessible. Relationships between collection items held in different institutions can be recognised with greater depth than is currently possible. The automated recognition of these relationships will allow user searches to return results which are richer in content and more complete in terms of the story they tell. Information gained from these types of user journeys will contribute to a greater public understanding of the Australian experience of war. Promoting and enabling interoperability, sharing and reuse of data across the distributed national collection begins with a shared ontology.

---

[14] http://www.awm.gov.au/collection/AWM4/ was originally indexed in *Access* but the data has now been migrated to *MIMSY XG* - It is published to the web via the *MIMSY* database.

[15] http://lucene.apache.org/solr/features.html

In August and November 2013 several key institutions working on similar First World War projects including the Australian War Memorial met to discuss how we should proceed with developing linked open data. How do we develop a vocabulary of Australia's experience of the First World War at the national level? Do we as cultural institutions develop our own ontology from scratch or investigate the use of existing First World War LOD to possibly include Australian context and vocabulary?

From the meetings it was discussed if it was best to expand on an existing ontology, perhaps extend the existing European ontology, for example, the ontology "ww1 LOD" Europeans data model. However, the existing data doesn't deal with people and there were no Australian centric terms. For example, in the existing WW1 LOD Gallipoli is just a place. However, for Australia it means a lot more and there was no markup available synonymous with other meanings.

Australian Cultural institutions are looking for people, unit, place and event linked data as a means of also linking their various projects and making them more discoverable online. The meetings were the first step in discussing and exploring possibilities for our institutions around using LOD. To start enabling the definition and discovery of relationships between people, places, events and objects of the First World War and work together to define data and reach agreement on requirements for linking and online interactions. It was also an important meeting to understand what each institution is developing in relation to the First World War including what kinds of data will be created and what entities within the data and the data sets are intended to be brought together.

There was consensus on the need for an Australian First World War standard reference dataset that could be used to bind together historical collections dealing with the war. The publishing of a common vocabulary if adopted would help institutions across Australia enrich their online records and provide a means for all cultural institutions to expose their collections using common relationships and links.

Enabling the definition and discovery of relationships between people, places, events and objects begins with the publishing of the URI[16]s. The URIs will be linked with related objects across First World War datasets using a published vocabulary. It is hoped this will provide semantically rich searches that will enable our users to more easily find and use our online primary sources.

Establishing unique identifiers (people, places, units, events) across the institutions is a big job and in dealing with LOD it is important on limited budgets to not bite off too much. Given the very specific nature of the First World War and the lack of existing specific Australian WW1 ontologies, it was agreed the Memorial would deliver a vocabulary for Australian Imperial Force Units, Place and Conflict. This in itself proved difficult as the Memorial could not just rely on our existing collection thesaurus, which having been created from many hands and decades of cataloguing, was incomplete and inconsistent for the purpose of creating an Australian WW1 Military unit vocabulary. We also had 17,000 duplicate records to research and clean from the integration of the two databases.

During 2014 the Memorial has worked on delivering a clean set of place and unit data for web publishing. The unit data is arranged in a branch structure based primarily on the collection of Official Commander's Unit war diaries[17] but also using a

[16] http://en.wikipedia.org/wiki/Data_URI_scheme
[17] http://www.awm.gov.au/collection/AWM4/

variety of other authoritative sources. The Unit structure is organised by conflict, nation, and context and can be linked with date data to provide an order of battle. This data was cleaned internally by closing off the existing structure, releasing the new structure for curators to use, and merging the old unit entries into the new. All the Memorial's collections will be linked to this new First World War Unit structure when completed. This data is intended to be published as URIs in 2014. For the unit structure to work as LOD, however, requires its adoption by other cultural and heritage institutions.

A published set of names related to all Australians who served in the First World War as URIs that all our respective institutions can link our data to would open up a massive amount of information and opportunities to understand the story of Australians who served. The *Anzac Connections* project is also working on merging people data but more in relation to particular projects – *Anzac Connections*, Indigenous service, and prominent people[18]. Working with this data is time consuming and requires human research and checking. First World War service numbers and names were not unique, sometimes when merging data curators need to rely on a third piece of information to distinguish between people. While the Memorial holds many First World War people name datasets we do not have a publishable set representing all Australians who served.

USER EXPERIENCE

To gain a better understanding of how people might be engaged with the *Anzac Connections* project and its development a User Analyst consultant was appointed to research and produce a report on how the public might be engaged in the project. The research aimed to develop an understanding of who uses the Memorial's website, to identify the types of users who may be interested in the developments and in using crowd sourced functionalities and to further understand the public's information needs and general motivations towards participating in developing the Memorial's collections and their context.

The consultant's engagement will contribute to the design and development of the new functionality of the website, and has led the project to explore ways to attract users to the engagement tools that are being developed and provide feedback for future developments. The report recognised strong themes of pride, belonging and identity in the public and identified this is as a powerful motivating tool in encouraging users to take a more active interest the Memorial's website as a whole.

Visitors to the Memorial's website are generally seeking something. They are researchers with varying degrees of skill. The research agenda will range from non-specific, serendipitous discovery to highly specific data discovery and acquisition. The research skills will vary from basic to advanced, from unfamiliar with the site to very familiar. Knowledge of military history will range from basic to advanced, from unfamiliar with the information on our site to very familiar. Frequency of visits will vary from incidental to frequent.

The Memorial is looking to enhance access for all researchers. Success on the site will encourage people to come back. The project will deliver collections online for the broadest public access. The collections are delivered via a platform that uses a consistent display with simple, intuitive navigation. The Memorial will progressively add newly digitised collections to the site over the centenary period.

[18] Page under construction link to be supplied.

Through making data available on the web in a more usable and open format the Memorial is creating the opportunity for researchers to quantifiably analyse our content. We are providing the ability to looking for patterns, anomalies and interconnections. To better understand the history and events of the First World War. To scientifically study trends in the data for purposes not even related to history, perhaps the incidence of Spanish Influenza, the weather or the technology of war and its performance. The website should help the user to develop skills in researching the First World War by asserting relationships and showing how elements of the content link together. Users should be able to create their own stories using our data and sources.

Ultimately we would like to attract more visitors to our site and encourage people to share this information, contribute and create new knowledge. The success of the project will depend on how well discoverability has been enhanced. If people who are unfamiliar with the site can easily find what they want to help their research with varying level of skill, knowledge, and interest then this project will be considered successful.

The Memorial is also seeking an emotional connection, an empathy with the content on our site; a personal connection to a discovery, a person or other aspect of our content that might lead to a sharing a discovery with someone else. Crowd sourced information will help the Memorial and others and hopefully build a community of people who are interested in participating on our website. The project aims to bring more people to the site and encourage them to stay longer, searching more widely and more deeply.

CONCLUSION

The Memorial's Anzac Connections project is designed to create a meaningful and sustainable presence on the web for the centenary. It brings to the website original primary content related to the First World War by delivering new significant historic digitised collections and creating new and improved data. Through making our data more extractable, usable and discoverable the Memorial is seeking to make information to be combined in ways yet to be imagined.

The Memorial is also seeking to incorporate the information of other linked data providers into its local description. This will enable a fuller more value added description. There are many significant First World War collections held across the country from the National Archives of Australia's First World War Service Records, The State Library of New South Wales's manuscript collections, Trove's newspaper collection and every library, archive or museum across the country who holds collections large and small related to the First World War. It is hoped that the centenary of the First World War becomes a catalyst for cultural institutions to link and share data and experiment with LOD.

BIBLIOGRAPHY

Juha Törnroos, Eetu Mäkelä, Thea Lindquist, Eero Hyvönen. Leveraging linked data to enhance subject access - a case study of the University of Colorado Boulder's World War 1 collection online. In *World Library and Information Congress: 78th IFLA General Conference and Assembly, Helsinki*. IFLA, http://conference.ifla.org/past-wlic/2012/117-lindquist-en.pdf, August 2012.

Juha Törnroos, Eetu Mäkelä, Thea Lindquist, and Eero Hyvönen. *World War 1 as Linked Open Data*
http://www.semantic-web-journal.net/system/files/swj458.pdf

Linked Open Data in Libraries, Archives, & Museums
LODLAM - http://lodlam.net

Eric Lease Morgan and LiAM. *Linked Archival Metadata: A Guidebook*. Version 0.99, April 23, 2014
http://sites.tufts.edu/liam/
http://infomotions.com/sandbox/liam/tmp/guidebook.pdf

W3C Wiki
http://www.w3.org/wiki/Main_Page

Wickett, Karen. *Collection item metadata relationships*. Thesis, University of Illinois at Urbana-Champaign, 2012
http://hdl.handle.net/2142/42198

Wikipedia,
http://en.wikipedia.org/wiki/Data_URI_scheme

Jennifer Zaino. *Linked Open Data In Action In World War I* Showcase Project
July 13, 2012

http://semanticweb.com/linked-open-data-in-action-in-world-war-i-showcase-project_b30723