



## The Australian Newspapers service and user interaction through text correction

**Pam Gatenby**

Assistant Director General, Collections Management  
National Library of Australia  
Canberra, Australia  
E-mail: [pgatenby@nla.gov.au](mailto:pgatenby@nla.gov.au)

**Meeting:** 99. ICADS

---

*WORLD LIBRARY AND INFORMATION CONGRESS: 75TH IFLA GENERAL CONFERENCE AND COUNCIL*  
23-27 August 2009, Milan, Italy  
<http://www.ifla.org/annual-conference/ifla75/index.htm>

---

### **Abstract:**

*The “Australian Newspapers” beta service was released to the public as a free service by the National Library of Australia in July 2008. From the outset of the newspaper digitisation project in January 2007 it was a priority for the Library to provide a high level of data quality. To improve the quality of text generated electronically using OCR software, we came up with the idea of enabling the public to assist with correcting the text. The beta service incorporating text correction functionality was released to the public with no promotion or publicity but the response was immediate. During the first 6 months several thousand members of the public found the beta service, largely through social network services, and a text correcting community of around 1300 people quickly developed and 2 million lines of text in 100,000 articles corrected in the first 6 months. The paper outlines some of the issues we considered in developing the user interaction features of the “Australian Newspapers” service, the user research we conducted into what motivates people to correct text, and some of the issues we have learnt. Based on the Library’s experience, text correction by the public has proved to be a successful and sustainable activity to enhance data quality.*

### **1. Introduction**

#### Outline of presentation

My presentation today is about the online Australian Newspaper service and the enthusiastic interaction with it by people who are correcting the electronically generated text.

I’ll outline some of the issues we considered in developing the service, the user research we have conducted, the activity of users, what motivates them, and some of the lessons we have learnt.

Before I start, I would like to acknowledge the work of the Australian Newspaper Project Manager, Rose Holley in carrying out and analysing the research I will refer to in my talk. Also, my presentation draws heavily on a detailed paper Rose presented at a conference held by the IMPACT project in the Hague in April 2009 on OCR in mass digitisation. Rose's paper, titled *Many hands make light work*, is about our experience with user interaction with our online newspapers service and it is available on the National Library's web site.

As well as text correction, the Australian Newspaper service supports user tagging and the addition of comments but I won't be discussing these today – if you are interested in our experience in these areas, this is covered in Rose's paper.

### The value of digitised newspapers

It goes without saying that newspapers are an invaluable, unique source of information about the history and culture of a nation – but in their original print form they are tedious and difficult to use. By digitising newspapers and making the content searchable online, their extraordinary power to support a wide range of research and to seduce those with a curiosity about the past, is unleashed. No topic is out-of-scope, as Angus Trumble, Australian scholar and author of a number of books and articles on an impressively wide range of topics, suggests in his blog *The Trumble Diaries*.

Writing with reference to the online Australian Newspapers service, he says:

“I cannot think of any resource for the study of Australian history that has in my lifetime come as close to providing almost overnight such an enormous sweep of access ...”

“... if you want to monitor the imminence of the colonial reputation of Queen Maud of Norway, or find and document the career of a particular racehorse, or study the incoming shipments to the Sydney agents of certain Paris milliners ... or find someone who has hitherto been completely invisible to most if not all other published sources—[the] *Australian Newspapers* ... [service] will henceforth need to be your first stop”.

We know that researchers from different fields of enquiry have discovered the service and are using it to assist with their research - for instance, one of our text correctors commented that the service “is the best thing that has ever happened to me in twenty years of family history research”; another academic is using it to discover words that originated in Australia; Professor Ian Fraser, immunologist, medical researcher and Australian of the Year in 2006, is using the service to research the history of influenza in Australia; and a government funded research project with which we are collaborating, is using the service to study climate change and weather in Australia up to 1900.

## **2. Background of the Australian Newspapers Digitisation Program**

The *Australian Newspapers* beta service was released to the public as a free service by the National Library of Australia in July 2008. It is our good news story in innovative service delivery and in demonstrating the enormous potential digitisation and social networking technologies have for connecting the public with library collections.

During stage 1 which will conclude in mid-2011, the service aims to provide online access to a major daily newspaper published in each of the eight states and territories up to 1954, when

copyright takes effect. That will equal about 40 million articles (4.4 million newspaper pages). Attention will then turn to regional and provincial titles. Content is being added progressively to the service – to date around 5.8 million articles (538,500 newspaper pages) are available for searching.

The program to digitise the newspapers included in the service is managed by the National Library and involves collaboration with the Australian state and territory libraries. Scanning and Optical Character Recognition processing is carried out offsite on contract but quality assessment is carried out in-house.

The software supporting the Search and Delivery System (which I will show you soon) and that used for workflow management and quality assurance was developed by the Library. We are happy to share the code for the Search and Delivery System and have made it available through our open source project repository at <http://code.nla.gov.au>.

The public website for the newspaper digitisation program (<http://www.nla.gov.au/ndp>) includes a comprehensive range of information; for instance, standards and workflows used; systems architecture, coverage and progress; development plans; public presentations and outcomes of research. It is a truly remarkable source of information and I would recommend it to those wanting to know more about our program or about newspaper digitisation generally.

### **3. Some features of the Australian Newspapers service**

Before I go any further I'll give you a quick overview of how the Australian Newspapers service works. This is the Home Page. From here, you can search articles using keywords, phrases and dates or search for a particular issue by selecting the newspaper title and date (year, month and day). You can also browse newspaper issues – by selecting the Browse tab - by title, state of publication, date or category of article – that is, advertising, news, family notices, or detailed lists. You will also notice at the foot of the Home Page users have the option to sign in, the 5 top text correctors are identified, and recent comments and tags are shown. Having conducted a search for an article the results returned are sorted by relevance and can be sorted by date. You can also choose to refine the reach results by newspaper title, or by category of article, whether or not it is illustrated, by decade and by word count.

When you select an article to read it is presented in the context of the page, with search terms highlighted and with the user interaction functions – tagging, text correcting and comments – presented on the left side. I'll return to this later. Where a tag has been assigned to an article it appears as a link on the article and the tag can be searched by activating the link. If the article has been corrected the date of last correction is also shown. Note also that each article is assigned a Permanent Identifier (PI) for citation purposes.

### **4. Involving users: text correction**

From the outset of the newspaper digitisation project in January 2007 it was a priority for the Library to provide a high level of data quality. However, we were aware that this would be a challenge, as while OCR processing works well for documents with a modern, consistent typeface and standard format, the nature of historic newspapers means that OCR accuracy can be low. This is because they have varying fonts and print quality as well as high article density with little white space between text. After experimenting with different approaches to

achieving better OCR results, we concluded that we would not be able to reach the quality of OCR text that we desired - so we came up with the idea of enabling the public to assist with correcting the text.

Some of the benefits of allowing the public to alter OCR text that we identified were:

- improvement to data quality and keyword searching for all users;
- meeting user expectations of a quality service;
- building new virtual user communities and social networks based on our collection;
- gaining experience in innovative use of web 2.0 technologies and assessing their potential for other discovery services.

Some of the risks identified were:

- it was new territory: there were no models to follow and we were not sure how difficult technically it would be;
- potential vandalism of text;
- large amounts of text correction activity could compromise service performance;
- users might not take up text correction so development time would be wasted.

As well as identifying benefits and risks, there were a number of management and technical issues we needed to consider – for instance, whether moderation of text correction was required; how to monitor user activity; how to manage the changed text; and if vandalism occurred, how to roll back to previous versions.

In order to test our ideas concerning text correction we released a prototype of the search and delivery system to our state library partners and then undertook user testing on the beta system that was developed incorporating their feedback on the prototype.

The beta service was released to the public on 25 July 2008 with no promotion or publicity but the response was immediate. During the first 6 months several thousand members of the public found the beta service, largely through social network services especially genealogy forums, and became active searchers, taggers and text correctors. A text correcting community of around 1300 people quickly developed and 2 million lines of text in 100,000 articles were corrected in the first 6 months. Feedback from users was actively sought following release of the beta service and after the first six months it was compiled, analysed and made public.

Before I report on the feedback, I'd like to show you how text correction works.

This slide shows the text of an article with user interaction functions displayed in the box on the left. In the Electronically Translated Text box there is information on why mistakes occur and how to correct them and a link to previous changes made on the article is provided.

To fix the errors in the article – shown by the red circle – you click on *Fix this text* which takes you to an editing screen where you can edit the OCR text of the article line by line. You can also add missing text to lines but the amount you can add is restricted to deter vandalism. As you move from line to line the first word of the corresponding line in the original text is highlighted so you know where you are. The control buttons are also shown on this slide.

The text correction functionality shown here is an improvement on what was offered in the beta service – it is the first enhancement we have made in response to user feedback.

## 5. Feedback on the beta service

Feedback on the beta service during its first 6 months was received from more than 600 individual users who generally made several suggestions. The response was overwhelmingly positive with many stating that the service already exceeded their expectations. From the feedback we learnt for instance, that:

- 49% of registered users were correcting text and 78% of users were based in Australia but there was also a growing international community; and
- many users found the text correction rewarding and/or addictive - one genealogist in need of serious help described the effect in a posting to the Genealogy.com forum in the following way:

*“While going through a whole month in a slightly obsessive crazed mind searching Australian newspapers beta online, I just realised the kilos I’ve stacked on in just one month. I can’t seem to snap out of it; from dawn to dusk I seem to be in this website craving to find more on my ancestors – all the gritty stories. Housework seems to have taken a backburner and meals are starting to come out of cans ... is there an AA for genealogy junkies?”*

We also learnt from the feedback that:

- users were actively correcting much more text than we had expected
- the top ten text correctors were correcting significantly more text than all other users with some individuals spending up to 45 hours a week on the activity;
- for many users, imperfect data still holds huge value - a fragment of a newspaper page, or incorrect OCR is better than getting nothing at all; and
- 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 some of our users.

## 6. What motivates text correctors?

In considering what motivates correctors, the key factors that emerged from the feedback were the desire to do something worthwhile that is interesting and which benefits others, and public recognition. A definite desire to be part of a team or virtual community and to support user profiles also emerged.

Some specific suggestions for maintaining motivation included:

- the ability to correct text more easily and quickly – and keen correctors would like to be given a list of articles to correct or topics to cover;
- providing support for personal activity logs so users can keep track of their activity on the service;
- ranking correctors so they can compare their amount of correcting with others - from the outset we have featured a correctors’ hall of fame on the homepage of the service which showcases our top 5 correctors, but it seems this is not enough;

- providing detailed instructions or guidelines for correction;
- the ability to do global text corrections – for instance, “winks and spirits” to “wines and spirits”;
- some really keen ones want to be able to moderate the work of others and to identify work that is required;
- being kept informed about developments with the service; and
- adding social networking features so users can communicate and interact with each other – the idea of a user forum came up alot.

Following on from feedback received on the beta service, we were interested to find out more about the text correctors so we sent a questionnaire to the people who were the top five over the previous 6 months. (The same people had remained in the top five each month.) All responded enthusiastically. This table provides some information about them.

	<b>Julie</b>	<b>Lyn and Maurice</b>	<b>Mick</b>	<b>Catherine</b>	<b>Fay</b>
<b>Interests</b>	Family/local history	Family/local history; shipping	Family history/early Aust. history	Doing something to help other people	Family history research
<b>Age and status</b>	31-45; stay at home mum	55-62 (retired couple)	41-60 (retired)	31-45 (working full-time)	61-80 (retired)
<b>Activity</b>	15-45 hours correcting per week	15 hours per week	12 hours per week	15 house per week	Varies
<b>Why do it?</b>	Enjoys it. A great way to learn about history. A service to the community	Sick of doing housework!	It benefits me and other people	Want to do something useful; finds the content fascinating	Need something to do in my spare time. It benefits me and others and I enjoy the challenge
<b>Will you continue?</b>	Yes – a “must do” mission”	Yes – it helps us and other people	Not sure	Yes	Yes
<b>What would keep you motivated?</b>	More papers added	Working on specific projects/topics	More papers added	Being given ideas on topics to correct	?

It is interesting to note that these dedicated text correctors all identify the common good as a key motivating factor and between them identified the same two things that would keep them motivated - the addition of more papers and working on specific topics that need fixing!

## 7. Some comments from our users

To give you a feel for the reception of the Australian Newspapers beta service by people using it, I have selected some representative comments from the hundreds received.

'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!'
'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.'
'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 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.'
'Wow – well I got sucked in! I can see why everyone is editing the OCR text... its compulsive!'
'The decision to let members of the public edit & correct articles must have been a difficult one, but it's a great idea.'
"Thank you! You lot are so cool!"

## 8. Some statistics on user contribution to text correction

As I've mentioned, the Australian Newspapers service has not been formally launched or promoted yet – but we have plans to do so soon – but this does not seem to have had a negative impact on use of the service. These figures show the level of activity at 4 August 2009, one year after the beta service was released.

	At 4 August 2009
Number of pages in service	538,334
Number of articles in service	5.8 million
Lines of text corrected	4.7 million
Number of articles corrected	216,093
Top corrector	237,901 lines of text
Number of comments added	3,441
Number of tags added	105,028
Unique visitors to site	492,000
Total keyword searches	3.3 million
Source of searches	55% from Google; 21% on the service; 10% from NLA web site

## **9. What's next?**

Based on our experience to date, text correction has proved to be a successful and sustainable activity to enhance data. We are now considering the future potential of data enhancement and its implications for other discovery services. Two issues that we are particularly interested in exploring further are the relationships between enhancement techniques, e.g. tagging, comments and text correction, and the commonalities and differences between user activities in digital full-text collections and digital image collections. We would also like to work out how to measure the extent and success of user interaction activities. In the meantime, we will continue to develop the Australian Newspapers service as resources permit and to implement other suggestions received from our users.

## **10. Conclusion**

In conclusion, the Australian Newspapers service has clearly demonstrated that users want to engage with full text newspapers in new and exciting ways that web 2.0 technologies can enable. Without publicity, the service rapidly harnessed an active group of users who are enthusiastically enhancing and improving the data through text correction. Users have demonstrated a willingness to work towards the 'common good', to volunteer their time, 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. Having no moderation and being open and free like the internet has not presented any issues for the service so far.

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 and engaged user base - a major asset for a library wanting to remain relevant and visible in the digital age.