Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers

Rose Holley

March 2009
National Library of Australia
ISBN 978-0-642-27694-0

About the author:
Rose Holley has worked in the digital library environment for many years and is a specialist in digitisation of cultural heritage materials and delivery of digital resources to users. She has worked in the UK, New Zealand and most recently Australia. She was appointed in 2007 to manage the Australian Newspapers Digitisation Program at the National Library of Australia.

Acknowledgements from the author:
I would like to acknowledge the creativity, enthusiasm and hard work of the ANDP team over the last two years. They worked tirelessly to achieve the Library's goal of free public online access to full-text searchable historic Australian newspapers and have been pivotal to its success. The team was small, close knit and high achieving. The combination of each person's specialist skills, knowledge, and ideas in combination with a shared vision for the future has lead to a very successful outcome. Special thanks to the team who were: Kent Fitch (Lead Architect), Ninh Nguyen (Programmer), Bronwyn Lee (Business Analyst), Mark Raadgever (Project Assistant), Cathy Pilgrim (Director) and myself (Project Manager, IT Manager). I would also like to thank the team for their support and encouragement towards me as the Project/IT Manager.

Contents: **Page**

1. Background of the ANDP

The Australian Newspapers Digitisation Program (ANDP) is an ongoing large scale national collaborative program that is digitising historic Australian newspapers from 1803 onwards. Initially a major daily newspaper published prior to 1954 from each of the eight states/territories is being digitised. The digitised images and full-text are being made available via the web to the public as a free full-text searchable service. The Australian Newspapers service was released as a beta version to the public in July 2008 and now has 3.5 million articles available (360,000 newspaper pages). The program aims to deliver over 40 million articles (4.4 million newspaper pages) by 2011. The National Library of Australia is leading and managing the program with every State and Territory Library in Australia being a contributor. The Library has developed software to enable workflow management and quality assurance of the newspaper digitisation process (the Newspapers Content Management System) and for delivery of the newspapers to the public (the Search and Delivery System http://ndpbeta.nla.gov.au[1]). The Library is also funding and has set up the national infrastructure including storage for all newspaper data. Project documentation is available from the public website http://www.nla.gov.au/ndp.[2] The ANDP team is small and is made up of the Project Director, Project Manager, Project Assistant, Lead System Architect, Programmer, Business Analyst, and Quality Assurance Assistants.

2. Striving for Quality

Of all the cultural heritage items that can be digitised, newspapers are one of the most challenging types. In addition the extent of newspapers for any given country is large, making the digitisation of a corpus costly. The first newspaper was not published in Australia until 1803 making the extent much smaller than that of European countries. It is also fortunate that any Australian newspaper published prior to 1954 is out of copyright. This has enabled the Library to freely digitise newspapers from 1803-1954. However early Australian newspapers are of very poor print quality due to the first printing presses being those withdrawn from service in England and sent to Australia, and due to there being a lack of suitable paper in the new colony. Australian Newspapers have unique and internationally sought after content since the early newspapers also served as Government Gazettes and listed all the convicts brought to Australia on convict ships, the name of the ship they were transported on, their movements around the prison colonies, their Ticket of Leave, Certificate of Pardon, Certificate of Freedom, and Absolute Pardon's. Convicts could leave the colonies after their sentences were completed or after being granted an Absolute Pardon. Departures were announced in the Sydney Gazette's 'Notice of Intent' column. The 'Shipping News' section of early Australian papers is very significant for genealogists providing unique information about the movement of convicts and other people. The settling of Australia by the British is well documented in the newspapers, and also the treatment of the indigenous aboriginal peoples at that time. All of these factors make Australian newspapers a rich resource for researchers.

The Library first began to digitise newspapers in 1996 under the Australian Co-operative Digitisation Project (ACDP) but this was not continued for two reasons. Firstly there was concern about the quality of the digital output from the newspapers particularly the poor Optical Character Recognition results (OCR) results, and secondly the huge cost involved in undertaking enough newspaper digitisation for the results to be useful to the public. Eleven years later once OCR technologies had radically improved and the Library had secured a significant amount of funding ($10 million), digitisation of newspapers became the top priority for the Library.

The ANDP team had from the outset in January 2007 decided to make a considerable investment in software development in order to be able to quality assure digital outputs to ensure they met minimum standards and to future proof the files in case they could be improved further in the future. Contributors had all expressed concern that the digital outputs (image quality, OCR text) may not be good enough to enable adequate full text retrieval or to meet user expectations. The ANDP team had regularly brainstormed and reviewed ideas to improve the quality of outputs and had implemented a number of ideas to achieve this[3]. At this stage the team were assuming that quality of data entirely relied on the Library's digitisation technique. It had not yet occurred to the team or the contributors that the public may play a role in improving and enhancing the quality of the data. Therefore actions that were taken to improve quality for users were:

- Enhancement of image files for OCR
- De-skew of each page to make text horizontal
- Testing to see which file types give best OCR results (bi-tonal or greyscale)
- Testing with dictionaries during OCR process
- Re-keying title, subtitles, author and first four lines of article text
- Applying categories to articles (for use in relevancy ranking of and narrowing of results list)

A further action was also identified - to use an OCR-error 'confusion matrix' and language modelling after the OCR process as an automated method of improving OCR quality. This has not yet been tested. The Library undertook a 50,000 page pilot of good, bad and average quality pages so that the end results for users could be reviewed by contributors and the quality assessed. The Library and the scanning/OCR contractors were in agreement that the highest possible quality had been achieved from relatively poor original material, but the quality of the OCR was still not as good as was desired. The suggested way for further improvement would be for manual intervention in the OCR process or for further re-keying of data. It was at this point that a member of the IT team suggested that the public may be able to assist for free. He was firmly of the opinion that the public would want to get involved in correcting OCR text and would do this for the 'common good'. If this was the case it may be effective for a large scale program such as this.

3. Benefits, risks and issues of public collaborative OCR text correction

The team discussed benefits, risks and issues of allowing users to directly alter text. The potential benefits identified were:

- Data quality is improved for all users
- Keyword searching is improved for all users
- The community becomes involved and engaged in enhancing and enriching the resource
- Users become empowered
- Innovative thinking is demonstrated
- The method to improve data quality is cost effective
- If successful the method could potentially be applied to other full text projects
- Web 2.0 technologies could be utilised and the service would have a 'cool factor'
- The resource is more likely to meet user expectations if digital content is improved
- New virtual user communities and social networks would be built

The risks identified and their mitigation strategies are as follows:

Risk	Mitigation/Solution
No such functionality has been implemented by a Library before and there are no policies in place	• Develop a disclaimer and terms of use for the service • Test in a beta version
Potential vandalism of text	• User can always see original image of page so will not unknowingly view vandalised text • Roll back to original OCR raw version on identified cases/entire database • Disable correction functionality if vandalism occurs • Make login mandatory instead of optional and then block specific users • Rely on the public to either not vandalise or to report vandalism • Develop a moderation module if required once adequate testing has taken place in beta
Large amounts of text correction activity compromise database/service	• Make beta a soft launch, no press releases, just word of mouth so that usage increases gradually
Users don't do text correction and development time is wasted	• Test in beta for several months • Implement basic correction only without a moderation module until it is proven that users want to be involved. Do further development after proven. • Provide information about how it will help everyone else if users participate • Encourage contributor libraries to promote it to their users • Promote service
Users don't understand the concept of text correction	• Give users adequate time to understand the concept and make beta available for several months. • Don't use the term 'OCR' ; instead use a term that was understood in user testing 'electronically created text' • Liken concept to that of Wikipedia • Improve interface and add help
Users are put off using the service because of seeing all the raw OCR needing correction which may be 'gibberish' and clutter screen	• Put a frame splitter in beta so users can hide or show raw OCR and change the width of it on screen. Therefore some users can view newspaper page image only and some the full screen of OCR text.
Users are confused about the difference between text correction, adding comments and adding tags and how to do it	• User testing before release of beta • Terms tested on users • Make functions clearer on interface • Add 'what is this' on interface • Add help text • Populate some articles with examples of the features before releasing Beta
Not sure how technically difficult it will be to implement text correction and how long it will take	• Allocate 2 weeks for development work and see what can be achieved in this time.
Not sure what an effective user interface for text correction will be like and how it can be developed since this has never been done before.	• Use expert user interface designer • Test interface with public users

Can we prevent robots and spam from compromising the service?	• Verify users with a captcha before any corrections can be made. • Stay up to date on captcha developments to stop robots.
Text correction once implemented may require high level of staff support and generate further work.	• Test in beta and review the level of ongoing support and resource needed.

The issues and technical questions we asked ourselves included:

- If moderation is required who will do it (Library, contributors, users), how will it be done, and how time consuming will it be?
- Will the page move down as the text corrector moves through the corrected text of long articles? How can we synchronise the page with the corrected text since the user needs to see it in the right place at the right time?
- Should users be correcting by character, word, line, paragraph, or article, and how should we save these corrections?
- Should there be a quick single word/line correction mode and also a 'power use' mode for those users who want to correct entire articles line by line?
- Should login be mandatory or optional? Mandatory will enable tracking of users but may put off casual users who just want to change a few words.
- How will we know what users and doing and gather useful statistics?
- Can we measure the improvements made to data?
- Should we keep all changes made by users and enable searching across all changes or just latest changes?
- Which versions should the user see before correcting – the latest only, the original or all versions?
- If there is vandalism, how will we roll back and will it be at article level or whole database?
- Should users be able to correct the title, subtitle and first four lines of article text that have already been re-keyed by contractors to 99.5 % accuracy?
- Should users be able to search the corrected text only or specify if they want to search raw text and corrected?
- Should multiple users be able to correct articles an infinite number of times or can an article be 'finished' and locked down?
- If a poor quality page is replaced in the database all the text corrections and tags will be lost.
- Should users be allowed to add missing lines to articles, change line breaks, or line formatting (i.e. denote text as italic or heading?
- How can users transcribe display advertisements when the words have been treated as images (not text) so there are no lines to correct?
- How can we match the old words with the corrected 'new' words so that word highlighting will still work in searches?
- How will users actually do this on the interface, how will we develop and test user interface mechanisms?
- Should users have keyboard shortcuts for example enter at end of line and F12 to move to next word/skip over words?
- Is it possible to have correction "candidate" words – for example if the text was OCR'd as "nniagme" we may suggest "imagine". But how could we identify the candidates and how to suggest they can be changed (dropdown?).
- Is it possible/desirable to prevent some words being changed too radically for example not possible to change 'imagine' to 'telephone', because it is so visually different, that it is unlikely that an OCR mistake could have been made?

4. Implementation of Text Correction

The Library agreed that the ANDP team should develop the Australian Newspapers service in an innovative way, implementing anything that would enable users to interact with and use the data in new ways. The team decided that the best way to do this would be to release a prototype to Contributors first for comment, followed by a beta version for public use. Feedback from users could then be gathered. The ANDP team developed a prototype search and delivery system in 6 weeks and released this for testing to contributors (Australian State and Territory Libraries) in December 2007. This included the text correction facility. Positive feedback was received from the contributing libraries though all queried if a moderation module was going to be developed and if so assumed that the Library would moderate public text correction. After the feedback a Search and Delivery beta system was developed incorporating the same functionality as the prototype. This took 8 weeks. A further 8 weeks was spent on user interface design. This included four rounds of user testing using a representative sample of the public. The user testing covered all the main functions of the service and also was our first glimpse into what the public would think about being able to correct text. They were shown a screen where they would be able to do text correction and asked what they would expect to be able to do and how at this stage. User's responses to text correction were as follows:

- The entire concept of text correction was new and they didn't understand it until it was explained.
- They did not expect to be able to do anything like this.
- If text correction was possible surely it would only be librarians allowed to do it?
- They did not expect to be able to do more things than you can do in Google. If it's possible why doesn't Google do it?
- Users don't understand what OCR is or that the searching works on the OCR.
- The term they agreed was meaningful instead of 'OCR text' was 'electronically translated text'.
- They looked at the OCR text or the image but not both even though both showed on the same screen.
- They did not initially understand the relationship between the text and the image or how to synchronise the text with the image.
- When they were told they would be able to do text correction and how they were firstly amazed, secondly impressed, and third keen to have a go.
- Once they had tried some text correction they likened it to the user's involvement and enhancement of content in Wikipedia and it made more sense to them.
- They had confusion about the difference between correcting text, adding comments and adding tags to articles.
- They assumed some form of moderation would take place and queried how this would happen.

The feedback was very useful in order to help us develop a user interface. The search and delivery beta system was released to the public on 25 July 2008. There was no promotion during this time since the Library was concerned about getting overwhelming use and feedback for the service. During the next 6 months 3.5 million articles were added to the service and public users were actively encouraged to give the Library feedback and suggestions for enhancements. During this time several thousand members of the public found the beta service and became active searchers, taggers and text correctors. A text correcting community of around 1300 people quickly developed in the first six months 2 million lines of text in 100,000 articles was corrected. Feedback from users was compiled and made public in January 2009.

5. How Text Correction works

The way the text correction works behind the scenes is as follows:

1. The raw OCR text for each article on a page is supplied to the Library from the contractor in an ALTO XML page level file. Most of the information from the ALTO file (paragraphs, line breaks, bounding box co-ordinates of each word, fonts/styles picked up by the OCR engine), is retained in the SQL database with the exception of the OCR confidence at the word level.
2. The text re-keyed by the contractor (title, subtitle, first 4 lines of article) for each article is supplied to the Library in A METS/MODS file at Newspaper Issue Level. This information is stored in the SQL database.
3. The public corrections to the article text are stored directly in the SQL database (they do not go into ALTO or METS/MODS file).
4. The storage structure in the SQL database was designed to minimise space and speed of parsing (for later extraction), so the text information is stored in a binary (non standard) format, not as an XML file.
5. For each word position the bounding box (coordinates) are recorded in SQL and the source of the word:
    a. Source 1 - the original word from raw OCR engine
    b. Source 2 – re-keyed words from contractor (by heuristically matching the re-keyed title/1st 4 lines with the raw OCR text).
    c. Source 3 - reconstructed hyphenated words (e.g. if the OCR records as 2 words "sing-" and "ing", we also store "singing" at the same position as "sing-" - this aids phrase searching)
    d. Source 4 – 'public' OCR text corrections (there can be any number of these as the same word is corrected and re-corrected)
6. The search index is Lucene. Each of the words at the same position is indexed in a Lucene database full text index. Lucene knows the word position for each word in an article but does know the other information (that is in SQL).
7. Lucene is good at storing word positions and can be used as a way of implementing synonym search. So, for example, you can index in Lucene:
   "The Chemist{/Pharmacist/Pharmacy} shop{retailer} was closed{shut}" - this is a phrase of 5 word positions and would phrase-match queries such as:
   "chemist shop was shut"
   "pharmacy retailer was shut"
   "pharmacist shop was closed"
   This feature although powerful also means that text corrected from say A to B still shows up in a search for A. Users have noticed and questioned this. But it provides robustness, so that if someone vandalised text it would still be discoverable using the words that had been replaced.
8. In the public delivery system the full text rendition shows, in preference order: the latest correction, if any; if not the contractor re-keyed text, if none the raw OCR text.
9. The Lucene index is periodically updated and the search system has to reopen the updated Lucene index to see the updates. The number of updates that can take place is not an issue (e.g. 1 update every 2 minutes or 10,000 updates in 2 minutes); however the actual act of updating is an issue. Lucene excels in a read-only environment and struggles with regular updates. However the Lucene community is working to improve real-time handling of updates.

The text correction user interface operates as follows:

1. When a line of text is right clicked or the "help fix this text" mouse over is clicked, JavaScript in the browser puts the text of the line into an input text box. The user can overtype/edit the text.

2. When the user hits 'enter' on the keyboard the cursor is moved to the next line and a new input box is created.
3. When 'save' is clicked, JavaScript on the client creates a list of changed lines and sends them to the server, using AJAX (JavaScript issuing a HTTP request). Each changed line is preceded with the XY co-ordinates of the 1st word in the line which lets the server match updated lines.
4. The server code matches up the existing and changed lines. For each changed line, it tries to 'line up' the old and new words so that the new words can be highlighted on the page image - that is, the system uses the bounding box of old words to correctly highlight new words. This is a hard problem in general which we solve heuristically by matching the longest runs of matching characters. But it is far from perfect, and doesn't work very well when the number of words in a line is significantly changed (words inserted/combined/deleted).
5. The server does not allow lines to be inserted or deleted.

Fig 1. User Interface - Australian Newspapers beta at article level. Enhancement options (text correction, tags and comments) on the left side of screen for the article being viewed.

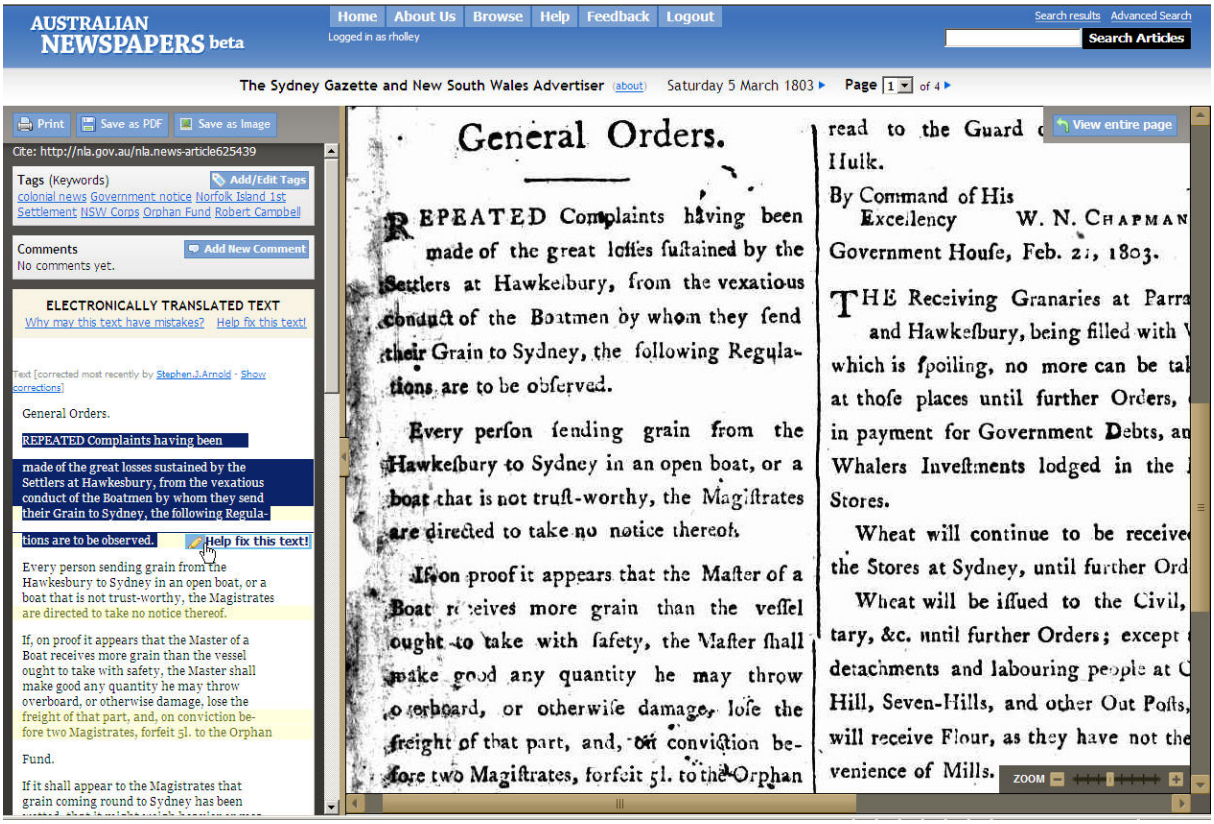


Fig 2. Show all text corrections. Users can see the history of corrections on the article

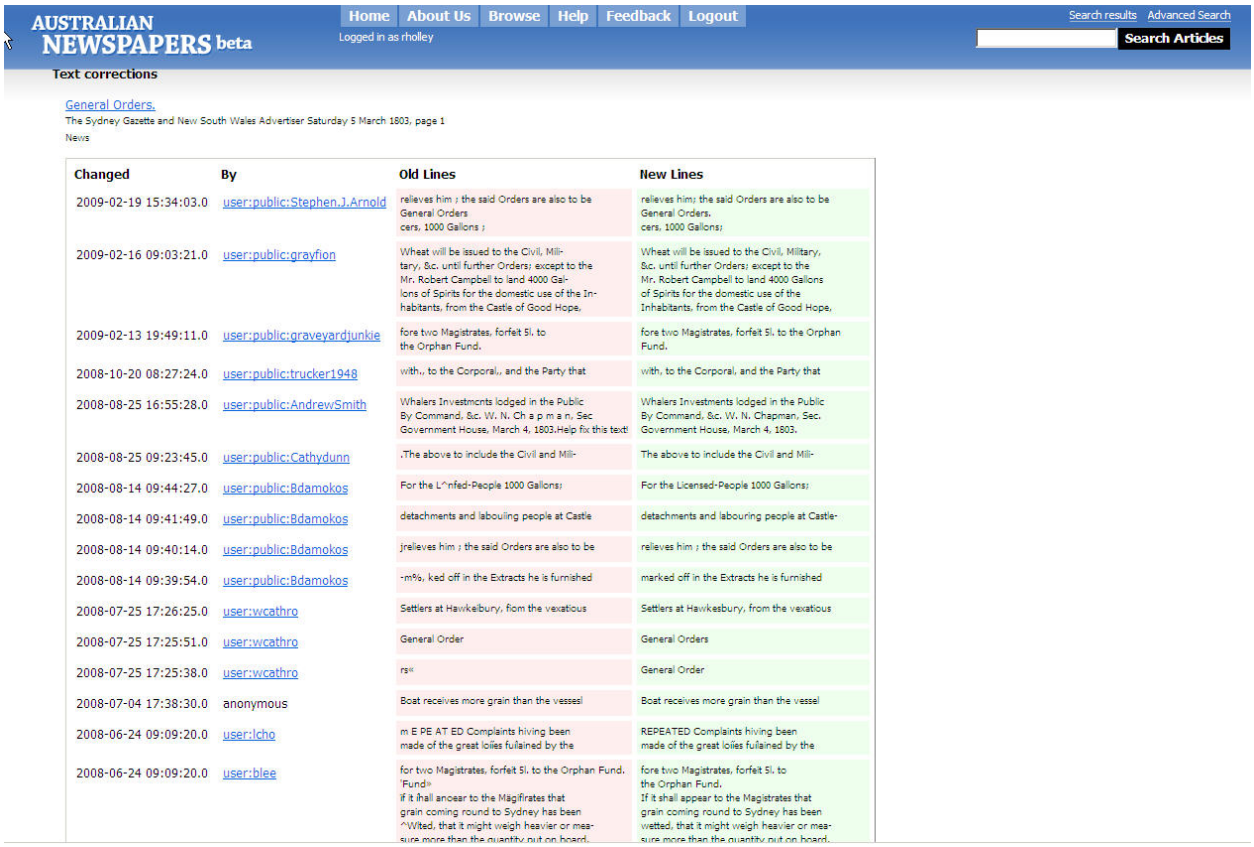
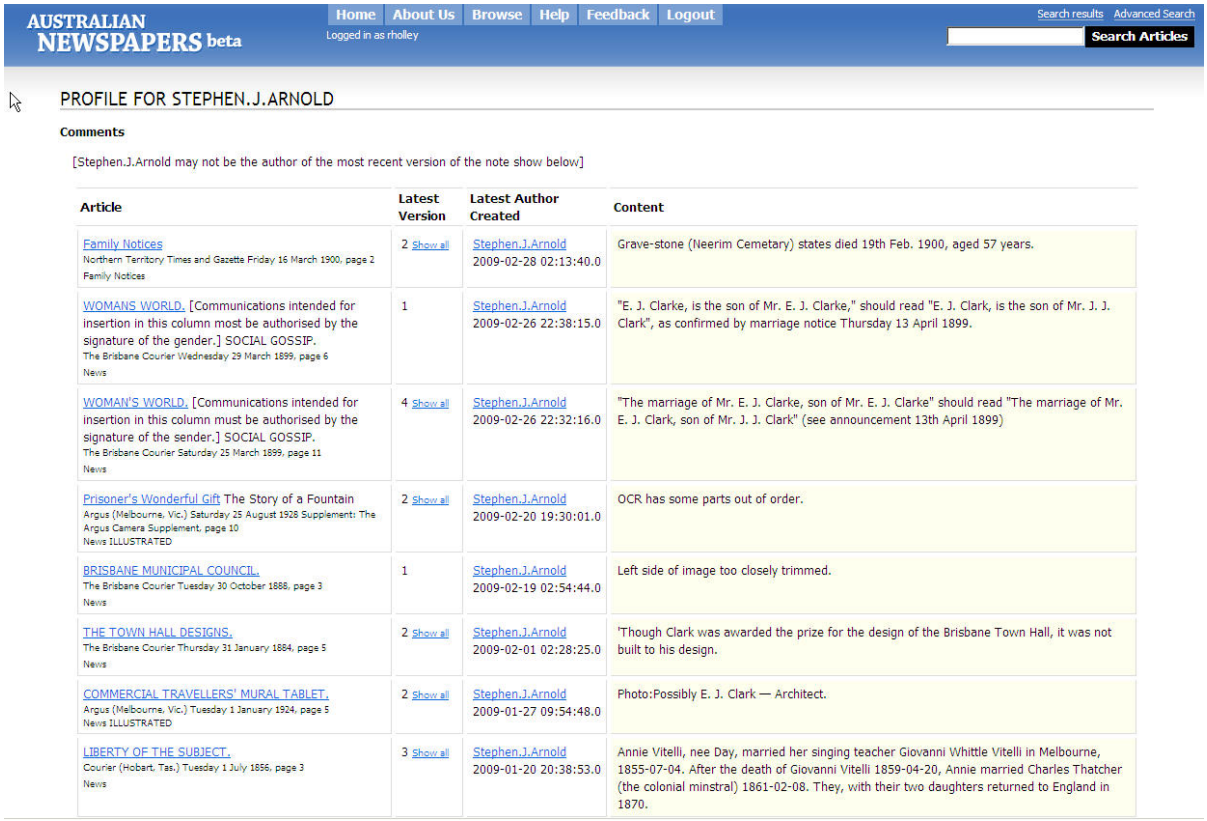


Fig 3. User activity. Users can view their own (and others) recent activity of commenting, correcting and tagging.



Tag cloud

'Bendigo Mac' (8) 'Black Harry' (3) 'The man with the iron mask' Annie Vitelli (2) Bird Lovers Charles Thatcher Clarence Holt (2) Clarissa Holt Colonial Architects (53) Donald MacDonald (5) E J Clark Edward James Clark - Architect (82) FARJEON Benjamin Leopold (182) FARJEON Eleanor FARJEON Harry (2) FARJEON Joseph Jefferson (18) Fiction G V Brooke Geo C Inskip - Architect (13) J J Clark John James Clark - Architect (174) Levee (4) Maria Morton Painter (2) May Holt obituary (2) Osman Pasha Otago Gold Fields (4)

Text corrections

Stephen.J.Arnold has contributed corrections to **26416** lines; most recently:

Article	Changed	Old Lines	New Lines
OPENING OF THE EXHIBITION. The Brisbane Courier Thursday 21 August 1884, page 5 News	2009-03-09 19:41:37.0	excess of pips. Thee othei exhibits; which also atti acted gi cat attention carne fiom the gardens of Mi J Holmns of Ballal da Camp.	excess of pips. The other exhibits; which also attracted gi cat attention came from the gardens of Mr. J.Holmes of Ballard's Camp.
OPENING OF THE EXHIBITION. The Brisbane Courier Thursday 21 August 1884, page 5 News	2009-03-09 19:40:04.0	thev carno to inspect the oi anges Tor the timo of the yeal nothu g could havo been bettor except inoio extended competition One dish in particular calls, foi attention It is a non compctihio entry b} Mr James Voller, of the Samford load The oranges aro seedling consist of largo luscious fruit, tho flax oui is unexceptionable, and they aio fieo from an excess of pips lue othei exhibits; which	thev came to inspect the oranges. For the time of the year nothing could have been better except more extended competition. One dish in particular calls, for attention. It is a non competive entry by Mr. James Voller, of the Samford-road. The oranges are seedling consist of large luscious fruit, tho flavour is unexceptionable, and they are free from an excess of pips. Thee othei exhibits; which
OPENING OF THE EXHIBITION. The Brisbane Courier Thursday 21 August 1884, page 5 News	2009-03-09 19:37:01.0	There w as onl} one pineapple for competition, and this was not deemed woithvof a priz» A moic cheonng sight greeted the judges when	There w as only one pineapple for competition, and this was not deemed worthy of a prize. A more cheeeing sight greeted the judges when
OPENING OF THE EXHIBITION. The Brisbane Courier Thursday 21 August 1884, page 5 News	2009-03-09 19:36:15.0	hnni'somol} a ranged basket of cut flowers and a b-ulal botiq ict composed of the purest and choicest white floweis, which came fiom Ash grove, aio ver} striking It was a pit} i ot to see ruoio wild floweis, as such exhibits are within the reach of bo many, and thoso sent in did not appoai to have iecelvod much cale fiom the gatheicis The next division in the horticultural section comes utitict the heading	hansomely aranged basket of cut flowers and a bridal bouquet composed of the purest and choicest white flowers, which came fiom Ash-grove, are very striking. It was a pity not to see more wild floweis, as such exhibits are within the reach of so many, and those sent in did not appear to have received much care from the gatherers. The next division in the horticultural section comes under the heading
OPENING OF THE EXHIBITION. The Brisbane Courier Thursday 21 August 1884, page 5 News	2009-03-09 19:33:20.0	haps as well that in future sonic idea should be given as to the w that a hand bouquet »hould atfa n ns exhibitor appen red woefully abroad in this malta of dimension, which chiefly ciused the tno onginl j uors to dispute. A	haps as well that in future some idea should be given as to the what a hand bouquet should attain as exhibitors appeared woefully abroad in this matter of dimension, which chiefly caused the two original jurors to dispute. A
OPENING OF THE EXHIBITION. The Brisbane Courier Thursday 21 August 1884, page 5 News	2009-03-09 19:31:33.0	final decision difficult in icfei euee to numbers 1053 and 1052 a third judge being called in and deciding in favour of the foi mci It is per	final decision difficult in reference to numbers 1083 and 1082 a third judge being called in and deciding in favour of the former. It is per-

6. Activity of Users and Feedback in the beta service

During the first 6 months of beta use August 2008 – January 2009 several methods of monitoring user activity and gaining feedback were utilised:

- A feedback survey located within the beta service.
- A web 'contact us' form on ANDP website and linked from beta 'about' and 'feedback' page.
- Observation of user activity and use of the system by gathering statistics and watching user activity.
- Comments posted by users to blogs and forums.
- Comments from Facebook fans on the Australian Newspapers Facebook Fan page.
- Feedback from users in State and Territory libraries gathered via ANPlan [4] (Australian Newspaper Plan) members.
- Analysis of feedback and statistics by ANDP team.
- Direct contact with specific users by e-mail or phone. (Direct phone and e-mail contact from users to ANDP team was not generally encouraged due to there being no operational service support, however sometimes it was useful for the ANDP team to make direct contact with users.)

The ANDP team received feedback from more than 600 individual users during August – December 2008. Individual users generally made several suggestions and provided multiple comments within their feedback, resulting in thousands of specific items being raised. Many of the comments were repeated by different users. It was a significant task to compile, analyse and review all of the feedback received. 340 individuals responded to the survey feedback form and 260 users contacted the team by other means. ANPlan members reported on feedback from users in State and Territory libraries at a meeting held at the Library on 28 November 2008.

An overwhelmingly positive response to the Australian Newspapers beta service was received from users, with many stating that it already exceeded their expectations. A summary of the activity and responses regarding collaborative text correction is below:

- Users did not expect to be able to correct OCR text and it was initially difficult to convey the concept, purpose and technique of doing this.
- Once users realised they could correct text they were overwhelming positive about the idea.
- When users tried text correction many said they found it addictive and rewarding and were actively correcting much more text than we had expected.
- Text correctors had 2 modes of operation
    a) Correct the odd word here and there as they read articles
    b) Correct the entire article line by line.
- Serious text correctors usually did regular sessions of between 1-3 hrs of text correction at a time.
- Many of the text correctors found out about the service from genealogy blogs, forums and noticeboards.
- Initially 50% of the corrections were done by anonymous users and 50% by registered users but by the end of 6 months 80% of corrections were being done by registered users.
- By the end of 6 months 1300 users were in the virtual text correcting community.
- Only 49% of registered users were correcting text. The other half of registered users appeared to be tagging articles instead.
- Users who correct text do not necessarily also tag articles.
- In the first month of use over 200,000 lines of text was corrected in 12,000 articles, by the end of 6 months 2 millions lines of text had been corrected in 100,000 articles.
- At no point since release of beta has there been a time when text correction is not taking place. It continues 24 hours a day 7 days a week.
- 78% of users were based in Australia but there was also a growing international community with users in the United Kingdom, United States of America, New Zealand and Canada. One of the top ten correctors was based in USA.
- The top ten text correctors were correcting significantly more text than all other users spending up to 45 hours a week on the activity. The top corrector at the end of 6 months had corrected 101,481 lines in 2594 articles. The same correctors remained in the top five for the first 6 months.
- No vandalism of text was detected in 6 months so no roll back to previous versions or moderation was required.
- Some users suggested that users should moderate each other (rather than the Library moderating) and have the ability to report and correct issues.
- Users had many suggestions for the enhancement of the text correction functionality including improvement of the text correction method, a power user mode for whole articles and the ability to add in missing lines of text.
- Users said they were motivated by addition of new content, the knowledge they were helping other people as well as themselves, the idea they were helping to record Australian history accurately, helping a good cause, being able to make a small but effective contribution to the big picture.
- Users said things that would increase their motivation would be public recognition, public ranking tables, user profiles, and the ability for them to communicate with other text correctors.

Some quotes received from users about text correction were:

'I like being able to see the OCR text and the original side by side. This is a smart move to entice people to correct text.'
'OCR text correction is great! I think I just found my new hobby!'

‘It’s looking like it will be very cool and the text fixing and tagging is quite addictive.’
‘I was not aware that user editing was going to be included, but what a great idea.’
‘An interesting way of using interested readers “labour”! I really like it.’
‘A wonderful tool - the amount of user control is very surprising but refreshing.’
‘It is wonderful that you have made it available before it is "finished". Also kudos on incorporating features - especially text-correcting - unavailable in any other product. Fantastic work!’
‘You're providing a great service and I've been very happy to correct the errors I've found in the articles I've been viewing, thanks.’
‘A great idea. I would sooner have online access, and search possibilities, even with the mistakes, and am happy to correct the scanned text as a quid pro quo.’
‘Would be happy to spend an hour a week correcting pages.’
‘Thank you! You lot are so cool!’
‘I am especially impressed by the creation of crude OCR (and its implications for searching) and the invitation to end-users to assist editing/correcting the text. The National Library is to be congratulated on its pioneering work in this area.’
‘Wonderful!!! First article I came across was gobbledegook, but correcting the text is very intuitive, no problems.’
‘Great development (and very addictive)’
‘This is a really useful project. And I like the invitation to correct text - many hands and all that’
‘I appreciate the great work being done on this project and pleased to help where possible.’
‘I love this site and have made many corrections.’
‘This service is absolutely brilliant and I'm sure it will take off with its own community of modern day sub-editors.’
‘The Australian Newspapers digitisation program is the best thing that has ever happened to me in twenty years of family history research. Is there anything I can do to show my appreciation?’
‘Thank you all for your efforts. This is a wonderful resource and so valuable for folk who can only do research on-line. I look forward to continuing search successes and am happily correcting text when I can.’
‘Would like to say this is a great initiative although I think there should be a warning about using this site and its possible addictive effects! I have a great deal of trouble getting back to what I should be doing at times.’
‘Correcting electronically translated article text is a worthwhile and enjoyable task for retired or interested people. Thank you for a great concept.’
‘I think this service is fantastic. I have found some very useful information with my searches and am expecting to find more as the text is improved. I am happy to contribute by editing.’
‘Wow – well I got sucked in! I can see why everyone is editing the OCR text… its compulsive!’
‘I applaud the capability for readers to correct the text.’
‘It is great that the NLA is breaking new ground in allowing the public to help them correct the OCR'd text.’
‘The decision to let members of the public edit & correct articles must have been a difficult one, but it's a great idea.’
‘There's an idea in Sociology of social capital: referring to the organisations out there that do positive work in the community and make it what it is. Help poor people, lonely people, sick people etc but

also develop the sense of community and do a great deal for the psychological and sociological health of people. Australian Newspapers Beta is one such constructive entity. Helping people come to terms with their past, helping them define themselves. It's a really big thing, I think and is a contribution to the health of the community.'

As at 18 February 2009 there have been 2.2 million lines and 104,000 articles corrected (see monthly tables below).

Fig 4. – Text corrections by line by month

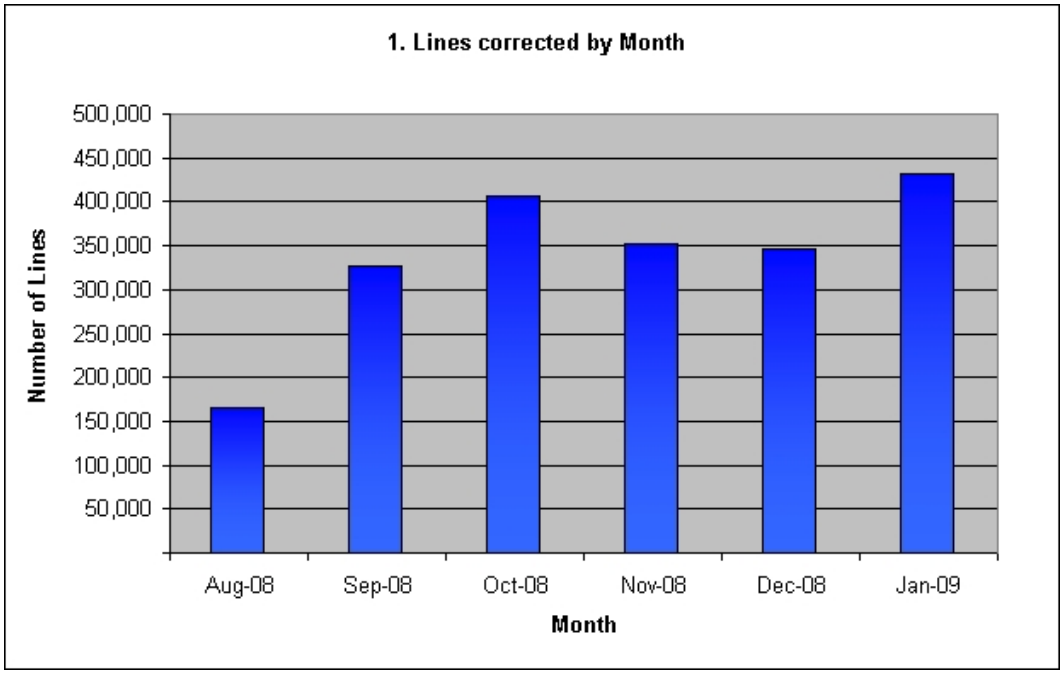


Fig 5. Text corrections by number of articles by month

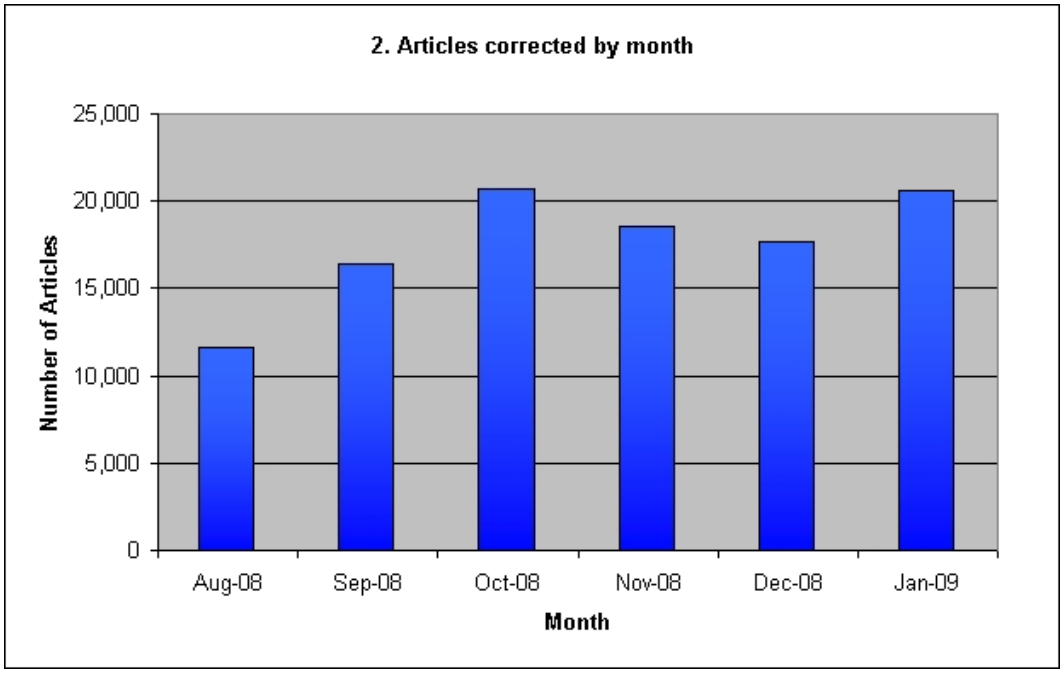
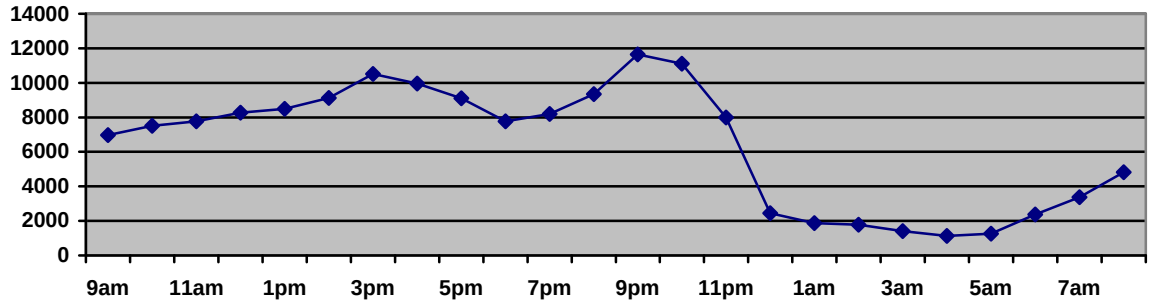


Fig 6: Text corrections by time of day 4 August 2008 – 3 November 2008 (number of times 'save OCR corrections' is clicked).



OCR text correction rises steadily throughout the day peaking at 3pm and 9pm (with a small dip around from 6-7pm as users presumably get their evening meal), and surprisingly continues throughout the night (though this may also be due to overseas users). OCR text correction is occurring 24 hours a day.

Fig 7: Sample of Text correction activity on an individual article 2 November 2008

Time	User	Activity
12:23	user 1	Corrects text in the article but doesn't know how to insert the pound symbol
12:24	user 1	Creates a comment on the same article saying that he doesn't know how to enter the pound symbol.
19:13	user 2	Enters the pound symbol in the article.
23:09	user 3	Spots an error in the correction by user 1, and fixes it.

Fig 8: Top text correctors ranked by number of lines over 6 month period 25 July 2008– 22 January 2009

Ranked out of 1285	ID	Lines corrected	Articles corrected
1	Jhempenstall	101,481	2594
2	Cmdevine	90,823	1585
3	Fwalker13	80,437	642
4	Mrbh	79,248	1439
5	Maurielyn	72,129	1192
6	John F Hall	59,111	1632
7	Jdickson2	28,796	2407
8	JamesGibney	25,106	479

Note: All other registered users have achieved considerably less than the top 6 users. Top 6 users have been consistently active for 6 months.

Fig 9: Increase in usage of Australian Newspapers and features:

	At Nov 2008	At Feb 2009
Number of registered users	1488	2994
Lines of text corrected	1 million	2.2 million
Number of articles corrected	60,000	104,000
Top corrector	60,000 lines in 2000 articles	103,000 lines in 2600 articles
Number of comments added	800	1806
Number of tags added	18,000	43,000
Unique visitors to site	94,000	205,000
% of users in Australia	78%	75%
Content in service	367,000 pages 3.5 million articles	367,000 pages 3.5 million articles

The comment functionality has been not been used very much in comparison to the tagging or OCR correction features. Comments were implemented primarily for researchers so that extra information about articles could be added. It was originally intended to call the feature 'annotations', this was changed to 'notes' but in user testing there was still a lack of understanding of what the feature was for and so it was released in beta labelled as 'comments'. The minimal use of the comments feature could be because of the lack of understanding of the feature. Also comments cannot be edited or deleted at this stage and the user has no facility to view all comments or search comments. Comments can be viewed only when an article that has comments on it has been found. On reviewing comments 3 things have been noted:

- Users are using comments as a mechanism to communicate with other users and ask questions (e.g. how do I put in the pound symbol in OCR correction?) This could be because there is no forum or online enquiry system at present, or because the users think this is what 'add comment' means.
- Feedback shows that users want to be able to paste hyperlinks in the comment field to link to other related sources, and they are unable to do this at present. This would indicate that users want a related article/resource feature.
- Some users have added comments about the article and used the feature as we intended.

Tags can also be added to individual articles. The ANDP team were initially dubious about the value of adding tags to full-text articles. Tagging has mainly been utilised on images, books and web pages to date. In theory if you can search for full-text it is less likely that you will need to add tags. A basic tagging facility was therefore added (individuals can add, edit or delete their tags, but searching and management of tags is not yet available). Surprisingly, a very large number of tags (14,270) were added by users in the first few weeks and this continued to increase exponentially until after 6 months 43,000 had been added. The majority of these tags are personal names and are possibly being used so that users can track their family research. 12,210 of the tags are unique and have only been used once. Of the tags created only 34 have been used more than 100 times. The most used tag is LRRSA and has been used 1613 times. This tag is being used by a group to track their research. The majority of users are tagging 1-5 times and often using the same tag (a family name). The most tagged articles all have 50 tags since that was a limit initially imposed. This has now been lifted in response to user feedback.

Fig 10: Summary of beta tagging statistics 23 Feb 2009

Total amount of tags added	46230
Number of distinct tags created	16534
Number of tags used 1-5 times	15390
Number of tags used more than 100 times	34
Total amount of articles tagged	19354
Amount of articles with more than 10 tags associated with them	289
Highest number of usages of a tag	LRRSA = 1613 Murder = 478 Bendigo = 412

The relationship between tagging, text correction and adding comments is still not fully understood by the ANDP team. All three actions enhance the data but why a user would choose to tag instead of correct, or correct and also tag or correct and not tag needs further research. Users also want the ANDP team to provide them with some guidelines on using and creating these features.

7. Profiles and Motivations of Text Correctors

We were interested to find out more about the text correctors. We wondered who they were, where they were located, what motivated them, would they do more or less as time went on, and why were they addicted to correcting text? The top five text correctors over the last 6 months were identified (the same people had remained in the top five each month) and were each sent a questionnaire[5]. All responded immediately and their answers were very interesting. Below are the brief profiles of the top 5 text correctors.

Julie (Australian)

Located: rural town near Bendigo, Victoria.
Interests: Family history, local history.
Age and status : 31-45 (stay at home mum)
Focus on single newspaper title 'The Argus', region around Bendigo and own family names. Always adds tags to articles as well as correcting text (some articles need more than 50 tags). Spends 15-45 hrs per week.
Why she does it: "I enjoy the correction - it's a great way to learn more about past history and things of interest whilst doing a 'service to the community' by correcting text for the benefit of others".
Why keep doing it: " The knowledge that you are doing something that will benefit future people that wish to access articles on their family history"
Will you carry on indefinitely? "Yes"
Is it addictive? "Yes. A must do mission"
What would keep you motivated? "More papers added"

Lyn and Maurie (Australian)
Located: Brisbane, Queensland. Interests: Family history, local history, shipping. Age and status: 55-62 (retired couple) Background: Working together on research. Not tagging. 15 hrs per week. Also transcribing shipping lists for http://mariners.records.nsw.gov.au Why they do it: "We are sick of doing housework" Why keep doing it: "Because it's addictive. It helps us and other people" Will you carry on indefinitely? "Yes" Is it addictive? "Yes, hard not to correct errors when you see them" What would keep you motivated? "Working on specific projects/topics that need help"
Mick (Australian)
Located: Sydney, New South Wales Interests: Family history, early Australian history, local history. Age and status: 41-60 (retired) Background: 12 hrs per week. Tags only occasionally. Why he does it: "I have recently retired from IT and thought that I could be of some assistance to the project. It benefits me and other people. It helps with family research" Why keep doing it: " I am interested in family history and early Australian history and newspapers are an excellent source of information" Will you carry on indefinitely? "Not sure" Is it addictive? "No" What would keep you motivated? "More papers added"
Catherine (Australian)
Located: Washington DC, USA (for 10 yrs) Interests: Doing something to help other people (not researching own family history) Age and status: 31-45 (working full-time) Background: 15 hrs per week. Never adds tags. Doesn't watch much TV. Also helping transcribe BMD (www.freebmd.org.uk). Why she does it: "I enjoy typing, want to do something useful and find the content fascinating" Why keep doing it: "To benefit others. The content and historical perspective is fascinating" Will you carry on indefinitely? "Yes" Is it addictive? "No, but I enjoy it" What would keep you motivated? "Being given ideas on topics to explore and correct"
Fay (Australian)
Located: Brisbane, Queensland Interests: Family history research. Age and status: 61-80 (retired) Background: Adds tags as well. Time spent varies Why she does it: "I need something to do in my spare time. It benefits me and others" Why keep doing it: "I enjoy the challenge and I enjoy doing it" Will you carry on indefinitely? "Yes" Is it addictive? "Yes, a challenge and I enjoy it" What would keep you motivated? "?"

When text correctors were asked what motivated them they suggested some of the usual motivating factors that motivate people to do any type of thing. John Jorgensen gives a good list of 21 motivating factors in his blog [6]. What text correctors said matched several of the points in his list, for example:

- **Pleasure** – The activity is pleasurable so creates eager and productive people.
- **Short and long term goals** – Many people have their own research goals. This guides their actions and the amount of work they do.
- **Concentrating on outcomes** – We made it clear what our desired overall outcome was (to improve the data quality for everyone) and left it up to the community to help achieve this in their own ways and they responded well.
- **Trust and respect** – We gave people a high level of trust and respect which motivated them to do their best.
- **Create challenges** – Lots of people liked the idea of having the opportunity to face new and difficult problems and were enthusiastic about it.

In addition text correctors identified these unique motivational factors:

- **Australian history** - Helping to provide an accurate record (sometimes linked to local history research)
- **Family names** - Helping myself and others find family names (sometimes linked to family history research)
- **Worthy cause -** I consider this a useful cause to volunteer for (linked to helping the Australian community, the Library, themselves)

Motivational tactics that we did not use but which text correctors said if we did would increase their motivation were:

- **Detailed instructions** - If you want a specific result, give us specific instructions. We will work better when we know exactly what's expected.
- **Team Spirit** - Create an online environment of camaraderie. We'll work more effectively when we feel like part of team or virtual community. We don't want to let others down.
- **Recognize achievement** - Make a point to recognize achievements one-on-one and also in group settings. We like to think we are being noticed and are making a difference. Show us how we fit into the big picture.
- **Raising the bar –** The more we do the more you should expect us to do. We'll do a lot more if you give us a lot more content. That would be our highest motivational factor.

We noticed in our communication with text correctors that a large proportion was family history researchers. These people are highly motivated to learn new skills in order to get the information they need. They also have a sense of responsibility towards other genealogists to help not only themselves but other people where possible. They realise the importance of information resources and researchers connecting with each other. The release of Australian Newspapers beta and the ability to text correct was immediately reported and discussed in forums internationally and this is how many of the users heard of the service.

Fig 11. Discussion on a genealogy forum about the release of Australian Newspapers beta and text correction

Although family history researchers are often retired and not of the web 2.0 generation they have come to grips with technology to get the information they need and to network online. There is a list circulating on genealogy blogs[7] of 'things done, or to do'. It is easy to see from the list the amount of new technology genealogists need to learn and their sense of community spirit. Here are some items from a list of 104 things:

- Join Facebook.
- Join the Genea-Bloggers Group on Facebook.
- Post messages on a surname message board.
- Respond to messages on a message board or forum.
- Join a Rootsweb mailing list.
- Create a family website.
- Create a genealogy blog.
- Upload a gedcom file to the internet.
- Google my name.
- Used a digital camera to "copy" photos or records.
- Be overwhelmed by available genealogy technology.
- Perform a random act of genealogical kindness.
- Perform a record lookup for someone else.
- Do indexing for Family Search Indexing or another genealogy project.
- Help someone find an ancestor using records you had never used for your own research.
- Teach someone else how to find their roots.
- Find a convict ancestor who was transported from the UK.
- Find a cousin in Australia (or other foreign country).

I have no doubt it won't be long before 'correct text in Australian newspapers' is added to the list if it has not been already. Genealogy is a passion not a hobby for these people.

It is also interesting to note that Australians have always had a strong sense of community and helping each other. This goes back to the pioneering spirit of first Australians and living in an environment that was and still is harsh and unforgiving, with natural disasters occurring in each state on a regular basis. The sense of making history, being a part of history and recording history is a very important to most Australians. So Australian genealogists are perhaps even more community spirited than other nations and perhaps this is why the text correction has been so successful.

For those seeking to do voluntary work, correcting Australian Newspapers is good because it requires very little prior skill or knowledge (just use of a keyboard and access to a computer). It can be done from home at times that suit the volunteer, and the volunteer can control how much work they do. Most consider it to be a good cause and can see that what they are doing will benefit Australians now and in the future including school children, academics, family and local history researchers. It gives structure and purpose within people's lives and those who have recently retired have commented

how beneficial this is in the transition between full-time work and retirement. With many Australians retiring in their fifties there is an increasing population willing and able to do work like this.

8. Requests from the Community for further involvement and engagement

The beta service clearly demonstrated to the ANDP team that users were more than ready and willing to volunteer and work together to enhance and improve the quality of the data. Many users were delighted with the service after waiting so long for newspapers to finally be digitised and accessible on the internet. Users also had imaginative ideas for how the service and their role in it could be improved. It was very clear that users wanted more engagement and involvement and in return they could offer us their services for free to improve and enrich the data which would benefit the entire community. This was a very exciting step forward for the Library and confirmed what we had learned in 2007 from the Picture Australia Flickr project [8]. In that project in the first year more than 23,000 digital photographs were submitted to Picture Australia, confirming that people want to contribute and provide additional value to Australia's collections, thereby offering a new level of interpretation to Australian history.

The requests for enhancements covered tags, OCR text correction, comments and social networking. The relationship between these four areas was close for some users but separate for others. The needs for all four areas are outlined below.

1) Text Correction
Users want to be able to correct text more easily and quickly. The current function is clunky and basic and there isn't a 'power user' mode. The ability to correct whole articles at a time and to be able to add missing lines into articles is needed. If text is hard to correct /read they could flag this to other users to get a second opinion. Knowing what articles have been corrected and seeing them flagged in the search results list as corrected would be useful. Some keen correctors who are not doing their own research would like to be given a list of articles to correct or topics/newspapers/geographic areas to cover. Correctors would like to be able to see where they rank in the entire text correcting community so they can compare their amount of text correction with others and see the big picture. Showing the top five text correctors on the home page is not enough. If top correctors could be acknowledged, rewarded or thanked in some way this would increase their motivation. This could include certificates, publicity, and public ranking tables.

Establishing guidelines for text correctors for example how to deal with italics, em dashes, pound symbols, missing lines is desirable. Reducing the time delay in text corrections being updated to the database would stop accidental double correction of the same article. Some correctors would like to help build up the community and take on a responsible role in the community for example to be able to moderate the work of others, or help develop guidelines. Advanced users have suggested being able to do global OCR text corrections or report global corrections needed for example change all occurrences of 'winks and spirits' to 'wines and spirits', change all occurrences of 'streefc' to street. Several users suggested post OCR processing is done on material before it is released to the public for correcting using the confusion matrix and language modeling technique that the team had identified but not yet tested or implemented.

2) Tagging
40,000 tags were added to articles in the first 6 months and users want guidelines for creation of tags – particularly for personal names. Many users feel that in order for tags to be useful to the overall community (rather than just the individual who entered them) there should be guidelines and users could volunteer to moderate tags, and then have the ability to edit tags for consistency. There is particular interest in having tags for names, places and events that will be useful for everyone.

Currently no searching of tags is enabled and the tag cloud is huge. Users thought it a priority that tags should be searchable and consistent. The usage of tags for full text material is looking very different to tagging for image collections mainly due to the use of personal names dominating everything else, and a usable 'cloud' not developing due to a tag rarely being used more than five times. Users also appear to be using tags to track their family research and save articles for later referral. The top tag in use is being used by a railway group to identify their research articles, so is useful for other members of that group but not to the general public. Users want the tags largely to be of benefit to everyone so think consistency, guidelines and moderation is the key to this.

3) Comments

Adding comments (annotations) is not a function that has been well used to date due to a lack of understanding about its purpose and usefulness and its relation to tags and corrected text. It was implemented for researchers as an annotation mechanism. Users have been using the comment facility to communicate with each other due to the lack of a forum or any other means of users to contact each other. Users also want to be able to put URL links into comments. At present the service does not support linking between articles in the database or external to the database and this is another way to enrich the data by adding a further layer. The comments are also not searchable or editable and users said this was important.

4) Social Networking

Some users would find it useful to have social networking features so that they can communicate, help, share, work and interact with each other. It would also make them feel part of the virtual community assisting with the 'greater good' as well as their own personal research. Suggestions were that users could have a user profile and choose whether this is public or private, and also to have a forum as part of the service. The online forum does not necessarily need to be managed by the Library. Three clear virtual communities have developed – taggers, text correctors, searchers. Some users are in all communities but most are not. In addition users who have been working as volunteer indexers in the hard copy world are wondering how they can contribute their skills in the online environment. As yet we are unsure.

5) Searching the enhanced data

The existing keyword searching takes place across the raw and corrected OCR text together. Now users have a greater understanding of how they are enhancing the data they want to be able to specifically search across layers for example on corrected text only, tags only, comments only and also combinations of these in conjunction with raw OCR text.

6) Personal activity logs

Being able to keep track of their activity and history (searches, text correction, tagging, adding comments) is important to many users. Being able to retrieve articles easily that they have previously used is a high priority and many users stated they are using tags to do this since there was no other way. Many users said having a complete record of their text correction activity was important to them (rather than just most recent activity) and knowing where they were in overall ranking also mattered.

9. Measuring increase in text accuracy of articles

When pages are supplied to the Library from the OCR contractor the accuracy of OCR text is not being measured due to difficulties with being able to do this [9]. However the page level OCR engine 'word confidence' figure is provided in the ALTO file. It is unknown how closely this OCR engine word confidence level may match the real word accuracy. The figure provided for word confidence is between 0 and 1, where 0 is considered very low confidence and 1 very high confidence. Of the

360,000 pages currently in the Australian Newspapers service the page confidence levels are as follows:

7.1% of pages have a confidence of < 0.5
18.5% of pages have a confidence of < 0.6
61.0% of pages have a confidence of < 0.7
87.1% of pages have a confidence of <0.8

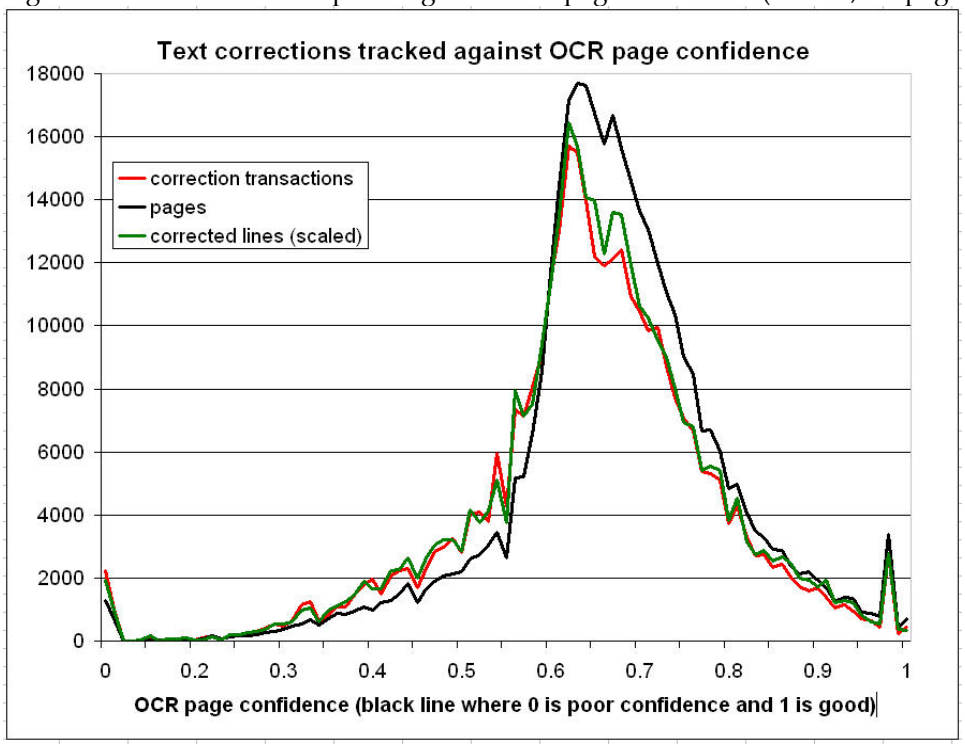
Therefore 42.5% of pages have an "average" confidence between 0.6 and 0.7

The ANDP team would like to be able to measure the improvement in overall accuracy of articles or the corpus as a whole now that public text correction is taking place. However this is still not possible to do due to lack of resource. It could be done simply by comparing words in an article with words in a dictionary before and then after text correction as a comparison.

In the meantime, as a matter of interest, all of the OCR text corrections in articles on a page (by line and by transaction i.e. clicking save) have been plotted against the existing OCR engine provided page confidence levels, for the entire 360,000 pages. We wanted to see if the lower the confidence the higher the correction transactions. The corrected lines have been scaled back by a factor of 7 so that they are more easily compared. The graph shows that corrections are "above" the page number curve for low confidence, and "below" the page number curve for high confidence, and about the same for mid-confidence pages (between about 0.6). So, lower confidence pages tend to attract slightly more corrections proportionally than higher confidence pages, but the effect isn't that pronounced. Pages with very low confidence of between 0.3 and 0.55 make up 10.6% of the corpus and they get 16.7% of the corrections, pages with high confidence between 0.75 and 1 make up 20.4% of the corpus and they get just 18.4% of the corrections. 69% of the corpus is of average confidence between 0.55 and 0.75 and these pages get 64% of the corrections.

It would be entirely feasible as some users have suggested to actively "dish up" the articles on pages that have a low page confidence if we wanted to target these for correction.

Fig 12: Text corrections compared against OCR page confidence (for 360,000 pages) March 2009



10. Future Potential

Text correction and tagging have proved to be successful and sustainable activities that enhance data. The Library is now considering the future potential of data enhancement and its implications for other services. Some key areas under discussion are:

- Resourcing ongoing IT development of Australian Newspapers to meet the needs of users and moving from a beta service to a version 1.
- Gaining a greater understanding of the relationships between enhancement techniques e.g. tagging, comments and text correction and commonalities'/differences between user activities in digital full-text collections to digital image collections.
- Application of similar web 2.0 functions in other National Library digital collections e.g. tagging, comments
- Integration of all National Library digital content under a single business IT architecture
- Developing user interfaces with the input of users that will fully meet user needs
- Sustaining ongoing engagement and involvement with users and virtual communities e.g. text correctors and genealogists, and increasing/maintaining motivation of users
- Actively promoting the service and harnessing the energy of volunteers on a mass scale
- Continuing to move towards 'mass digitisation' and expanding content available
- The technique/method/reality of post processing of raw OCR text using language modelling and confusion matrix, before the text is delivered to the public for text correction. This technique has the potential to increase the accuracy of text and that combined with public correction has a very good chance of achieving very accurate data indeed.

It seems likely that many more people would volunteer to correct text once the system for text correction is enhanced and publicity has taken place. This is based on observations from the experience of the FamilySearch Indexing Project [10] , where volunteers transcribe births, deaths and marriage records. User involvement in this project grew rapidly:

August 2005	FamilySearch Indexing on web introduced.
January 2006	2,004 online volunteers
January 2007	23,000 online volunteers
January 2009	160,000 online volunteers

In 2008 over 115 million names were indexed by volunteers of the FamilySearch Indexing project. Australia has over 21 million residents and over half of the households have access to the internet [11] so there is a large pool of potential volunteers.

Questions that the ANDP team do not have answers to at this stage are:

- Can collaborative text correction activity can be applied as successfully to other full text resources. Perhaps the high level of motivation only exists for newspapers because the core users are all genealogists?
- Is the activity sustainable long-term and if other libraries or large organisations like Google follow suit will activity decline or increase?
- Is there potential for organisations to join together and pool digital resource requiring text correction in a single place for the public to access?
- How can we measure overall improvement to OCR text when it is so difficult to measure accuracy levels of raw OCR text?
- Is the amount of text correction activity directly linked to the addition of new content?

- So far the benefits have been very cost effective and have outweighed the risks and issues but will this continue?
- Will moderation of text corrections be required?
- What is the relationship between full text tagging and text correction and what can we learn from this about our users?

The Library intends to build on the knowledge it is gaining from its experiences in this area. It intends to share the knowledge and software it has developed in the international arena, and to continue developing the Australian Newspapers service so that it continues to exceed the expectations of users.

11. Lessons Learnt

- OCR text correction involves people changing something from what they think is wrong to what they think is right therefore implying that incorrect data will be deleted, which concerns some people. It is better to refer to the process as text enhancement or enrichment since actually layers of corrections are all being saved with no data being deleted. The enhancement of the data also includes the addition of tags and comments. Users have a desire to search for data enhancements separately or in combination with original data.
- Just because it hasn't been done before doesn't mean it won't work and isn't a good idea, you won't know until you try it.
- Groups of people can achieve amazing things together and should not be under estimated.
- The public have more spare time than any single institution will ever have in people resource.
- When a Library appears to give up some control over its content it embodies in its public a great deal of trust. The trust is honoured and grows as users become actively involved with the content and feel a sense of responsibility working for the institution/service/common good.
- Control of data is not 'given up' if you consider the activities of users as 'enhancements', which do not compromise your original data and to the public are clearly viewable or searchable as separate data.
- Many people are highly motivated to do voluntary work and it gives them a sense of fulfilment and purpose.
- Imperfect data still holds huge value. A fragment of a newspaper page, or incorrect OCR that is freely available online is always better in the eyes of the users than getting nothing at all.
- With newspapers, quantity is more important than quality. Having the entire run of a paper even if it is in poor condition is better than having only the best pages available according to our users.
- Being transparent about processes and the development path increases the public's trust in us and their sense of knowing what's going on.
- Actively seeking feedback from the public and developing a prototype and beta version resulted in suggestions from users that were innovative, fresh, and viable and helped shape development of the service to better meet user needs.
- Genealogists are some of the most dedicated and motivated users, making great effort to find and learn how to use online resources. They network with each other in an online world that has no boundaries and it is important for them to undertake random or organised acts of kindness that will help other genealogists find names. They have taken to text correction like ducks to water.
- The social impact the service is having in the community and to individuals is equally as important to users as the improvement to the data. The Library has been unable to quantitatively measure either thing.
- Having no moderation and being open and free like the internet has raised many questions but has so far resulted in bringing more advantages than issues to the program.

- Improving quality of OCR text can still only be effectively done by manual intervention as far as we are aware. The Australian public are better at correcting text than OCR contractors based in developing countries since Australians speak English as their first language and can use context and historical knowledge to make educated guesses about what blurred or illegible words should actually be, especially where it is a place or personal name.

12. Conclusion

The Australian Newspapers beta service has clearly demonstrated that users want to engage and be involved with full text newspaper data in new and exciting ways. The use of web 2.0 technologies can enable this. Without publicity, 'how-to' tutorials or even a familiar and refined interface or concept, the service still rapidly harnessed an active group of users who are enthusiastically enhancing and improving the data by use of the text correction, tagging and comments functions. Users have demonstrated a willingness to work towards the 'common good', to volunteer their time, energy, skill, knowledge and ideas and to be involved long term in a program of national historic significance. The collaborative activity from this new community is enhancing the quality of the data and therefore the accuracy of full-text searching in a way that the National Library of Australia could never have achieved using its own resources alone.

The primary motivator for embarking upon collaborative text correction was to improve data quality and this has been a success. Another outcome is that the Library is now beginning to understand that engaging users in services, empowering them to make a difference, and building social networking communities is almost if not equally as important to the users as having high quality data. Giving control to users and entrusting the community to have such a crucial role in the development of a service helps build a dedicated, responsible, engaged and committed user base.

People want to work together to achieve amazing things and we are in a position to enable this by use of innovative technologies including not just text correction tools but social networking tools as well. As Barack Obama recently stated "We must not under-estimate the power of people who join together … they can accomplish amazing things". We have indeed proven that "many hands make light work".

13. Notes and References

[1] Australian Newspapers beta: http://ndpbeta.nla.gov.au

[2] Australian Newspapers Digitisation Program website: http://nla.gov.au/ndp

[3] Holley, Rose (2009). How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. D-Lib Magazine, March/April 2009, vol. 15 no 3/4. ISSN: 1082-9873 doi:10.1045/march2009-holley
URL: http://www.dlib.org/dlib/march09/holley/03holley.html

[4] ANPlan. Australian Newspaper Plan http://www.nla.gov.au/anplan

[5] Questionnaire for text correctors:
http://www.nla.gov.au/ndp/project_details/documents/ANDP_Questionsfortextcorrectorsv2.pdf

[6] Jorgensen, John (2007). 21 Proven Motivation Tactics. Published online in Pick the Brain, August 23 2007. http://www.pickthebrain.com/blog/21-proven-motivation-tactics/

[7] Genealogy Blogs with 'list of things to do'
http://blog.epcrowe.com/2009/01/07/104-genealogy-things-done-to-do-not-going-there/
http://geniaus.blogspot.com/2009/01/99-genealogy-things-meme.html

[8] Picture Australia Flickr project:
http://www.nla.gov.au/pub/gateways/issues/80/story01.html

http://www.nla.gov.au/pub/gateways/issues/90/story05.html

[9] Holley, Rose (2009). How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. D-Lib Magazine, March/April 2009, vol. 15 no 3/4. ISSN: 1082-9873 doi:10.1045/march2009-holley
URL: http://www.dlib.org/dlib/march09/holley/03holley.html

[10] FamilySearch Indexing http://www.familysearchindexing.org/home.jsf

[11] Australian Social Trends, 2008. 4102.0 - Internet Access at Home
http://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/4102.0Chapter10002008