MEASURING THE ANZACS: LESSONS FROM DIGITAL HISTORY WRIT LARGE

Evan Roberts

INTRODUCTION

The centenary of World War One (WW1) was commemorated by an appropriately globally diverse range of public and scholarly explorations of the conflict. Digital initiatives by archives and museums were a significant aspect of the commemoration, facilitating historical research and community engagement with artefacts, records and exhibits. In Australia, Canada and Great Britain, as well as in New Zealand, an important public history initiative was completing the imaging of the vast majority of soldiers' personnel files from the conflict.[1] These files constitute an important resource for historians, with significant detail on the pre-war, wartime and post-war lives of around 10 per cent of the population. In New Zealand (and Australia), WW1 personnel files are even more important as a source for social history, because of the contemporary destruction of census manuscripts.[2]

The vast size of the New Zealand collection – 3.7 million page images – means that significant additional indexing is required to make the collection more useful for research. However, the information in the personnel files is largely handwritten, typed in small font, or has poor contrast and resolution. Thus, human transcription on a large scale is needed to index additional information in the personnel files beyond full names and serial numbers.

Measuring the Anzacs is a crowd-sourcing project launched in 2015 to address this need. The goal of the project is to work with volunteers in New Zealand and around the world to complete the transcription of key elements of all 140,000 personnel files that were digitised by early 2015. Launched on the Zooniverse platform in October 2015, the project achieved substantial public engagement for its first two years. Software challenges prompted a fallow period, followed by redesign and relaunch in October 2019. The achievements and challenges of the project over the past seven years hold important lessons for sustaining online history projects with ambitious goals. In this article, I give an overview of the history of the project, demonstrate some of the research potential of a large transcription project, and conclude with reflections on how digital public history is more successful in conjunction with the in-person engagement of an interested public.

[1] Kris Inwood and J. Andrew Ross, 'Big Data and the Military: First World War Personnel Records in Australia, Britain, Canada, New Zealand and British Africa', *Australian Historical Studies*, vol. 47, no. 3 (2016), pp. 430–42.
[2] Len Cook, 'Constraints and Conflicts in Access to Official Statistics and Statistical Records', in Brad Patterson (ed.), *New Zealand Archives Futures: Essays in Honour of Michael Hoare*, Wellington, 1996, pp. 173–80; Erik Olssen and Maureen Hickey, *Class and Occupation: The New Zealand Reality*, Dunedin, 2005, p. 29.

BACKGROUND

World War One was a modern war in its application of industrial power to fighting, and in the scale of bureaucracy required to manage the war effort. In New Zealand nearly 10 per cent of the population served: around 100,000 men and several hundred women. Effective military operations required officials to know the background of people who had served, and document their service to manage it. At the conclusion of service the military had to either manage the dead – burying and informing – or for the living, contribute to the administration of post-service benefits. The technology of the time meant that these operations were performed on paper, requiring nearly 1,000 staff by the end of 1917.[3] While more than 3.7 million page images existed in the collection in 2015, the paper files had been considerably trimmed since the war. Personal correspondence from family members to the Defence Force survives infrequently in the New Zealand records.[4] Conversely, some files were large enough that information on men's service was split across multiple files and existed as separate archival entries at Archives New Zealand.[5] In the 1960s parts of the personnel files were microfilmed, with some of the original paper files discarded on the completion of microfilming.

Unrestricted public access to the files began in 2005 when the Defence Force transferred paper WW1 service files for 122,000 servicemen and women to Archives New Zealand (hereafter Archives). In practice original files had to be copied to protect the original paper, and to reduce demands on microfilm readers in the Wellington office. Individuals could request copies of only two personnel files per week, but copies made for others could be examined. When making copies by request, Archives typically produced three copies: one for the requester and two for others to view onsite at an Archives office. Thus, practical access to the files expanded slowly but at a rate significantly behind demand. In 2008 the reproduction workflow moved to a digital platform, although at first scanned copies were not available within the archival search system. From 2010 scanned files became available in the Archives catalogue attached to individual records, making remote access to records straightforward. Thus, a significant number of New Zealand files were being digitised in response to public requests, while peer countries Australia, Canada and Britain had determined efforts underway to digitise all records. In 2013 the New Zealand government adopted the goal of scanning all WW1 personnel records by the 100th anniversary of the start of the war.[6] Although overtime had to be worked in the last month to achieve the goal, nearly all the records of men and women who had served in WW1 were publicly available by August 2014.[7]

---

[3] Kathryn Hunter, 'Red Ink, Blue Ink, Blood and Tears? War Records and Nation-making in Australia and New Zealand', *Rethinking History*, vol. 22, no 3 (2018), p. 407.
[4] Ibid.
[5] These files have a different 'R number', a unique number assigned to every distinct item held by Archives New Zealand.
[6] Paul Easton, 'Kiwi Soldiers' Diaries go Online', *Dominion Post*, 29 August 2013.
https://www.stuff.co.nz/dominion-post/news/9103269/Kiwi-soldiers-diaries-go-online
[7] Archives New Zealand online manager Alan Ferris told me in a March 2015 conversation that the scanning team worked 'round the clock' in three shifts over the last month (July 2014) to complete the project.

Reflecting their origins in a military filing system, the personnel files were indexed only by surname and serial number. Published embarkation lists and other lists of military units allowed researchers to identify the files of men who had served in particular units, and begin reconstructing their histories. Genealogists, relying on names, and military historians, organising their work around identifiable military units, are reasonably well served by this indexing. However, the research potential of the files is considerably greater if other text in the files can be transcribed and indexed for electronic searches.

Databases benefit from standardisation and making similar information available for every person. In the New Zealand Expeditionary Force, five standard forms describing people's lives, physical characteristics and military career are readily available. The events summarised on the standard documents provide the context for the significant number of other papers in the files, which average 26 pages over 140,000 records. The key documents are:

- *Attestations* describing social, economic, family and demographic characteristics, and prior military service. Attestations changed format incrementally during the war, with more than 30 different versions.
- *History Sheets* summarising administratively relevant experiences a soldier had, including being wounded, and the dates of his service abroad.
- *Statements of Service* providing more detail on the units a man served in, his promotions and 'reductions', and extensive documentation of misconduct and punishment.
- *Active Casualty Forms* providing detail on wounds, sickness, transfers between hospitals, and deaths in service.
- *Death Notifications* for men who died in service, and for most men who died before 1990.
- *Ballots* indicating if a soldier was conscripted.
- *Miscellaneous papers* describing hospital stays and post-service pension assessments, and other correspondence.

Transcription of key documents in the personnel files could provide a critical resource for understanding New Zealand social history in the first half of the twentieth century. The files look both forward and back from WW1. Because men enlisted in early adulthood and were required to name next of kin, the files provide information on parents' background, international and internal migration, and marriage and family relationships. We estimate, conservatively, that on average at least three additional people are named in each file, meaning that around one-third of the New Zealand population alive in 1914 are mentioned in the records.[8] Looking forward, the WW1 records often extend to the end of men's lives, with evidence of death for people dying as late as the early 1990s.[9]

[8] Some of those named were dead (e.g., parents listed for identification purposes), and others were resident overseas. Because many people who served had siblings, not all names of others are unique.

[9] For further discussion of how the records can be linked to birth and death records, see Kris Inwood, Les Oxley and Evan Roberts, 'The mortality risk of being overweight in the twentieth century: Evidence from two cohorts of New Zealand men', *Explorations in Economic History* (2020), https://doi.org/10.1016/j.eeh.2022.101472.

The size of the WW1 personnel file collection makes it infeasible to apply for grant funding to complete transcription of even just the key documents in the files. With 140,000 files at an average data-entry time of 30 minutes for the key documents, around $2,100,000 would be required to fund complete transcription. Realising the significant research potential of a complete transcription, our research group took advantage of an opportunity to launch a crowd-sourcing effort to transcribe the files. With several papers published based on smaller samples from WW1 and related data, we saw immense opportunity in a larger-scale transcription to answer questions about health and mortality in New Zealand from the late nineteenth century onwards.[10]

DEVELOPING MEASURING THE ANZACS

Our opportunity to apply crowd-sourcing techniques to the transcription of the WW1 files was serendipitous. The New Zealand WW1 files became available at around the same time as a major citizen science organisation launched an initiative to extend their work into humanities research.

In October 2014, shortly after Archives New Zealand completed scanning all the files, the Zooniverse citizen science group made a call for proposals for handwriting transcription projects to launch new transcription-specific citizen science software.[11] The origins of Zooniverse lay in the accelerating volume of digital images from astronomical telescopes in the early twenty-first century quickly outgrowing researchers' capacity to identify and classify galaxies on the images they had collected. Computer recognition of the shapes of galaxies in these images was poor, but people could be taught to identify the different galaxy morphologies with a short tutorial, even with no prior astronomical training. In 2007 astronomers at Oxford launched the Galaxy Zoo project, asking people to simply identify whether a galaxy had the shape of a spiral or an ellipse. To overcome the potential for inaccurate or malicious classifications, images were presented randomly and every image was viewed by multiple people. Thus, the final classification of a galaxy's morphology was probabilistic. Trained researchers verified a selection of the volunteer-classified images to establish the accuracy of volunteers' work.[12]

A similar 'data deluge' was building in ecological research in the same era. Rapidly declining costs for digital cameras, and rapid increases in the storage capacity of digital media, meant that camera trap technology was deployed on an increasing

[10] Kris Inwood, Les Oxley and Evan Roberts, 'Physical Growth and Ethnic Inequality in New Zealand Prisons, 1840–1975', *The History of the Family*, vol. 20, no. 2 (2015), pp. 249–69; Kris Inwood and Evan Roberts, 'Longitudinal Studies of Human Growth and Health: A Review of Recent Historical Research', *Journal of Economic Surveys*, vol. 24, no. 5 (2010), pp. 801–40; Kris Inwood, Les Oxley and Evan Roberts, 'Physical Stature in Nineteenth Century New Zealand – a preliminary interpretation', *Australian Economic History Review*, vol. 50, no. 3 (2010), pp. 262–83; Evan Roberts and Pamela Wood, 'Birth Weight and Adult Health in Historical Perspective: Evidence from a New Zealand Cohort, 1907–1922', *Social Science & Medicine*, vol. 107, no. 4 (2014), pp. 154–61.

[11] The call for proposals can be seen at https://sites.google.com/a/umn.edu/zoomanities/

[12] Chris J. Lintott et al., 'Galaxy Zoo: Morphologies Derived From Visual Inspection of Galaxies From the Sloan Digital Sky Survey', *Monthly Notices of the Royal Astronomical Society*, vol. 389, no. 3 (2008), pp. 1179–89.

scale from around 2006.[13] Animals come in more shapes and sizes than galaxies, but the research challenge was isomorphic: ask volunteers to complete a short online tutorial, and then carry out simple, repetitive tasks to identify which animals were present in a given image. Recognising the similarity between these research problems across multiple fields, the Citizen Science Alliance launched Zooniverse in 2009 to support general purpose software for citizen science. The University of Oxford, Johns Hopkins University, the Adler Planetarium and (from 2010) the University of Minnesota were key institutional supporters of Zooniverse.

The crux of the Zooniverse approach to citizen science is threefold:

1. Volunteer contributions are broken down into the smallest feasible tasks that are useful
2. Every task is carried out by multiple volunteers
3. Experienced researchers verify the accuracy of a sample of volunteers' work.

Taken together, this approach means that casual web surfers can make useful contributions, the scope for malicious actors to harm the data is limited, and with attention to verification and validation, the resulting data is scientifically valid. The accuracy of volunteer contributions in Zooniverse projects is high. At the time we began developing the software for transcribing handwriting, colleagues in ecology found that volunteer classifications were 98 per cent accurate, although accuracy was lower for rare species.[14] In physics and ecology, it was common for classifications of each image to be made by between 5 and 20 volunteers.

The positive experience that physicists and ecologists had with citizen science prompted Zooniverse to explore how the approach could be extended to the humanities. At the time, in 2014, similar issues of managing and using a growing volume of digital images were accruing in the cultural sector. Many museums, for example, had photographed thousands of specimen cards or accession documents. Like the WW1 personnel files, these were electronically indexed on a couple of key fields, but the documents often contained significant additional information on the provenance of objects. Transcribing this material and providing the transcriptions along with images of the object had the potential to expand access to museum collections for researchers and interested citizens alike.

The Zooniverse approach to crowd-sourcing transcription differed slightly from the then-current state of the art in utilising volunteers. Many libraries placed scanned images from their collection online, and with simple web forms provided a place for volunteers to transcribe whole pages from letters and diaries. However, this approach often led to pages only being partially transcribed and staff often had to complete the transcriptions.

---

[13] Thomas E. Kucera and Reginald H. Barrett, 'A History of Camera Trapping', in Allan F. O'Connell, James D. Nichols and K. Ullas Karanth (eds), *Camera Traps in Animal Ecology*, Springer, Tokyo, 2011, pp. 9–26.
[14] Alexandra Swanson et al., 'A Generalized Approach for Producing, Quantifying, and Validating Citizen Science Data From Wildlife Images', *Conservation Biology*, vol. 30, no. 3 (2016), pp. 520–31.

Our proposal to work with Archives New Zealand and Auckland Museum on a Zooniverse approach to transcribing the New Zealand personnel files was selected in early 2015 as one of four initial humanities transcription projects for Zooniverse. These projects would also serve as test cases for a new 'software stack' being developed by Zooniverse to support transcription. The other projects in the initial group were transcribing weather logs from Arctic voyages, records from the Emigrant Savings Bank in New York, and natural history collection labels from the University of Minnesota's Bell Museum.

The software, called Scribe, was intended to deal with the slightly different ways in which we anticipated volunteers and researchers would want to interact with handwritten documents. In particular, we assumed that people would want to be able to view and transcribe sequential images related to the same subject. For physics and ecology projects, it had been appropriate to randomise the display of images. But handwritten documents often occur over multiple related images: pages of a diary, letter, logbook, ledger or personnel file. Incorporating this grouping of individual images into 'subject sets' of related images was a major change in the Zooniverse approach, and necessitated new software. A second issue that required modification to the existing software arose from the fact that while radio telescope images were typically of one galaxy, and many camera trap images had just a few animals, handwriting occurs in multiple words and lines. Following the Zooniverse approach, we wanted to break down transcription into appropriately sized tasks – a word would be too small, but a line or a box delineating the answer to one question might be just right. At the same time, pages are made up of multiple lines or graphically delineated answers. The software would need to be able to capture volunteers' identification of a varying number of lines or boxes on every page.

Through the first nine months of 2015, our research team met regularly with Zooniverse software developers as the Scribe software was modified from existing code. Archives New Zealand provided access to their entire collection of images, and Auckland Museum consulted with us on how we might reach a volunteer audience when we launched in October 2015. The generous collaborative support of these institutions was critical in allowing us and the other new Zooniverse transcription projects to focus, on developing usable, responsive software.

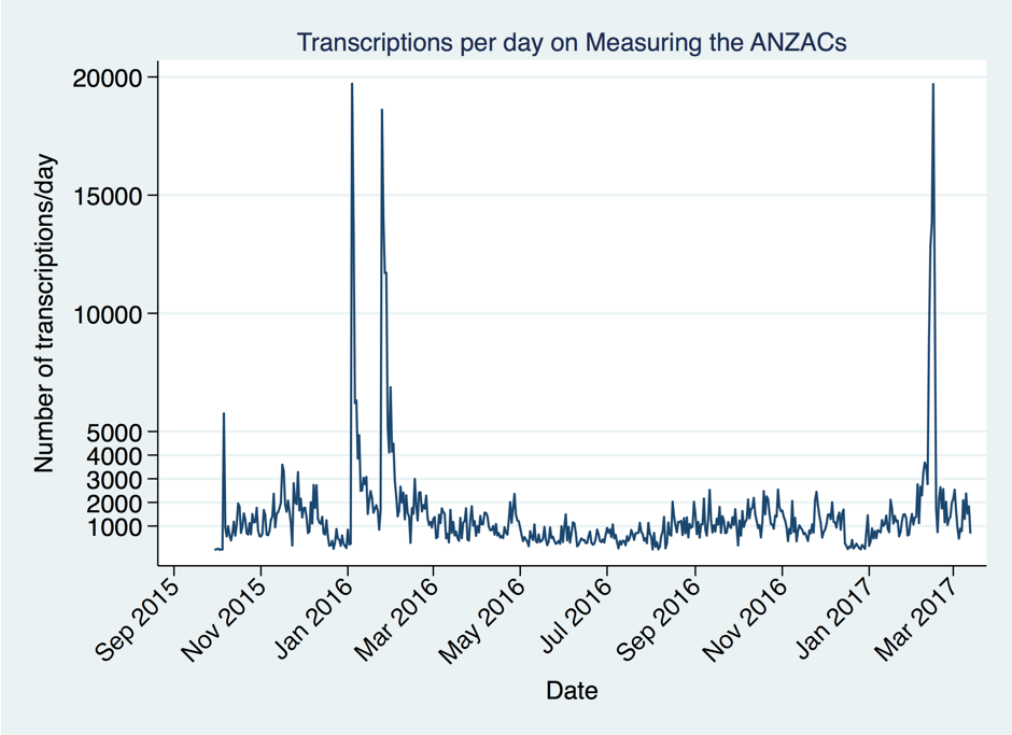
For the personnel files, our approach to designing the crowd-sourcing interface was heavily influenced by our experience teaching history students in high schools and universities. In the classroom it was natural to download a file from the Archives New Zealand site, and for students to work sequentially through the pages. It bears mentioning that the internal organisation of the personnel files reflects their origin as working documents. As researchers who have worked with government archives will know, the top document is often the last one that was 'touched' by clerical staff or contains summary information useful for filing or other disposition of the record. Pages that we were interested in, particularly the attestation at enlistment and death notifications, were often many pages into the file.

Building on the Zooniverse approach, we asked volunteers to first answer a simple question: 'What kind of page is this?' If they identified one of the five key document types, the software branched to allow the volunteer to enter information on the questions asked about each document. The most challenging document we were interested in was the attestation form. While the questions on the medical side of the form were relatively consistent throughout the war, our earlier research had suggested significant variation in the questions asked about social, demographic and occupational characteristics.[15] Using our earlier data, which contained enlistment dates, we asked an undergraduate research assistant to retrieve two samples of an attestation form from each month of the war. Over 51 months, we identified 31 distinct variations in question content and ordering. This bewildering variety made it difficult to design consistent software for recording the answers, and thus we had to ask volunteers to transcribe the question (e.g., 'Have you passed the Fourth Educational Standard or its Equivalent?') and then the answer. A beta test of the site in August 2015 was successful, requiring minimal software adjustments, and we prepared to launch in October 2015. Reviewing the data collected in the beta test, we determined that just three different transcriptions were likely to be necessary for accurate crowd-sourcing of handwriting.

LAUNCHING MEASURING THE ANZACS

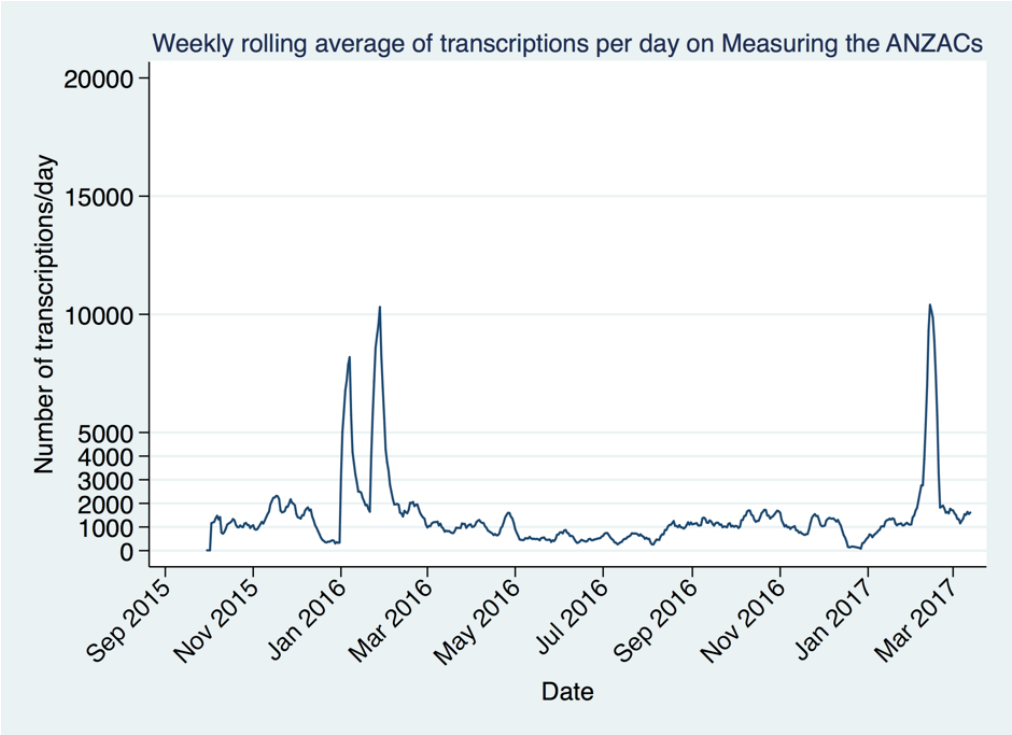
We launched the site to the public in October 2015. The Zooniverse volunteer list of more than a million volunteers meant that we had significant initial interest. On the first day, more than 5000 transcriptions were made. On the personnel files a transcription represents one answer to one question, as short as '6/6' for someone's eyesight, or as long as several hundred characters when people summarised the contents of a letter. On the basis of prior Zooniverse experience, we expected that only a small percentage of people who initially volunteered on the site would stick around and become long-term volunteers who would push us closer to the goal of completing the transcription of all key documents on all the files. Therefore, we put significant effort into volunteer engagement on discussion boards embedded on the the site. After completing a transcription, volunteers were able to click a button to generate a discussion board post about the image they had just transcribed. Volunteers could also visit the 'Talk' boards, and initiate discussion of other topics. Using material that volunteers uncovered, we posted regularly to social media to encourage interest in the site.

[15] Inwood, Oxley and Roberts, 'Physical Stature in Nineteenth Century New Zealand'.



Public attention to the project was boosted by media coverage on TV One and TV3 early in 2016, as can be seen in the graph of daily transcriptions. We had reached out to reporters upon launch in October 2015, anticipating that coverage of the site might be tied to news coverage on Armistice Day. However, in both cases the television reporters were planning other stories for 11 November. Print media in both New Zealand and Britain did give us coverage in November 2015, and we publicised the project at academic conferences in Europe and North America around the same time. Spikes in daily traffic around this time are dwarfed by the ephemeral boost from television news coverage in early 2016. The publicity we received on both major television channels in New Zealand was well-timed: reporters were able to place the story on relatively slow news days and we were given around three minutes' coverage on both channels. This coverage helped us regain traffic after a dip around the Christmas and New Year holiday. Interestingly, a long dip in site visits during the holidays had not been observed on the physics and ecology citizen science projects hosted by Zooniverse. The Australasian volunteers on Measuring the Anzacs project are probably also likely to take a break in summer. Not only was transcription different in the software required, but the patterns of user engagement were different. The threshold for making a useful contribution was higher, and the nature of the material meant that transcription of war records was less likely to be perceived as a relaxing diversion than classifying images of stars or animals.

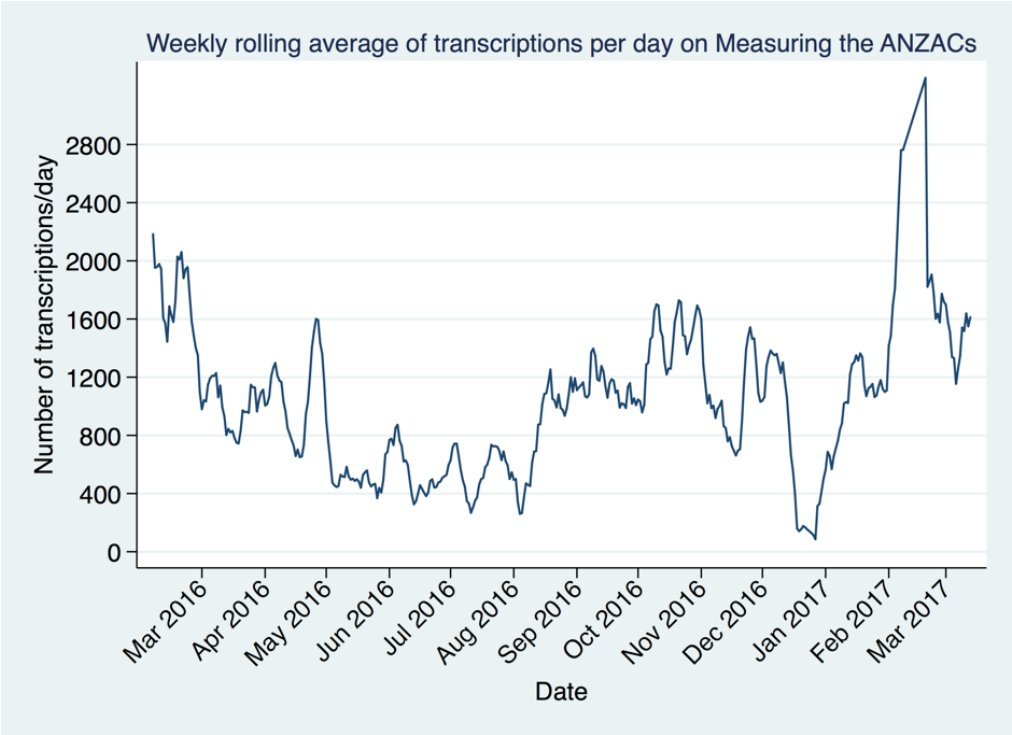
As we had anticipated, growing and sustaining the volunteer community for Measuring the Anzacs required in-person outreach to interested people. While we could not be sure *exactly* who our volunteers were in early 2016, email inquiries, interactions on the discussion boards, and the pattern of traffic over time suggested a sustained mix of genealogists, history buffs and students in universities and high schools in New Zealand, Australia and Britain.



Students are not truly volunteers.[16] They are in a very real sense, like soldiers, often conscripts. While they may gain something from their contributions, careful engagement with teachers is required to produce work that is mutually beneficial for both the project and the students. In fact, the most valuable student-centred assignments with crowd-sourced transcription place little value on the amount transcribed, and more on developing skills and interest. High school and university students today need to learn how to read cursive handwriting as a precondition for accurate transcription.[17] To support the development of these transcribing communities, we conducted several in-person workshops with university lecturers, high school teachers, and genealogists in New Zealand in July 2016. These workshops did not lead to massive one-day spikes in the number of 'answers' transcribed, as our television appearances had. But through several months in late 2016 the weekly amounts transcribed reached the highest sustained levels since

[16] Spencer Keralis, 'Disrupting Labor in Digital Humanities; or, The Classroom is Not Your Crowd', in Dorothy Kim and Jesse Stommel (eds), *Disrupting the Digital Humanities*, punctum books, 2018, pp. 273–94.
[17] Evan Roberts, 'Measuring the ANZACs: Exploring the Lives of World War I Soldiers in a Citizen Science Project', in Christopher J. Young et al. (eds), *Quick Hits for Teaching with Digital Humanities: Successful Strategies from Award-Winning Teachers*, Indiana University Press, Bloomington, 2020, pp. 262–70.

launch. As we had anticipated, our numbers took a seasonal dip around Christmas, before picking back up. We then received another significant, yet ephemeral, boost from coverage on the BBC World News website in January 2017. Just over a year after its launch, Measuring the Anzacs was the second most popular transcription site on Zooniverse, behind Shakespeare's World. Our daily number of transcriptions ranked us just outside the top 10 most active Zooniverse sites, out of nearly 100 different projects.



Unfortunately, these early successes in sustained engagement and a large volume of transcriptions being made contained the seeds of problems that would arise over the next year. Scribe had been developed as a collaboration between the Zooniverse and New York Public Library, but there was no ongoing funding for its maintenance or further development. Zooniverse software developers were prioritising the development of an online tool called the Project Builder that would allow people to build their own websites without needing to interact with software developers.[18] Until that point, every Zooniverse site had been custom-coded. However, significant abstract similarities between research problems in different fields meant that each successive customisation was a little simpler as code was reused by different projects. The large size of Measuring the Anzacs ' image collection and the number of transcriptions that had been contributed also presented problems for the Scribe software: we had launched with 10,471 people's files on the site. Building on the successful approach in other Zooniverse projects, Scribe was built to present volunteers with a random image that they had not seen

[18] https://www.zooniverse.org/lab

before, and which needed attention. Simplifying somewhat, a subject set was deemed to need attention in one of three situations:

1. It had not been transcribed
2. Volunteers had indicated the presence of text needing to be transcribed, but there was no transcription
3. Verification by additional transcribers was needed.

These conditions required the software to check a user's identity when they logged on, and then find the best image to present to them to help advance the goal of multiple transcriptions of every page. As the number of transcriptions grew, the site became increasingly unresponsive at this task. Teachers planning to use the site in class found the site difficult to work on unless the number of students using it at any one time was limited, contradicting the ideal of being able to have the class work simultaneously on related material.

Beyond these challenges, the Zooniverse leadership had determined that while Scribe was innovative and fit for purpose for smaller image collections, it struggled with the approximately 250,000 images we had initially loaded onto the site. The software being developed for the Project Builder was incorporating transcription tasks, and had demonstrated that it was capable of handling significantly higher numbers of simultaneous volunteers on the site. Faced with these challenges from mid-2017, we wound back our public engagement aimed at encouraging greater use of the site, while continuing to support and engage with the small number of volunteers who persisted in transcribing.

These software challenges were particularly disappointing, given the accuracy of the transcriptions we were collecting. The accuracy of volunteer transcriptions exceeded our expectations and was comparable to the accuracy seen in physics and ecology projects. Measuring accuracy in transcription involves comparison to a 'gold standard', which conveniently we had access to. The metadata provided by Archives New Zealand contained both the names and the serial numbers of the soldiers, and both these items were transcribed by volunteers in several places on each form. Although each page image was identifiably from a named file, we decided to ask volunteers to transcribe name and serial number information as a measure of transcriber accuracy, and to identify pages that had been placed in the wrong file.[19]

While less accurate than paid research assistants, the Zooniverse volunteers were very accurate. The share of volunteer entries for names and serial numbers that exactly matched the gold standard was lower than among two sets of research assistants our research team had employed between 2007 and 2010 (Table 1). However, when we look at useable accuracy by scoring the discrepancy between volunteer transcriptions and the gold standard, volunteer performance begins to look comparable to the work done by research assistants. Using the Jaro-Winkler

[19] Over thousands of pages, we have identified fewer than 10 misfiled pages, about which we have alerted Archives New Zealand.

'string comparator' to generate a similarity score, we find that Zooniverse volunteers and research assistants transcribe nearly the same proportion of records with a similarity score of > 0.9.[20] To give an example, if the correct name is 'Charles' and the volunteer types 'Cjarles', the similarity score is 0.91. The Jaro-Winkler similarity score penalises mistakes in the middle of the word less than mistakes in the first and last letters. On average, Zooniverse volunteers who made mistakes mistranscribed one or two characters, generally in the middle of the word.

	Attestations (NZ RAs)		**Casualty Rolls (Canadian RAs)**		**Zooniverse Volunteers**		
	Given	Last	Given	Last	Given	Last	Serial Number
Mean similarity score to truth (1 = absolute accuracy)	0.98	0.99	0.99	0.99	**0.98**	**0.98**	**0.98**
Proportion absolutely accurate	0.91	0.95	0.93	0.98	0.87	0.84	0.89
Proportion with similarity score > 0.9	0.96	0.98	0.95	0.99	0.95	0.95	0.95
Proportion with similarity score > 0.8	0.98	0.99	0.99	0.99	0.97	0.98	0.98
Proportion with > 3 words in string (more likely problematic transcriptions)	0.0030	0.0002	0.030	0.00	0.0029	0.0030	0.0030

Table 1. Accuracy of research assistants and volunteers

These results give us significant confidence that crowd-sourcing can be an efficient and accurate way to complete the transcription of all New Zealand's WW1 and South African War personnel files. In time, the same approach can be applied to World War Two. With the software challenges we encountered and the site limping along, we invested significant time in 2019 in rebuilding the site on the new Zooniverse software stack using the Project Builder. The Project Builder interface allowed us to build separate 'workflows' for the transcription of each of the key documents in the personnel files. The Statement of Services and History Sheet both have several elements that repeat a variable and unknown number of times. For example, people can be wounded or sick many times. Similarly, they can transfer units or be promoted a finite but unknown number of times. These varying unknowns require the volunteer to loop through the same questions about each instance of a sickness or promotion event. The Project Builder was capable of building workflows that could handle these complexities.

Learning from our experience in 2016, we decided that the most sustainable approach to re-engaging the New Zealand history and genealogical community was not to seek ephemeral media publicity, but to engage deeply with an interested community in person. When we formally relaunched the project in October 2019, we kept launch activities to a minimum and began developing pedagogical and

[20] William E. Winkler, 'Approximate String Comparator Search Strategies for Very Large Administrative Lists', Bureau of the Census, 2005.

family history materials with a view to sharing these with New Zealand groups in 2020.

If you are reading this article in the early 2020s, you can probably infer how these best-laid plans went awry! If you are reading it further into the future, it suffices to say that the disruptions to international travel, education and in-person gatherings brought on by the Covid-19 pandemic have made it impossible for our research team based abroad to run workshops in New Zealand. At the time of writing, we are planning to conduct workshops in late 2022 and mid-2023.

The goal of Measuring the Anzacs remains the same audacious but achievable one we had when we launched: to work with New Zealanders and other interested volunteers to complete a transcription of key documents from 140,000 personnel files. The final dataset will be made available as a searchable and download public dataset to support historical research of all kinds. With each personnel file taking around half an hour to completely transcribe, and three entries for each file required in our crowd-sourcing protocol, we need around 210,000 hours of volunteer work. In the first 18 months after Measuring the Anzacs' launch, volunteers contributed nearly 43,000 hours. With significant improvements to the software that limit the potential for duplicative effort, and a hosting platform that has proved robust under significant volunteer loads, the project is now ready to re-engage with the community and find a path to completing a significant public history resource for New Zealand.

The length of the path to completing the transcriptions can be expressed in several comprehensible, human terms, If every high school student in New Zealand and one of their parents transcribed one soldier each, the threefold transcription of all 140,000 files would be complete in a year.[21] Towards the other end of the age spectrum – where comprehension of cursive handwriting is likely better – if every New Zealander reaching retirement age transcribed the files of two soldiers, the transcription would be complete in a couple of years.[22] Perhaps more achievable would be around 1,500–2,000 people making a commitment to transcribe the file of one soldier every weekday for about a year. Put in any of these terms, the power of citizen science becomes comprehensible. Many hands can make audacious projects achievable.

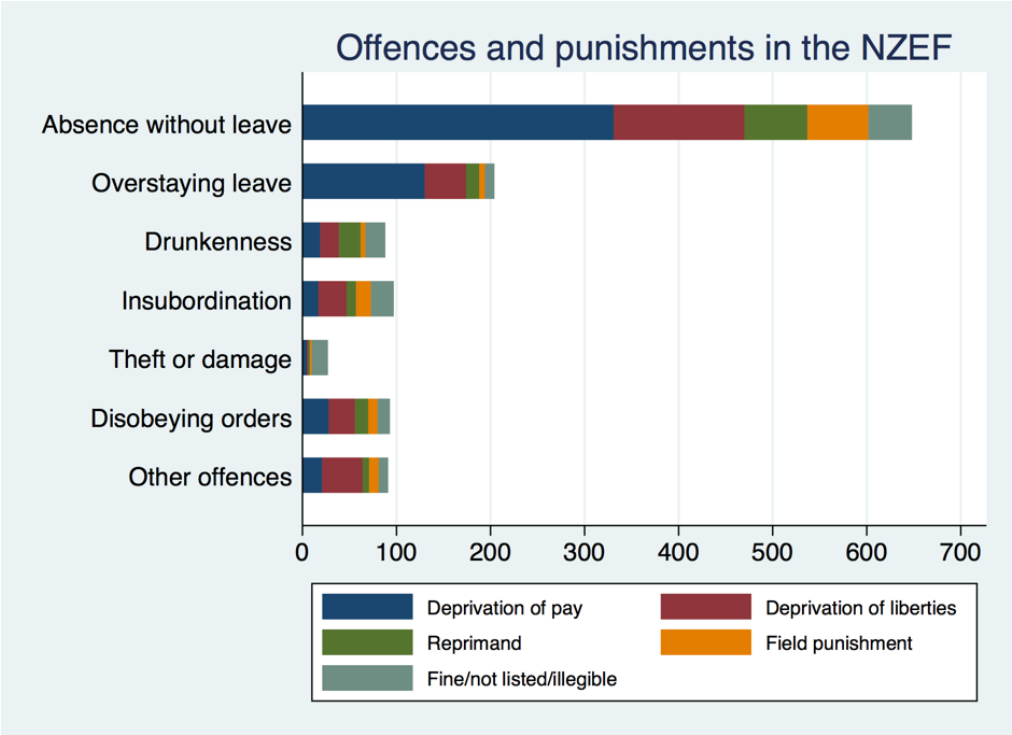
In writing our proposal for the project, we expressed the hope that transcription would uncover hidden stories that were otherwise inaccessible. Indexing elements of the personnel files that had not previously been explored could open new research areas. An example of this potential is the topic of misconduct, which is somewhat unexplored in military and social history, and rarely touched on by modern scholars.[23] Working with an extract of some of our earliest data, we

[21] There were 282,000 high school students in New Zealand in 2022: https://www.educationcounts.govt.nz/statistics/6028

[22] Estimated from population figures at https://www.stats.govt.nz/information-releases/national-population-estimates-at-30-june-2022/

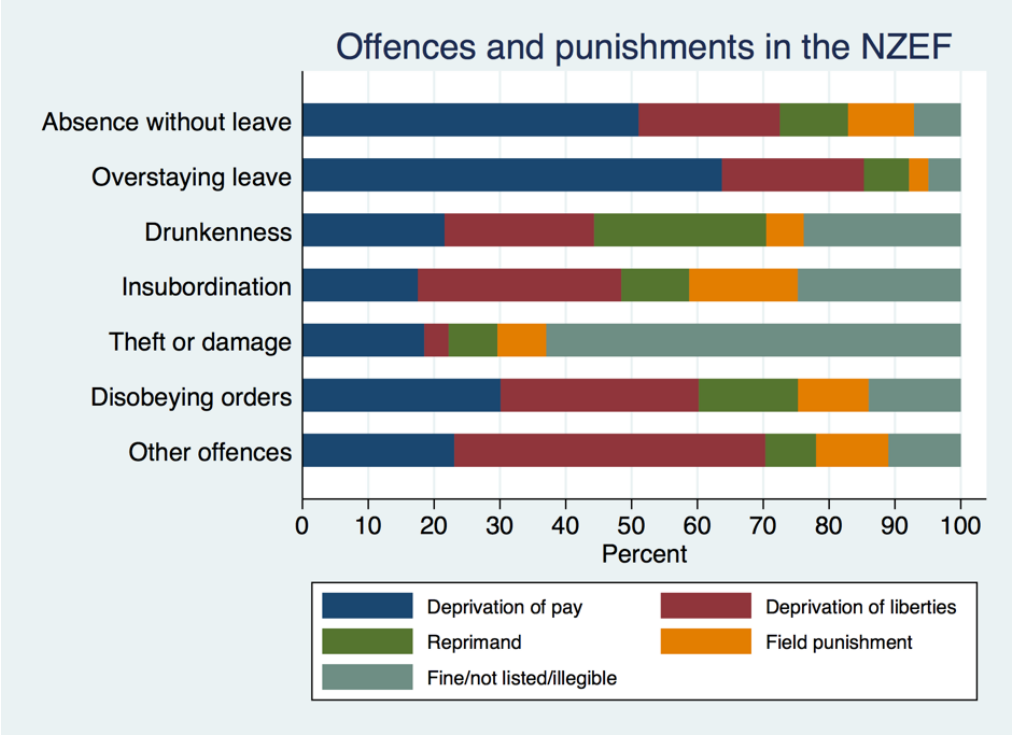
[23] Stephanie Booth-Kewley et al., 'Psychosocial Predictors of Military Misconduct', *Journal of Nervous and Mental Disease*, vol. 198, no. 2 (2010), pp. 91–8.

identified that 21 per cent of men had a misconduct citation listed. In short, misconduct was relatively common, perhaps complicating narratives of heroism and duty that are sometimes attached to returned soldiers by the public and politicians.



We then examined the punishments that were meted out for different offences. The typical punishment for leave offences were deprivation of pay – a direct and related consequence. Property damage was typically punished with a fine. Insubordination and disobeying orders were most often punished with being confined to barracks, or other restrictions on movement. This initial examination suggests that punishments were often connected to the offence, and calibrated to minimise the impact on the fighting force. But the personnel files can tell us much more, if we can complete the transcription and see how misconduct contributed to different promotion or demotion paths. Even in the most serious instances of misconduct we reviewed, misconduct did not lead to dismissal. For example, one man convicted of sleeping on his post was sentenced to five years penal servitude, to be served after the war. Sadly, he died of wounds less than a year later.[24] We observed numerous files of men who rose to become officers despite several misconduct incidents.

[24] Personnel file of Hugh Christian Hansen (WWI 17/79). Archives NZ Item Code R16791295. Available: https://ndhadeliver.natlib.govt.nz/delivery/DeliveryManagerServlet?dps_pid=IE20074674



A complete transcription of the personnel files will allow New Zealand historians to tell more complete stories of what happened in WW1, and connect the men and women who served with hundreds of thousands of others alive at the time. The experience of Measuring the Anzacs demonstrates some of the challenges of crowd-sourcing on a large scale. The initial rush of volunteer enthusiasm in the first 18 months revealed problems in the software that took significant effort and time to resolve. Having prioritized the development of a system that allowed people to browse complete personnel files and experience the narrative of people's lives as if in the archive, we were surprised that this software feature contributed to some of our ongoing data challenges. With the project now relaunched on a stable software platform, our next challenge is finding a community of engaged volunteers, and maybe some willing conscripts, to bring the project to completion within the next few years – before we tackle World War Two.

Evan Roberts is an Assistant Professor in the History of Medicine and Population Studies programs at the University of Minnesota, with research interests in health and work since the nineteenth century in New Zealand and the United States. Evan grew up in Wellington, and graduated from Victoria University with a BA(Hons)/BSc before completing a PhD in History at the University of Minnesota in 2007.

.