IFLA WLIC 2013
SINGAPORE

Submitted on: July 31, 2013

'Singing for their supper': Trove, Australian newspapers, and the crowd

Marie-Louise Ayres
Resource Sharing Division, National Library of Australia, Canberra, Australia
E-mail address: mayres@nla.gov.au

Abstract:

Building on 20 years of national cooperation under the Australian Newspapers Plan (ANPlan), the National Library of Australia has delivered digitised newspapers to the Australian public since 2007, and incorporated newspaper delivery into Trove (www.trove.nla.gov.au) in 2009. Trove has won a number of awards for its user engagement features, most notably the ability for users to correct the computer-generated text imperfectly generated by Optical Character Recognition software. By July 2013, use of Trove dwarfed use of the Library's other online services, the Trove Newspapers zone dominated use of Trove, and Australians had corrected more than 100 million lines of text - the equivalent of 270 standard work years of crowd-sourced effort. With five years' experience, this form of user engagement is a mature part of the Library's service offering, and evaluation of its impacts and future is timely. The paper will summarise what we know about the motivations of those who engage in this activity, their patterns of engagement, how 'deep' the penetration of this form of engagement with the Library extends, and what current statistics suggest about where we are in the user engagement growth curve. The paper also considers the significant impact the success of crowd-sourcing has on the Library's rationale for offering this service (are we offering the service to address a huge task, or to build and maintain a broader community of Library supporters?), the opportunities and risks in having dramatically increased the number of Australians passionately engaged with the Library, and the ways in which success is changing the Library's thinking about service delivery and engagement with the public. For those considering or in the early days of offering a digitised newspaper service, the paper provides a view 'from the trenches' about the ways in which success can fundamentally reshape the questions libraries must ask themselves.

Keywords: Newspapers, digitisation, crowd-sourcing, engagement, community.

1 INTRODUCTION

The National Library of Australia's Trove (http://trove.nla.gov.au) has been recognised as an international leader in facilitating public access to documentary heritage. While Trove provides access to much more than digitised newspapers[i], it is the newspaper component of the service that has captured public attention, and that consistently accounts for more than three-quarters of all user visits. Similarly, while Trove offers a range of user engagement features, and use of each of these features continues to grow, it is Trove's newspaper text

correction features that have attracted the highest level of user engagement. Australians and members of the international public have now been correcting newspapers for more than 5 years. Library staff and academic researchers have gone beyond marvelling at the popularity of the service to thinking more deeply about what is actually occurring in the text correction landscape. Recent research has confirmed many of our 'hunches' about what this popularity means, confounded other expectations, and made us quite thoughtful about the future of text correction in particular, and where we should head more generally in relation to engaging with our online users.

2 BACKGROUND

The stand-alone digitised Australian newspapers service was launched in 2007 and integrated into Trove in 2009. However, the antecedents of the project extend back much earlier. Australia has a long history of national cooperation in library matters. A case in point is the Australian Newspapers Plan (ANPLAN), which was established in 1992.[ii] The Plan established a firm cooperative and collaborative basis for collection, management, preservation and access to Australian newspapers, and paved the way for a national approach to newspaper digitisation. In 2005, the Library was a partner in an application for Australian Research Council funding to digitise Australian newspapers to support Australian humanities and social sciences research[iii]. That application – which included some early thinking about the potential for university researchers to add value to digitised content – was unsuccessful, but proved useful when the Library shortly afterwards decided to commence a newspaper digitisation pilot, and to focus on the broadest possible audience in terms of user engagement.

These early thoughts about engaging users did not themselves arise in a vacuum. The Library was an early leader in collection digitisation and in the development of national discovery services for documentary heritage, with Picture Australia, Music Australia, the Register of Australian Archives and Manuscripts, Australia Dancing, Australian Research Online and the PANDORA web archive all now integrated into Trove. These services – built between 1997 and 2008 – all preceded Web 2.0 and the kinds of social media features we now take for granted. However, some did incorporate options for user engagement. Picture Australia established a relationship with Flickr as early as 2006, enabling members of the public to add their own photographs of Australian life to a designated Flickr pool which was then harvested into Picture Australia. In the same year, Australia Dancing incorporated a new 'Take part' service component, which invited registered users to contribute their own knowledge to the service. The former was very successful (more than 180,000 images have been uploaded to the Trove and predecessor pools and are available through Trove). While the Library had high hopes for the latter – believing that such a service would be taken up by the small but very dedicated dance research community – the promise of 'Take part' was not realised.

The Australian newspaper service was therefore launched in an environment in which the Library had long been thinking about ways to engage with users online, had experimented with several modes, and had experienced some successes and some failures. The newspaper correction features incorporated into the digitised newspapers service reflected this desire to engage with our users, as well as recognising that a crowd-sourced approach to correcting imperfect OCR'd newspaper text had the potential to improve search results for Trove searchers, and that the manual work required for this work could never be resourced within the Library's funding base.

3 TROVE USER ENGAGEMENT

Five years after Trove's release to the public, its success is a source of pride and delight for the National Library, and the raw statistical numbers are impressive:

- Trove accounts for more than 62% of all National Library of Australia pageviews, dwarfing use of the Library's own catalogue, website, and the Libraries Australia service, which is used by more than 1000 libraries;
- around 75% of all Trove visitors arrive at the site after finding a resource in Google, highlighting the importance of exposing Trove content to large search engines;
- an average of more than 60,000 unique users visit Trove every day, with spikes of up to 80,000 on a single day; and
- total visits have doubled in the last two years, with 1.8 million visits in June 2013.

This means that in just a few years, Trove has become the most common means by which Australians (and the world) encounter the National Library of Australia, and that the vast majority of those users arrive via a search engine, rather than by beginning their navigation at either Trove, of the National Library of Australia website.

Many of Trove's user engagement features are very popular. More than 100,000 users have registered to date, and more than 2 million tags and nearly 60,000 comments had been added to Trove resources. The fastest growing user engagement feature is Trove lists, which allows users to curate a set of Trove and other web resources into lists than can be made public or private: nearly 40,000 of these lists have been created on a wide variety of topics.

Text correction, however, stands head and shoulders above any other user engagement features. By July 2013, more than 100 million lines of newspaper text had been corrected by members of the public. We have estimated that this equates to more than 425,000 volunteer hours, or 270 standard Australian work years. Costed at the Library's lowest pay rate, this means that text correctors have contributed more than AU$17 million in value to the service[iv]. These are heady figures and it is interesting that despite Australia's success in this area, few digitised newspaper services offer users the ability to improve the quality of OCR'd newspaper text for the benefit of all.[v]

Statistics like these inevitably raise many questions, and in the last year or so, members of Library staff and independent researchers have been delving below those numbers to try to understand more about our users, especially those engaging in text correction activity. This research has been in three broad areas. Library staff have used Google analytics and Trove logs to try to understand more about our users and their patterns of engagement. The Library has also commissioned an independent evaluation of Trove user customer satisfaction, which reveals more Trove user demographics, what they use the service for, what they value in Trove, what they like and what they would like to see changed[vi]. Meanwhile, academic researchers are considering what motivates users to correct text, and are teasing out some of the 'meanings' behind crowdsourcing engagement with Trove content.

4 ENGAGED USERS: WHO ARE THEY?

While many social media sites gather vast quantities of information about their users, Trove does not. Registration for Trove users is optional, and only a minority of users register. Even when users register, we capture minimal – and absolutely no demographic – information

about them. While we can track registered user use of our engagement functions, we do not build 'profiles' of individual users. We do not know their names, age, gender, marital status, employment status, level of education, income bracket, address etc. We do not track or deduce what individual users are interested in (either through their use of Trove or by analysing sites they come from or go to from Trove), and therefore do not 'push' any suggestions to them, or offer any 'other users like you found this helpful' type services. The issue of whether - in the world of 'push' rather than 'search' - we should consider doing so, is a question for another day, and one which raises many ethical and professional dilemmas.

This does not mean we have no information about our user base. The small percentage of users who use our Contact Us help service[vii] – and therefore could be said to be engaged enough to seek personal assistance – can elect to give us some information about themselves: they can select from a list of user types, tell us their postcode, and tell us their country. A recent evaluation of Trove user satisfaction gave us our clearest picture yet of our user base. Putting these sources together, we have been able to conclude that:

- around 40% of all Trove visits are from international users (even when crawlers and bots are excluded). Most of these users are from the United Kingdom, United States and New Zealand. Tiny proportions of international users come from a number of Asian countries;
- Trove has achieved a truly national reach, with the proportion of users from each Australian state and territory, and the proportion of users in metropolitan, regional and rural parts of Australia aligning strongly with actual population distribution;
- 70% of Trove users are female;
- 65% of users are aged 50 or over; 34% are aged 60 or over; only 17% of users were aged under 40;
- 60% of users are employed; another quarter are retired, aligning strongly with Australian workforce participation;
- almost half of Trove users earn more than AU$40,000 p.a. compared to 28% of the general population; 34% earn more than AU$60,000;
- only 1% of Trove survey respondents are Indigenous (compared with 2% of the population), and only 4% speak languages other than English in their homes (compared with 15% of the general population);
- 45% of Trove users have or are completing a postgraduate qualification (compared to 2-3% of the general population); 85% had some form of tertiary qualification;
- almost half of Trove users consider family or local history as their primary reason for using Trove;
- 40% of Trove users say they use the service at least weekly, with a further 19% saying they use the site on a daily basis; and
- Trove users who correct text are more likely to be family historians, retired, and long-term (more than 1 year) Trove users.

Together, this means that the 'typical' Trove user is a very well educated, highly paid, English speaking employed woman aged fifty or over, with a significant or primary interest in family or local history, who visits the Trove website very frequently. Users of Trove newspapers are older than the average Trove user; only 13% of newspaper users are under 40 years or age. Correctors are also older, and are long-term, frequent users of the service.

Surveys recently conducted by Frederick Zarndt, Brian Geiger and Alyssa Pacey found that, users of the Cambridge Public Library and Californian Digital Library newspaper services shared a number of features[viii]. The majority of users identified themselves as genealogists, and a large majority use the site for family history purposes (85% for Cambridge; 61% for California). Around 40% of users asserted that they visit the sites at least weekly, with visits typically 60-70 minutes in length. On average, 77% of all users of these services are aged 50 and over, and 45% are aged 60 and over.

The 'typical' users of these three services look remarkably similar (although Trove's content and services are broader than the examples cited). Of course, 'typical' or 'average' obscure significant variations in user characteristics, reasons for using the service, which parts of the service are used most, whether the user feels 'connected' or not, and what the user would like to see change or develop in the future. While at least half of Trove users use the service for family history, 15% are undergraduate or postgraduate students or faculty pursuing academic research; these users do not feel so 'connected' to the Trove community, but frequently describe Trove as 'revolutionising' their research. A significant proportion of Trove users are actually librarians conducting searches to assist their clients. The Library's challenge for the future is to try to hit as many 'sweet spots' as possible; content and services that appeal to a broad range of user and usage types.

However, if we focus in on the text correcting community, Australian and international research seems clear: the largest audience for digitised newspapers is older users whose primary focus is on family or local history. Trove's text correction features appeal to older, and often retired, audiences primarily focused on family and local history.

While the National Library and other digitised newspaper providers can feel justifiably proud of the great leap forward their programs represent, and the Library is certainly delighted to have reached so many Australians scattered so widely across the continent, there are some uncomfortable questions for the Library, and questions that other libraries contemplating such a program may wish to consider.

Users inevitably age. Is it likely that we are currently seeing a 'wave' of genealogical interest that will wane as this generation ages? Or is it likely that genealogical research will continue to be a preoccupation for the over 50s, and that as the general population reaches that age, they will become more avid users of digitised newspaper services? Are there any scenarios we can foresee in which there will be substantial increases in use of these digitised newspaper services by those of school and university age, or those in early to mid-life stages? Are we content to appeal primarily to an older, well-educated, well-off audience? Do we need to accept that the kinds of content libraries currently provide, and the services we make available online are most attractive to that demographic – or do we need to think about what kinds of content, and what kinds of user engagement features would extend our reach to new audiences, to be truly socially inclusive?

5 WHAT MOTIVATES USERS

One of the ways to think about this issue is to consider what motivates our current users; in this context, what motivates our text correctors. Sultana Lubna Alam and John Campbell, both academics at the University of Canberra, used the Trove text correcting community as a case study for their investigation of crowdsourcing motivations in a study conducted in 2011 and 2012[ix]. They created a text correction motivation model (based on a number of models

for crowdsourcing motivation in a range of spheres) encompassing intrinsic and extrinsic motivations. Intrinsic motivations were classed as egoism-based (e.g. personal research interest), community-based (e.g. altruism and collectivism), and enjoyment-based (e.g. fun, enjoyable, pleasurable). Extrinsic motivations were primarily identified as social motivations (e.g. recognition, rewards, attribution and ownership). Trove text correctors were found to be primarily intrinsically motivated. Recognition and rewards – present in the Trove community but relatively muted – were not primary motivators. Instead, Trove text correctors are primarily motivated by their personal research interests, by the sense of being involved in something 'bigger than them' and 'of lasting value', and by a very strong sense of giving back or 'singing for their supper'. They enjoy their autonomy, find the correction task relaxing and enjoyable, have a strong sense of trust both in and by the Library, and feel that their work is valued. Alam and Campbell note that longitudinal studies of crowd-sourced workers will provide insights into motivational dynamism. They have since conducted further research into governance of crowdsourcing communities, and the motivations of institutions – such as the National Library of Australia – which engage in these forms of engagement[x].

6 PATTERNS OF USER ENGAGEMENT

Paul Hagon, a Senior Web Designer at the National Library of Australia, has used Trove transaction logs to investigate patterns of user engagement[xi]. Hagon used Google Analytics to show that Trove has a large number of repeat visitors, and that average visits last nearly nine minutes, compared to three minutes for the Library's catalogue and one for the Library's website. Australian visitors stay even longer, averaging twelve to fourteen minutes and viewing fourteen pages per visit. Trove's revisit rate and the length of average visits are very high by website industry standards, demonstrating that Trove content is very attractive to Australian visitors. This gives us a very broad impression of user engagement with the service, but is less helpful in terms of understanding what patterns there may be around our user engagement features, specifically text correction. To glean this information, Hagon used Trove's extensive transaction logs.

He established that text corrections by registered users comprise around 85% of corrections, with the remaining 15% of corrections by unregistered users, or perhaps by registered users correcting without having signed in. Patterns of engagement can only be established for registered users. The high correlation between user registration and text correction seems to indicate that text correctors feel they 'belong' to the service, that they have forged a relationship with the Library and the service. The high correlation also means that we can have a reasonable degree of certainty that the activities of our registered users act as a good proxy for user engagement patterns across the board.

Hagon's research established that a very small number of people – the top 100 – have undertaken 43% of all corrections. This means that the top 100 correctors have, between them, corrected more than *41 million* lines of text. A larger group of people – the top 1000 – have made 81% of all corrections, and 96% of all corrections have been made by the top 5000 registered correctors.

50% of all correctors have corrected less than 100 lines (at the Library's estimate of around 15 seconds per line corrected, this means half of all correctors have spent a total of less than half an hour correcting text) and 75% of users have corrected less than 500 lines (somewhere around 2 hours of their time). Our super-correctors – those who have corrected more than 1

million lines of text – account for just 0.01% of our users. Their contributions of more than 4000 hours apiece place them in a separate 'class' of engagement. For example, John Warren, Trove's top text corrector, was recently profiled in the Friends of the National Library of Australia's *Newsletter*[xii]. Mr Warren is a retiree who now spends between 6 and 8 hours a day correcting text, and regards this as his work. He has equipped himself with a large screen to facilitate zooming, and while Mr Warren focuses his work on items relating to his family history and the communities in which his forbears lived, he always corrects the full text of any article he touches.

The results of this research were quite surprising, and are certainly provoking a rethink about our user engagement strategy. Although we knew that some correctors were much more committed than others, we did not have a sound understanding of the relatively small number of people responsible for the majority of corrections. To put this into perspective, the National Library of Australia has around 80 onsite volunteers who work either front of house (especially as exhibition guides) or behind the scenes in collection areas. The 'super-correctors' are therefore not much more numerous than the onsite volunteer group. Similarly, the 1000 people who have made more than 80% of corrections is roughly half the number of Friends of the National Library. The 5000 is not very different to the 3,163 reading the Library's *Magazine*. In some ways, thinking about groups of this size is much more manageable than huge figures such as 100 million lines of text corrected, or 100,000 registered users. We care about all of our users – no matter how frequent or how intensive – but it may be that recognising that we are dealing with a more 'human' number of intensely engaged users may free us to think about new ways in which we can interact with them.

7 WHERE NEXT FOR TEXT CORRECTION?

We have also recently recognised that while the Trove newspaper content, visits to the Trove website and text corrections have continued to grow, the annual rate of growth in correction over the last two years has not kept pace with content growth. This suggests to us that we may have hit a point of critical mass, a point at which we are unlikely to gather many more really motivated text correctors, and that the 'market' for Australians willing to volunteer their precious time to this endeavour – at least with our current text correction functionality – may essentially be saturated. Does this matter? If there were no OCR improvement software solutions on the horizon, it would certainly matter, as an ever-growing body of content would not be optimised for searching, and we would certainly have to develop new ways of engaging with a broader audience to achieve the same aims.[xiii] But if we can apply software solutions to solve the original problem of less than perfect Optical Character Recognition (and it seems likely this will be possible within the next five years), will it matter if we no longer have this work for our dedicated digital volunteers to do?

8 WHAT IF WE TURNED THE QUESTION AROUND?

I argue that flat-lining of user engagement does matter if you are the national library of a nation with a small population scattered over a very large landmass, trying to reach out to and engage with distance communities.

The Library's original reasons for offering the newspaper text correction facility included our understanding that we could potentially harness the crowd to improve the quality of OCR'd newspaper search results, and our desire to engage with users in new ways in the online environment. As a result we now have a small number of Australians extremely highly engaged with the National Library's work, a larger number highly engaged, and a very high

number who are engaged enough to contribute smaller quantities of their time for the public good through Trove. Many of these people had no prior connection with the National Library, and our recent user evaluation suggests they also had weak connections with their local libraries, at least for regional and rural users. They are effectively a new audience – albeit looking demographically very much like our more traditional onsite audience – and represent a wider support base for the National Library, and for all the libraries and organisations contributing to Trove. I believe that maintaining our current users' engagement and growing the engaged user audience are now primary goals for the Library.

If we can use software to improve our newspaper text, what value-adding user engagement features might tempt our huge and growing audience of older and retired Australians learning about their family and community histories to give us and their fellow Australians some of their time? Transcription for digitised manuscripts material, similar to the National Archives of Australia's Hive[xiv] service? More value adding of the kind undertaking by special interest groups around topics as diverse as early Australian climate history and Australian South Sea Islander heritage? The ability to add geo-tags to large numbers of resources in a fun, easy and relaxing way? Opportunities for older Australians to curate their own online exhibitions so that the stories they find can be presented in interesting and visually attractive ways? All of these are on the National Library's short, medium and long-term drawing boards, and a number are already being implemented by our sister services, Europeana, the Digital Public Library of America, and Digital NZ.

But what of the audiences we are not reaching – the young, the less affluent, the less well educated, Indigenous Australians, and the large proportion of the Australian population for whom English is not the primary language spoken in the home? Given that the current Trove audience looks remarkably like the current Australian library profession, [xv]reaching these potential users will be even more challenging. This is, however, an aspiration that – as a profession stewarding documentary heritage and charged with making that heritage available for all who wish to learn and enjoy – we must strive to achieve.

---

[i] Trove provides access to more than 350 million resources, of which one-third are freely available online, one-third are licensed resources available to patrons of libraries subscribing to the resource, and one-third require additional steps for access. Trove provides access to books, journals, research outputs, pictures, manuscripts, maps, sound recordings, moving image, archived websites and realia.

[ii] Pamela Gatenby, Assistant Director-General, Collections Management at the National Library of Australia to December 2012 presented a paper on the history of ANPLAN to the 2008 IFLA conference, also held in Singapore. Her paper is available on the Library's website: http://www.nla.gov.au/content/the-australian-newspaper-plan-anplan

[iii] Slightly before the United Kingdom's JISC newspaper digitisation program, which was undertaken as a public-private partnership.

[iv] Other ways of estimating the value of text correction are available. Frederick Zarndt and Brian Geiger used a hypothesised outsourcing cost of 50 cents per 1000 characters corrected to arrive at a figure of $1.4 million for Trove's then 70 million lines of corrected text in their excellent study of newspaper text correction, presented at the November 2012 Digital Library Federation Forum, and available at: http://www.diglib.org/forums/2012forum/no-tempest-in-my-teapot-analysis-of-crowdsourced-data-and-user-experiences-at-the-california-digital-newspaper-collection/

[v] As discussed recently by Rose Holley, former Trove Manager and now working on a similar project at the National Archives of Australia. See: http://rose-holley.blogspot.com.au/2013/04/crowdsourcing-text-correction-and.html

[vi] Conducted by Gundabluey Pty Ltd in May and June 2013. The Library expects to complete full analysis of the results by late 2013. The survey collected extensive qualitative responses from 28 users, and detailed quantitative responses from 1086 users. The size of the sample was sufficient to give the Library a significant level of confidence in the result.

[vii] The Trove team responds to around 3000 enquiries per annum. In addition, more than 500 suggestions for newspaper titles which should be digitised, or questions about which titles are in the pipeline, are referred to the newspaper digitisation team.

[viii] Their results were presented the June/July 2013 conference of the American Library Association. Slides are available at: http://www.slideshare.net/cowboyMontana/what-motivates-library-crowdsourcing-volunteers-20130630-ala-lita

[ix] Alam, Sultana Lubna and Campbell, John, 'Crowdsourcing Motivations in a not-for-profit GLAM context: the Australian Newspapers Digitisation Program', paper presented at the 23rd Australasian Conference on Information Systems, Geelong, 3-5 December 2012, and available at: http://trove.nla.gov.au/version/188319518

[x] Alam, Sultana Lubna and Campbell, John, 'Role of relational mechanisms in crowdsourcing governance: an interpretive analysis' (paper to be presented at the Nineteenth Americas Conference on Information Systems in Chicago, August 15-17, 2013), and 'Dynamic changes in organizational motivations to crowdsourcing for GLAMs' (paper offered for inclusion in the program of the Thirty-fourth International Conference on Information Systems, to be held in Milan, December 15-18, 2013).

[xi] Paul Hagon's research is available on the website: http://www.nla.gov.au/our-publications/staff-papers/trove-crowdsourcing-behaviour. Many other staff papers on Trove are also available on the website.

[xii] Sylvia Marchant, 'John Warren, champion text corrector', *Friends Newsletter*, June 2013.

[xiii] The National Library of Finland, for example, partnered with Microtask to develop Digitalkoot, incorporating a gamified approach to correcting digitised newspaper text (http://blog.microtask.com/2011/02/digitalkoot-crowdsourcing-finnish-cultural-heritage/). Two games, Mole Bridge and Mole Hunt, invited users to play online games to achieve levels of certainty over OCR'd words. The endeavour appealed to nationalist sentiments ('Start saving ... Finnish culture here'), but seems likely to have appealed to precisely the under 50s audience that is less engaged with Trove and other digitised newspaper services. A study of this quite different approach to the OCR problem would round out the studies mentioned in this paper.

[xiv] The National Archives of Australia's Hive community transcribes government records. More information about the service is available at: http://transcribe.naa.gov.au/

[xv] Like the 'typical' Trove user, the author of this paper is female, fifty years of age, has post-graduate qualifications and enjoys a comfortable income...but is not a family historian.