

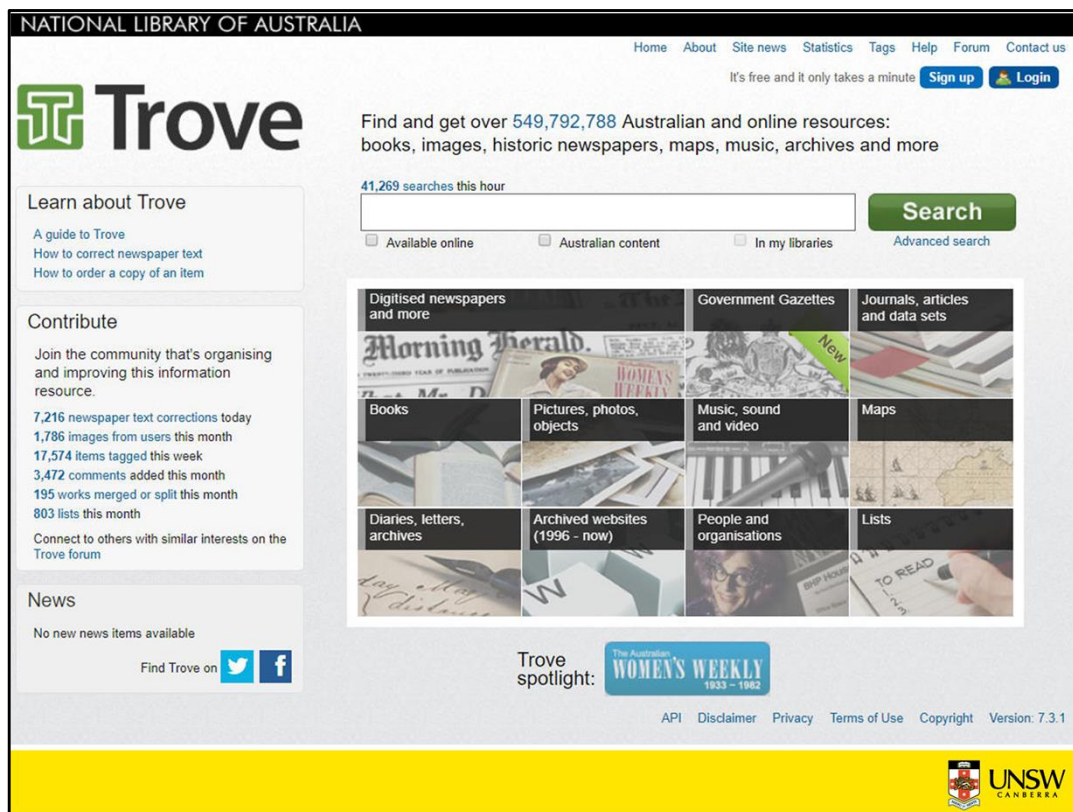


## **Crowdsourcing Based Curation and User Engagement in Digital Library Design**

**Rose Holley**  
Special Collections Curator  
[r.holley@adfa.edu.au](mailto:r.holley@adfa.edu.au)

25-27 October 2017, New Delhi

**UNESCO-NATIONAL DIGITAL LIBRARY OF INDIA  
INTERNATIONAL WORKSHOP ON KNOWLEDGE  
ENGINEERING FOR DIGITAL LIBRARY DESIGN**



This is the current interface of Trove. It has gone through some design changes in the last few years, but in essence is little different to the original design. The backbone of Trove is the union library catalogue of Australia which has physical and digital holdings from over 2,000 libraries in Australia. Trove also harvests data resources from other institutions such as Archives, Galleries and Museums and our national TV broadcaster the ABC, making this a really useful Australian information portal. But what makes Trove really well used by millions of Australians is the main body of content which is 211 million digitised Australian historic newspaper articles from 1,200 different newspapers from around Australia, dating from 1803 to 1955. The newspaper zone is also the area in which the majority of the user curation is occurring, and which has captured the interest of libraries internationally due to the public text correction activity.

# Digitised Australian Historic Newspapers

The screenshot displays the Trove website interface. At the top, it says 'NATIONAL LIBRARY OF AUSTRALIA'. Below that is the Trove logo and a search bar containing 'unesco new delhi'. The breadcrumb trail reads: 'Newspapers / Browse / The Canberra Times (ACT : 1926 - 1995) / Wed 7 Nov 1956 / Page 12 / India Rushes New Buildings for Unesco'. On the left, there is a sidebar with various icons for text correction, comments, tags, and sharing. The main content area shows a newspaper clipping with the headline 'India Rushes New Buildings for Unesco'. The article text describes the construction of modern buildings in New Delhi for the UNESCO conference. To the right of the main article, there are other newspaper snippets, including one titled 'Migrant Fined For Fighting' and another for 'A. G. MILDREN'. At the bottom right, there is a small advertisement for a used car.

This slide shows the interface for Australian Newspapers which is a sub interface within Trove. I have searched for an article on Unesco New Delhi and I am now seeing the article in the newspaper viewer. You can see on the left hand side the OCR'd text with user options and on the right the image of the newspaper page. Both text size and image can be increased in size.

A user can undertake a number of activities on the article by hovering their mouse over the icons on the left hand side, including correction of text by clicking the 'fix this text' button, adding comments, tags, adding the article to a list, sharing or downloading the article and also ordering a high resolution copy. I will not go into detail about this user interface because you can try it yourself. It is really easy to use.

Image zooming and panning was based on an open-source library inspired by the Google map viewer, which was the idea of Kent Fitch of Project Computing, the reliable system architect and programmer we worked with.

## Newspapers project team 2008

Rose Holley : Project  
Manager IT and  
Digitisation



Mark Raadgever: Quality  
Assurance and  
Stakeholder Engagement



Kent Fitch: Project  
Computing  
System Architect,  
Programmer



Alexi Paschalidis: Oxide  
Interactive. User interface  
design and testing



Newspaper beta users and text correctors: functionality and design



Ninh Nguyen:  
Programmer



Brownyn Lee:  
Business Analyst



Our team was very small consisting of only 6 people, with the key IT roles being external contractors.

# Led by users: Functionality and interface design

AUSTRALIAN  
NEWSPAPERS beta

[Home](#) [About Us](#) [Browse](#) [Help](#) [Feedback](#) [Login / Signup](#)

## Feedback

To provide feedback about Australian Newspapers Beta, fill in our feedback form [here](#) (new window).

Please note that we are unable to provide individual responses to questions or comments entered in this form.

If you have an enquiry that requires a response please use the form located on the ANDP website at:  
[http://www.nla.gov.au/ndp/contact\\_us/](http://www.nla.gov.au/ndp/contact_us/) (new window).



Developed by the National Library of Australia as part of the [Australian Newspapers Digitisation Program](#)

You need customer insight to craft solutions that your audience wants to use. Knowing your audiences' expectations and aspirations is key to developing useful and usable experiences that'll keep them coming back. Have the ability to learn from your customers before, during and after development.....



Right up front we made two decisions, firstly to have the system live on the National Library website as we developed it, so we branded it 'beta', and secondly we wanted direct and active involvement from the users on its development. In fact for the first three years all development was actually led 100% by our users feedback.

What we did was to put a notice on our website 'testers wanted'. We expected perhaps 20 people to contact us, but after only a few days we had to take the notice down because so many people had emailed us. We also did not expect the use to take off as quickly as it did, so we had thousands of live users at a very early stage.

The way we engaged with users was to have a prominent link on every page saying 'contact us with feedback'.

## Enhancement requests 2008

| Task   | Summary   | Frequency<br>(1=little, 5=most) | Expected Complexity<br>(1=easy, 5=hard) | Score |
|--|---|---------------------------------|---|-------|
| Correct a few lines of OCR text in an article and save the corrections.  | Correcting OCR                                      | 3                               | 5                                       | 15    |
| When choosing to correct OCR text in an article (or add a note or tag) and not signed in; before being able to make the changes, be prompted to sign in or register (optional), choose to sign in then complete selected   | Being prompted to log in                            | 3                               | 5                                       | 15    |
| Add a comment (note) to an article.  | Adding a comment                                    | 3                               | 4                                       | 12    |
| When choosing to correct OCR text in an article (or add a note or tag) and not signed in; before being able to make the changes, be prompted to sign in or register (optional), decline to sign in or register, prove you are not a machine (using Captcha technology), then | Decline sign in and perform a Captcha               | 2                               | 5                                       | 10    |
| Search for articles from a single newspaper title (by using Advanced search, Refine search, search from newspaper title page or Simple search with Did you   | Search for articles inside a single newspaper title | 3                               | 3                                       | 9     |
| While viewing an article, save and print the article   | Save image/PDF                                      | 3                               | 3                                       | 9     |
| While viewing an article, save and print the PDF version.  | "   | 3                               | 3                                       | 9     |
| View tags that have been added to an article.  | View tags   | 3                               | 3                                       | 9     |
| Find issues published on a certain date (using Browse) and view one of them.   | Find issue published on a certain date              | 4                               | 2                                       | 8     |



We did not specifically ask for feedback on functionality or design, but most of the feedback was around these two topics. If we had more than 10 users within a few days ask us for the same functionality or interface design then we implemented the change immediately, for example to change the size of the font, or ability to add a missing line.

Most of the users were astounded by the fact their requests were implemented, and that we responded to them by email directly and immediately, and this seemed to engage them to a much higher level than we had expected. So at this point we had no software development plan and were making many small incremental changes on a daily basis led by users. You have to remember that nothing like this had ever been done before so we were feeling our way in the dark.

Quite quickly the users expectations shifted and instead of asking things like “can I do this?” they said “where do I do this”? Users initially seemed to have their functionality and design expectations based on Google as the benchmark, and when we surpassed Google they got really excited. We encouraged users not to place limits on their expectations and they quickly started to ask us for all kinds of things.

It was a very exciting time working on this and we knew we were onto something really important. I still consider it a great honour that I was able to have so much freedom to run with innovative ideas, make decisions, and have such excellent

customer focused people on my team to create the Australian Newspapers Service, which eventually morphed into Trove.

## OVERVIEW

---

### Prototyping

We'll be testing our prototypes for layout and functionality with HTML wireframe mockups. We will include elements of graphic design as the prototype evolves toward the final product. Halfway through, we will implement a more refined visual style and continue to test.

### User Sample

5 users will be tested during each round of testing. Users are recruited from the general public.

- Male or Female
- 18-60 years old
- Use the internet once a week or more

### Schedule

The number of iterations is limited by time - 4 days of testing have been allocated. Users have been scheduled 1 hour apart.

- **Thu 29 May** - Round 1 (5 users)
- **Tue 3 Jun** - Round 2 (5 users)
- **Fri 6 Jun** - Round 3 (5 users)
- **Wed 11 Jun** - Round 4 (5 users)

We also engaged two young students just out of University who had formed a web design company Oxide Interactive to undertake user case studies, prototyping, interface design and usability testing. There were four rounds of testing with five different users each time.



## National Library of Australia

25/7/08 - Newspapers Beta Interface Recommendations

**AUSTRALIAN NEWSPAPERS BETA**  
Explore our historical collection of Australian Newspapers spanning 1803 to 1954

Home | About Us | Browse | Help | Login

**FIND AN ARTICLE**  
Search Articles

**FIND AN ISSUE**

by Title

1. Colonial Times
2. Colonial Times and Tasmanian...
3. Hobart Town Daily Mercury
4. Hobart Town Gazette
5. Northern Territory Times

Show all titles

by Date

1804

| JAN | FEB | MAR | APR |
|-----|-----|-----|-----|
| 1   | 2   | 3   | 4   |
| 5   | 6   | 7   | 8   |
| 9   | 10  | 11  | 12  |
| 13  | 14  | 15  | 16  |
| 17  | 18  | 19  | 20  |
| 21  | 22  | 23  | 24  |
| 25  | 26  | 27  | 28  |
| 29  | 30  | 31  |     |

by State

NT WA SA TAS QLD NSW ACT VIC

USER LOGIN  
username:   
password:   
Login

TOP TEXT-CORRECTORS

1. dcho@nla.gov.au (2)
2. thong@nla.gov.au (1)
3. krt@nla.gov.au (0)
4. gpa@nla.gov.au (0)
5. alme@oxideinteractive.com.au (7)

RECENT NOTES

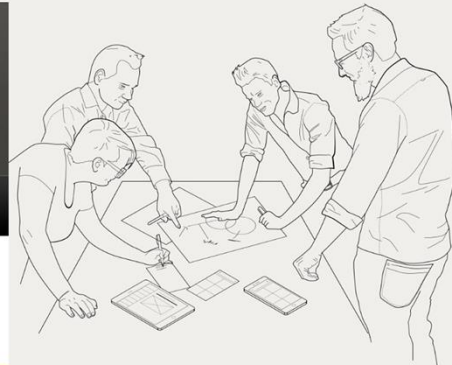
How times have changed!  
newspaper 20/5/08 12:28pm by gpa@nla.gov.au

How times have changed!  
newspaper 20/5/08 12:28pm by gpa@nla.gov.au

RECENT TAGS

Tornado Shark attack Urban Design  
Alternative energy Electric Cars  
Steam cars WWII Cars  
Baboon Eels & Broom

User research, prototyping and usability analysis of the Library's pioneering digital exploration of historical Australian newspapers.



Oxide basically just took a group of people off the streets in Canberra and sat them down in front of prototypes of the site and asked them to do various things like search for a newspaper, correct the text etc. and then modified the interface based on their feedback. At this time there was a strong expectation from the librarians who were contributing newspapers to the new service, that the functionality and interface design should be decided by librarians rather than random people off the street or real users, since they thought librarians would know better. We strongly resisted this and continued using our real users to lead us in development, which is why I think the interface was such a great success.

## Clay Shirky 'Cognitive Surplus'

*Here Comes Everybody: The Power of Organizing Without Organizations*



The free time that people have on their hands to engage in collaborative activities, specially as applies to web 2.0.

Wikipedia is an example of wide-scale deployment of cognitive surplus.

*“The Internet will transform the relationship between ordinary individuals and large, hierarchical institutions”*

*“The internet runs on love”*



People asked us why anyone would want to be involved in curating content when not being paid for it and I have two answers to that question.

Firstly expectations of users have changed in the digital age. With online technologies users can be easily enabled to help us gather, create, curate and add value to our resources, and they want to. Why shouldn't they? The gatekeeper role of the library has changed into that of enabler, giving people freedom to do what they want with our resources, add value and use them in new ways.

Secondly I was greatly inspired by Clay Shirky's book published in 2008 *'Here Comes Everybody'* which explains his idea that everyone has some 'cognitive surplus'. This is people's leisure time, most often spent watching TV, that could be harnessed and used instead on doing things that matter, for the common good, that require a bit of brainwork or creativity. The internet is the enabler for this activity allowing anyone to build online communities that can come together to achieve amazing things. He says *“As the Internet radically reduces the costs of collective action for everyone, it will transform the relationship between ordinary individuals and the large, hierarchical institutions”*. He also coined the phrase *“the internet runs on love”*. His idea of cognitive surplus certainly proved to be correct in 2008 with the involvement of millions of individuals in the creation of Wikipedia, the citizen science projects of Zooniverse and Australian Newspapers.

## Motivations of Newspaper Text Correctors

- ✓ Pleasure and fun
- ✓ Worthy cause
- ✓ It's a big challenge
- ✓ You've given us a high level of trust and respect to do this
- ✓ We understood the big picture and the expected outcome and wanted to help
- ✓ Interesting - It's about our Australian History
- ✓ Interesting – It's about my family
- ✓ Interesting – It's in my research or learning area

*'Many Hands Make Light Work' by Rose Holley 2009*



In 2009 I undertook some research on our text correctors behaviours and motivations, which I published in a report 'Many Hands Make Light Work'. You can see on this slide some of the main reasons why people wanted to be involved. The most cited reason being it is pleasure and fun.

## Motivations of Wikipedians

- ✓ **Fun** – enjoying the activity
- ✓ **Ideology** – expressing support for the underlying ideology of the activity (e.g. the belief that knowledge should be free)
- ✓ **Values** – expressing values to do with altruism and helping others
- ✓ **Social** – engaging with friends, taking part in activities viewed favourably by others
- ✓ **Understanding** – expanding knowledge through activities
- ✓ **Career** – gaining work experience and skills

*'What motivates Wikipedians' by Oded Nov 2007*



Interestingly what I found out about motivations correlated closely to a study carried out by Oded Nov in 2007 on the motivations of people contributing to Wikipedia. Again the top reason given for editing and creating articles was fun.

People also have a strong desire to help with something that is for the common good, in this case free knowledge, and in the case of the newspapers improving the word searching, especially so that family names can be found. Over 50% of the corrections are done on names appearing in the births, deaths and marriages sections and shipping news.

It's important to understand motivations of people so that you can create something where people will want to join in.

## Increasing motivation

- **Recognising achievements** of the group and individuals, acknowledgements and rewards (Hall of Fame, Progress Chart)
- **Team spirit** – online camaraderie – able to see and talk to others online
- **Raising the bar** – giving us more work and challenges
- **Guidance** - detailed instructions, guided topics, short term goals
- **Treat your top users well** they achieve more than 50% of the work




### Understand behaviours of volunteers

e.g. duration of work, time of work, engagement with others, personal goals



I have also researched ways that user motivation can be maintained or increased. This slide shows the key things that increase motivation of users. These include recognising their achievements, developing camaraderie and team spirit with online communication mechanisms, giving them more work, and guiding them in targeted projects or short term goals. It is very important to remember that a tiny percentage of all your users – the top 100, undertake the large majority of the work. Once these users are identified you need to pay close attention to any feedback they give you or functionality they ask for since these people are the kingpins and essential to the success of your service. In the next few slides I will give you some examples of the functionality our top users asked us to implement.

# Leader board – hall of fame – ranking tables



[Home](#) [About](#) [Site news](#) [Statistics](#) [Tags](#) [Help](#) [Forum](#) [Contact us](#) [Login](#)

[Newspapers:](#) / [Browse](#)

## Text correction hall of fame

Automatically extracting text from scans of old newspapers and magazines is extremely challenging. Although the best available Optical Character Recognition (OCR) software has been used, the condition of the images it has to process combined with the frequently small fonts used means that many errors of interpretation are made.

Thankfully, many people have stepped in to correct the text and in so doing have made a wonderful contribution to this resource, and helped to improve search results. The following table lists by month the people who've corrected the most lines of text.


**Total number of lines corrected for all time : 246,425,980**

### Leaderboard

All time
Newspaper articles
View

Total number of correctors **49,191**

| Rank | Username                      | Date started | Lines corrected |
|------|-------------------------------|--------------|-----------------|
| 1    | <a href="#">JohnWarren</a>    | Sept 2009    | 4,690,014       |
| 2    | <a href="#">noelwoodhouse</a> | April 2012   | 3,170,129       |
| 3    | <a href="#">NeilHamilton</a>  | Nov 2011     | 3,156,603       |
| 4    | <a href="#">John.F.Hall</a>   | Sep 2008     | 2,420,787       |
| 5    | <a href="#">annmanley</a>     | July 2009    | 2,277,278       |
| 6    | <a href="#">DonnaTelfer</a>   | Sep 2010     | 2,004,300       |
| 7    | <a href="#">Rhonda.M</a>      | Aug 2008     | 1,748,565       |
| 8    | <a href="#">maurielyn</a>     | Aug 2008     | 1,712,313       |
| 9    | <a href="#">C.Scheikowski</a> | April 2009   | 1,703,316       |



In our first beta version we only had a list of the top 5 correctors. Very quickly many users requested a complete ranking table because they wanted to see where they were in the overall big picture. Stakeholders thought we had started to make the text correcting deliberately competitive, but we had not, the rankings were entirely in response to our users asking for them.

It is interesting to note here, that there is a massive loyalty, since of the Top 10 correctors you can see that three have been contributing since first year of launch, (which is now 9 years ago) and three since the second year of launch. All except the Frankston Library joint account have been helping at consistently high levels for at least 6 years.

The ranking table also shows the big picture, which our users wanted to know. 246 million lines in total have been corrected. The top users have corrected 1.6 to 4.6 million lines total. It would be very hard now for a new person to get into the top 10 since these people have years behind them. There are about 50,000 registered correctors.

# Public user profiles

## Information about Trove user: **anmanley**

[View user profile in the Trove forum](#)


Tags [Recent comments](#) [Text corrections](#) [Recent merge/splits](#) [Lists](#)

### Tags

Display options

|             |                                 |                   |                             |                         |
|-------------|---------------------------------|-------------------|-----------------------------|-------------------------|
| Containing: | Display only:<br>50<br>top tags | Added:<br>anytime | By:<br>user:public:anmanley | <a href="#">Refresh</a> |
|-------------|---------------------------------|-------------------|-----------------------------|-------------------------|

**Advertising** also advertising also news American news Amy shipwreck Annie M Miller shipwreck Aramac  
Shipwreck boxing Catterthun Dayspring shipwreck **divorce** Dudley Calamity **Elingamite Shipwreck**  
**English News** Ethel Harris Exeter railway disaster **Family Notices** Fifeshire  
shipwreck Frederick Septimus Kelly **Funeral** Granville Train Disaster illegible Inquest **Matrimony merged**  
**columns** mostly illegible Murulla railway disaster Nemesis shipwreck **News** **Obituary** partly illegible  
**poetry** Redfern Railway Disaster Rodney launch capsized Roll of Honor s s Tasmania  
Shipwreck **shark attack** **shipwreck** slightly illegible Stockton Calamity Sumatra shipwreck swimming Sydenham  
Railway Accident **Tarana Railway Disaster** Text transcribed in comments **Thomas Ranken** Trevesa shipwreck  
Wahine Ferry Disaster **Wairarapa Shipwreck** Yacht Ripple capsized



The users wanted more than a ranking list, they wanted to be able to see the profile and interests of the users around them. Each user can see anyone else's activity including tags, comments, text corrections and lists. For users who have a tag cloud it is clear to see their interests. In this example the user is interested in shipwrecks, railway disasters, obituaries, family notices and poetry.

## Information about Trove user: annmanley

[View user profile in the Trove forum](#)

[Tags](#) [Recent comments](#) [Text corrections](#) [Recent merge/splits](#) [Lists](#)

### Text corrections

[Everything](#) [Newspapers](#) [Government Gazettes](#)

**2,277,348 line(s) corrected.**

#### Corrections by month

|               |        |
|---------------|--------|
| August 2017   | 16,307 |
| July 2017     | 15,743 |
| June 2017     | 5,035  |
| May 2017      | 2,702  |
| April 2017    | 8,956  |
| March 2017    | 2,498  |
| February 2017 | 857    |
| December 2016 | 563    |
| November 2016 | 3,971  |
| October 2016  | 10,572 |

### Hall o' fame ranking

| Rank | Corrector                     | Lines corrected |
|------|-------------------------------|-----------------|
| 1    | <a href="#">JohnWarren</a>    | 4,688,635       |
| 2    | <a href="#">noelwoodhouse</a> | 3,170,129       |
| 3    | <a href="#">NeilHamilton</a>  | 3,154,630       |
| 4    | <a href="#">John.F.Hall</a>   | 2,416,575       |
| 5    | <a href="#">annmanley</a>     | 2,277,348       |
| 6    | <a href="#">DonnaTelfer</a>   | 2,004,158       |
| 7    | <a href="#">Rhonda.M</a>      | 1,747,656       |



There is also more information and statistics on text correction in the profile, you can for example see how many lines per month this user has been doing, and that she now holds ranking 5 in the hall of fame, with nearly 3 million lines corrected overall.

This user is a champion for the service and is often out and about giving talks on the content of newspapers, search techniques and text correction.



Tags Recent comments Text corrections Recent merge/splits Lists

### Recent comments

Display options

Added start date: 21/07/2017 Added end date: 21/10/2017 Material type: Any type By: user:public:annr Refresh

Showing: 1 - 23 of 23

| When             | By        | About  | Comment  |
|------------------|-----------|--|--|
| 2017-08-26 12:12 | annmanley | Article: BUNDARRA. (Article), The Armidale Express and New England General Advertiser (NSW : 1856 - 1861; 1863 - 1889; 1891 - 1954), Tuesday 15 December 1885 page 2   | Robert Walter Murphy should read Richard Walter Murphy |
| 2017-08-19 15:07 | annmanley | Article: Barcardine AND General Budget WHAT'S THE WEEK? (Article), The Western Champion and General Advertiser for the Central-Western Districts (Barcardine, Qld. : 1892 - 1922), Saturday 9 December 1916 [Issue No.1299] page 7 | Ruby Riebelt should read Ruby Reibelt                  |
| 2017-08-19 14:55 | annmanley | Article: MRS. M. E. KELLY. (Detailed lists, results, guides), Warwick Daily News (Qld. : 1919 -1954), Saturday 2 April 1927 [Issue No.2430] page 6   | Mrs. P. Riebelt should read Mrs. P. Reibelt            |

### Text corrections

[Evening News \(Sydney, NSW : 1869 - 1931\), Fri 9 Oct 1914, Page 4 - AMBULANCE MEN'S DIARY](#)

| Date modified                            | Old Lines  | New Lines  |
|--|--|--|
| 3 years ago<br><a href="#">annmanley</a> | Two soldiers — Sullivan and Tonga — were brought back to us wounded. They told us one of their comrades named Street lay dead in the trenches. When daylight came, we fixed some iron railings round the graves of Ewell and Courtney. About dinner time, we had three cases of sunstroke to attend to. Two were very bad. It took us four hours to bring one round. Took all three to one of the destroyers, and had tea on board, which we much enjoyed. This boat conveyed us to Herbertshohe, where we met other men of our corps. | Two soldiers — Sullivan and Tonga — were brought back to us wounded. They told us one of their comrades named Street lay dead in the trenches. When daylight came, we fixed some iron railings round the graves of Ewell and Courtney. About dinner time, we had three cases of sunstroke to attend to. Two were very bad. It took us four hours to bring one round. Took all three to one of the destroyers, and had tea on board, which we much enjoyed. This boat conveyed us to Herbertshohe, where we met other men of our corps. |

We are looking at the profile of the same user on this slide and you can see in detail her comments and text corrections both before and after changes. Having this high level of transparency was important to us and the users, since there was no moderation on activity. Any changes a user makes go live immediately. If they correct a word to the correct spelling it is immediately searchable. By doing this it clearly shows the value of the activity to the user. Also it assumes that users can be trusted and when a high level of trust is gifted it is usually human nature to honour this.


Our strategy was to rely on the users to alert us if anything unexpected happened, which it never did in the 6 years I managed it. It would have been very difficult to moderate changes since there were thousands of users and the average activity of the top users is between 16 to 60,000 lines per month.

The screenshot shows a forum user profile for 'annmanley'. The profile is divided into several sections:

- Header:** Forum, What's New?, Back to Trove. Navigation links: Today's Posts, Trove FAQ, Forum FAQ, Calendar, Community, Forum Actions, Quick Links, Trove, Terms of Use.
- Member List:** annmanley
- Registration Notice:** If this is your first visit, be sure to check out the Forums FAQ and the Terms of Use. You have to register before you can post: click the register link above to proceed. To select the forum that you want to visit from the selection below.
- User Profile:**
  - annmanley** (Trove user)
  - About Me / Friends** tabs
  - Basic Information:**
    - About annmanley**
    - Biography:** Ann Manley: correcting papers fr July 2009; researching my family history for 30 years; grad Syd Uni; aka afm1
    - Location:** Narraweena, Sydney (Northern beaches)
    - Interests:** Genealogy (personally and as a volunteer helping others), swimming (participant and referee)
    - Occupation:** Volunteer, formerly systems analyst
    - Research:** So many - but any further info relating to all branches and twigs in my greater family tree.
  - Statistics:**
    - Total Posts:** 59
    - Posts Per Day:** 0.02
  - General Information:**
    - Last Activity:** 29-07-2017 02:46 AM
    - Join Date:** 10-07-2010
- Left Sidebar:**
  - Join Date: 10-07-2010
  - Last Activity: 29-07-2017 02:46 AM
  - Avatar: [Image]
  - 2 Friends (maurielyn, jhempenst)

This slide shows the same user in the Trove forum. Users can share as much information as they like about themselves, and communicate with the project team or with each other. Initially we did not build a communication mechanism into the service and we emailed users directly when they contacted us. However in the early days before we understood how much users wanted to talk to each other we observed a strange activity in the comments. Users were trying to connect with each other by leaving comments with their phone numbers and emails in articles that were heavily corrected, that many users would see. We were surprised by this and immediately set about investigating how to implement a forum, so that the virtual community as a whole could see each other, form sub interest groups, and build online camaraderie. Sub groups that developed where for example the people using knitting patterns, railway history enthusiasts, and local history groups. There were University researchers: one doing climate change and one doing influenza who galvanised groups of people to help them find and correct articles relevant to their research topics.

# Discussion: testing, enhancements, titles


Help [Signup](#) [Login](#)

---

Forum **What's New?** Back to Trove
Today's Posts Daily Group Message Daily Events Mark Forums Read
Search the Forums


---

Search [Search Results](#)

If this is your first visit, be sure to check out the [Forums FAQ](#) and the [Terms of Use](#). You have to [register](#) before you can post: click the register link above to proceed. To start viewing messages, select the forum that you want to visit from the selection below.

**Search:**  
Type: Posts; User: [annmanley](#) Page 1 of 3 [1](#) [2](#) [3](#)

| Search:   | Search took 0.00 seconds.    |
|---|------------------------------|
| 17-10-2014, 12:08 PM<br>Thread: When your corrections will not save by <a href="#">annmanley</a>  | Replies: 96<br>Views: 25,529 |
| <b>I have encountered a similar error in the last...</b>  |                              |
| I have encountered a similar error in the last couple of days since the recent update this week. It seems to occur when some special symbols are either encountered in the text, or alternatively,...     |                              |
| 29-05-2014, 04:12 PM<br>Thread: Digitised Newspapers - Advanced Search page Trove 6.1 by <a href="#">annmanley</a>  | Replies: 5<br>Views: 2,136   |
| <b>Digitised Newspapers - Advanced Search page Trove 6.1</b>  |                              |
| [I originally posted this to the Forum early this morning to a Thread under "Trove newspapers redevelopment" which it appears is only visible to persons who were invited to test same (in 2013). I am... |                              |
| 21-05-2014, 09:37 PM<br>Thread: Digitizing "The Age" by <a href="#">annmanley</a>   | Replies: 16<br>Views: 6,282  |
| <b>[QUOTE='mraadgev [NLA];4710']As far as I am aware...</b>   |                              |
| [QUOTE='mraadgev [NLA];4710']As far as I am aware the short answer to the question is funding - digitisation of the Age would cost about \$1m assuming that it is a similar number of pages to the...     |                              |



I added this slide to show the three main types of user discussion that went on in the forum. Firstly functionality – how to do things and error testing, secondly enhancement requests, and thirdly queries on which newspaper titles would be digitised.

## Key *good* decision points

- ❖ **Login not required** ....make it easy and open to all
- ❖ **No moderation** or vetting of activity .... trust your users to do the right thing
- ❖ **Edits are immediately live** and searchable ..... make it understandable and useful
- ❖ **Transparency of activity**....let all users see what has been done, by whom and when



Nick Youngson <http://nypphotographic.com>



So to summarise. The key *good* decision points we made in the crowdsourcing journey were:

Login is not required – make it as easy as possible for people to try it out without commitment or unnecessary barriers. As you have seen many users did decide later to register because otherwise they could not get in the ranking tables or see their own history of activity.

No moderation or vetting of activity. We placed a level of trust in the users, which they responded positively to. We relied on the user community to let us know if anything was wrong. However the original version of the text remains searchable if a user were to vandalise or delete it, and vandalism would be easy to reverse with rollback to before and after dates of changes.

All curation including corrections, tags and comments are immediately live within milliseconds *and* can be searched on. This really impressed users, helped them understand the big picture, and immediately improved the service for them.

A high level of transparency. Everyone can see all text changes to an article and who has done them and when – we retain the whole history and make it viewable. This also enabled roll back if required if anything went wrong, but it never did.

## Other things that work for crowdsourcing

- ❖ Options on how to work the task

e.g. provide next article we want you to correct in a theme, or choose your own;  
hard and long article vs easy and quick article..

- ❖ Show progress chart against target

in prominent place on interface.

- ❖ Share outcomes

- ❖ Give rewards, recognition, lapel pins



Limited edition Trove pin 1 of 100.



From my research into other crowdsourcing projects there are also another four things that ideally could be done to motivate and encourage users, which other sites have used successfully. These include having different options for users to work on your tasks, for example, for those not doing their own topics they generally prefer to be given work. We were asked lots of times by text correctors if we could define groups of articles for them to work on, instead of them having to search for their own.

Also it is important to have progress against the goal clearly visible to motivate users, if it is possible.

Users like to be kept informed about how their work fits into the big picture both in statistical format and through things like exciting discoveries and new research.

Lastly we had many requests for small tokens that would make users proud of their work and identifiable to others, mainly in the form of t-shirts, and lapel pins.

## What interests people – duck houses?

### UK MP Expenses Scandal 2009

- Power + Politics
- Money + Sex
- Scandal
- People - known and of interest
- Subject of interest and relevance to wide audience
- No experience/subject knowledge required to help
- For the public good



The luxury duck house cost tax payers £1,600



It is important to remember that no matter how good your functionality and interface is, the main goal of your crowdsourcing activity and the subject content are critical to its success. The goal has to be for the common good, to benefit as many people as possible, and the content has to be really interesting to a wide audience. Boring or niche topics will never attract a large body of people, of those for which a high level of prior knowledge is necessary.

Of all the crowdsourcing projects I have researched the one that galvanised a nation, even more so than Australian Newspapers is the UK Parliamentary Expenses Scandal in 2009, otherwise referred to as the duck house scandal. This was a major political scandal concerning expenses that MP's had claimed which included everything from the luxury duck house pictured to prostitutes and second homes. It is a perfect example of topics that make a project interesting. It covered power, politics, people, money, and sex. It appealed to a wide audience and was for the common good. In a matter of days, after the original documents were released under a Freedom of Information Act every single expense claim from every MP, which was about 400,000 barely legible handwritten pages, had been transcribed, tagged and flagged by the public in record time, in a crowdsourcing interface created by the Guardian newspaper.

The opening and identification of the information resulted in resignations,

sackings, prosecutions and imprisonments of members of the House of Commons and Lords.

## My top tips for crowdsourcing

| The Activity    | The Content                  | The Interface       | The volunteers      |
|-----------------|------------------------------|---------------------|---------------------|
| 👍 Clear goal    | 👍 Interesting                | 👍 Easy and fun      | 👍 Trust them        |
| 👍 Big Challenge | 👍 People-History-<br>Science | 👍 Reliable + quick  | 👍 Acknowledge       |
| 👍 Show Progress | 👍 Lots of it                 | 👍 Intuitive         | 👍 Rewards           |
| 👍 Show Results  | 👍 More of it                 | 👍 Different options | 👍 Build team spirit |

*'Crowdsourcing: How and why should libraries do it?' by Rose Holley 2010*



This table shows a summary of my top tips for crowdsourcing. More information can be found in my published paper 'Crowdsourcing, how and why should libraries do it?', with some samples of user interface designs.



## Success? When fun becomes addiction....

*"I would like to say Australian Newspapers 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."*

*"When I first joined Galaxy Zoo on 11<sup>th</sup> July 2007 my record longest continuous classifying period of time was 12 hours. I kept classifying when eating sandwiches as my lunch & dinner. It is deadly addictive."*

*"While going through a whole month in a slightly obsessive crazed mind searching Australian Newspapers 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 til dusk I seem to be in this website. Housework seems to have taken a backburner and meals are starting to come out of cans....."*



A measure of the success of your site is when it changes from being fun to being addictive. Both the Australian Newspapers and the Zooniverse citizen science projects have been described by many users as 'addictive' and both have retained a core group of loyal users for years, who are spending much longer per session and many more hours per week than you would expect. On both these projects some users are working as if it is a full time job.

Addiction comes when you have the perfect combination of interesting content and main goal and *all* the functionality and interface design points we have discussed, that have been developed based on user feedback and involvement. That is your service 100% meets the needs and expectations of your users.

## Cost benefit analysis

National Library of Australia –  
Newspaper Text Correction: 2013  
estimated that volunteer work was  
**valued at \$12 million**

My own calculations: 2017 the value  
is **in excess of \$44 million.**



Wikipedia: 2008 study  
estimated that **100 million  
hours** had been spent by  
volunteers creating Wikipedia  
articles in the first 6 years.

[https://en.wikipedia.org/wiki/Wikipedia\\_community](https://en.wikipedia.org/wiki/Wikipedia_community)

<https://www.nla.gov.au/sites/default/files/trove-crowdsourcing-behaviour.pdf>



What is the likely cost benefit in all this volunteer curation?

In April 2008, writer Clay Shirky and computer scientist Martin Wattenberg estimated the total time volunteer editors had spent creating Wikipedia at that point was roughly **100 million hours**.

In 2013 the National Library of Australia estimated that if staff had been employed to do text correction it would have cost in the vicinity of **\$12 million**.

In October 2017 my own calculations for the cost benefit of text correctors was **in excess of \$44 million**:

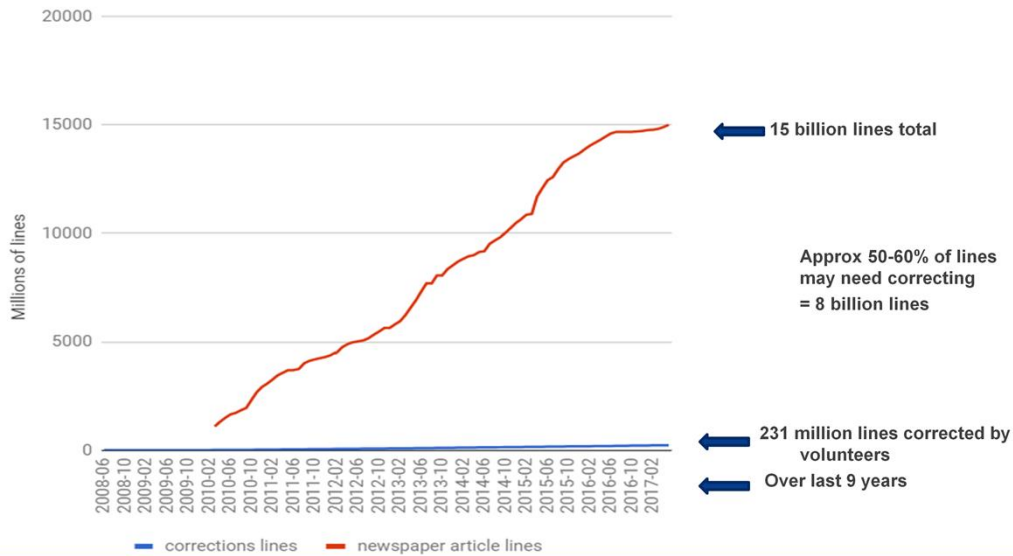
I calculated this by working out that in a single hour an average person can correct 95 lines (although top text correctors are doing double this amount). The 246 million total corrected lines from the last 9 years equate to 2.5 million hours work. At the Australian minimum wage of \$17.70 per hour this equals \$44 million. This figure does not include the overheads or office space to house approximately 8,000 workers each month.

But what does success really mean? In the case of Australian Newspapers the goal was to improve the quality of the search by increasing the overall accuracy of the corpus. Arguably this has not been significantly achieved as we will see shortly. However what has occurred – the creation of an active vibrant community that has become socially engaged across many physical and virtual communities was a

runaway success, even though very little thought had been given to this at the start and it was not even an articulated goal.

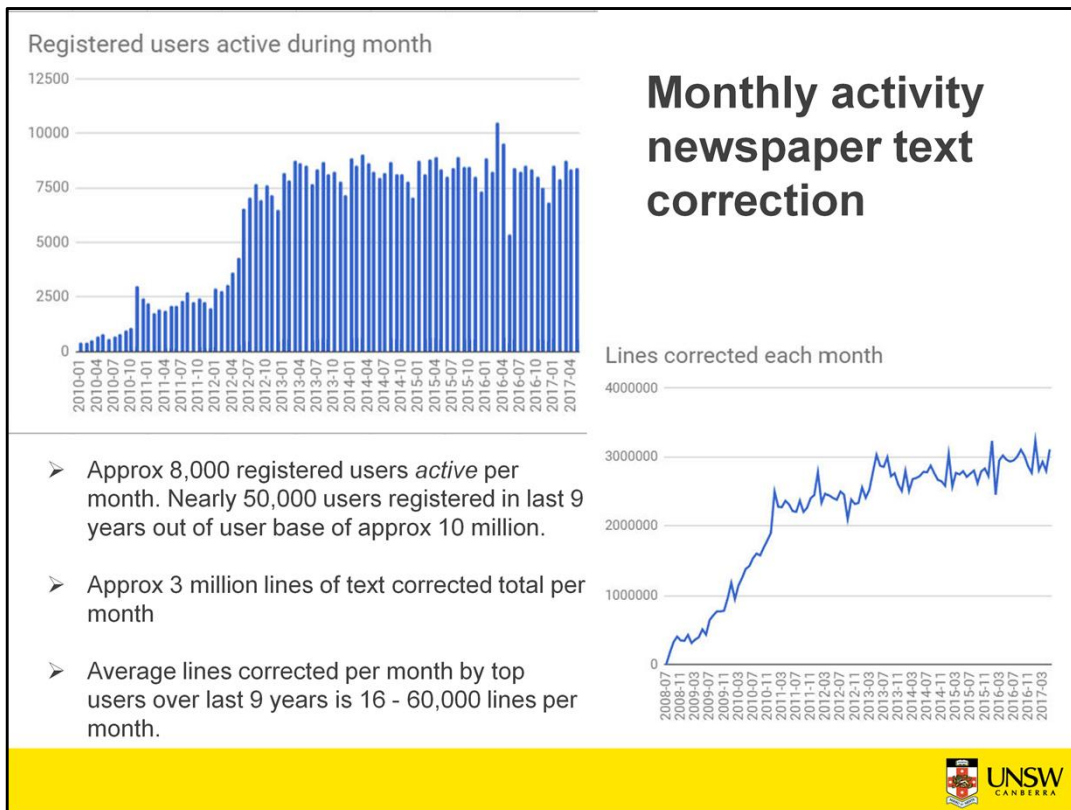
## Success? Australian Newspaper Text Corrections

Lines in articles compared to lines corrected



Let's look at this graph. The red line shows the increasing amount of newspaper articles being added to Trove over the last 9 years, measured by lines. By May 2017 this year there were 15 billion lines of newspaper text, which equates to 208 million articles. Based on research data I believe that perhaps half to two thirds of the articles have lines that need correcting which is approximately 8 billion lines. The blue line at the bottom shows the actual lines corrected by volunteers so far, which is 231 million. When we started this presentation that seemed a lot, but now we compare it to the goal and see it visually on the graph, it is but a drop in the ocean and doesn't seem very much at all. What this means is that the corpus has not been significantly improved in quality after all, which was the single aim of the crowdsourcing. This is the first time I have clearly understood this big picture myself and properly analysed the data. Don't get me wrong – the quality is not terrible, it's still above 90% character accuracy, it is just that we wanted it to be perfect, and it seems so hard to achieve this last little step of perfection.

So if we want to continue striving for perfection – what are our options to do this? Can we encourage more volunteers to join us or our existing users to do more corrections?



Let's examine more closely the activity of our text correctors. We have two charts here: the top shows the average number of users actively making corrections each month and the bottom shows the amount of lines corrected per month. You can see that during the first few years of the project the number of users and corrections was rising exponentially. When I was doing my first research on crowdsourcing in 2009 there was a direct correlation between amount of new content added and increasing activity of all users.

However you can see that in the last few years, despite more new content being added the activity is levelling out. It is currently remaining fairly consistent in that there are about 8,000 registered users active per month, correcting a total of 3 million lines per month. What we know from our ranking tables is that the top users are correcting much more than anyone else, which is up to 60,000 lines per month.

## Cognitive surplus saturation point?

**Wikipedia (established 17 years)**

2001 - 2007..... Volunteers grow exponentially (6 years)

2007 - 2011..... Volunteers plateau.

**By Nov 2011 31.7 million registered editors and 270,000 active monthly = 8%**

**128,267 registered active users per month in English edition of Wikipedia**

**Note: Actively involved volunteers are a tiny % of the total number of people using Wikipedia and Trove which is millions.**

**Trove newspaper text correctors (established 9 years)**

2008 - 2012..... Volunteers grow exponentially (5 years)

2012 - 2017..... Volunteers plateau

**By May 2017 50,000 registered text correctors and 8,000 active monthly = 16%**

**Zooniverse (established 9 years)**


2008 – 2014..... Volunteers grow exponentially 1 million registered users.

[https://en.wikipedia.org/wiki/Wikipedia\\_community](https://en.wikipedia.org/wiki/Wikipedia_community)




Research on other similar services including Wikipedia and Zooniverse show a similar pattern developing. It appears that at a certain point in time the cognitive surplus saturation point is reached. In Wikipedia the saturation point is 8% of registered users are active each month. It was originally thought that this could be because most of the encyclopaedia is now written. However since Australian Newspapers is showing the same pattern with saturation point at 16% of registered users, even when there is more content than ever before to correct, there may actually be something else happening. It is interesting to note that many of the volunteers who want to help with common good goals are actually involved in more than one crowdsourcing project and they divide their time equally between several worthy causes. I think further research on some of the established long term and large crowdsourcing projects is needed to get a better idea of what is actually happening here.

**And finally.....**



**Reset**  
Do you want to start over?  
**No** **Yes**

**If yes – what will you do differently?**



I am now nearing the end of this presentation so I want to address the question *“if you were starting over again from scratch, would you do anything differently now, based on what you have learnt in the last ten years and the advances in technology”?*

My answer is yes, however there is only one thing I would do differently.

Over the last ten years we have seen great advances in artificial intelligence, that is using computers to do things that previously could only be achieved by human intelligence using the mind, eye or hand. For example advances in facial recognition software which enables Facebook to recognise and tag you in other people’s photos.

Combined with these advances in intelligent software is the fact that we now have massive corpuses of newspaper text that has been corrected both by automated OCR software and manually by the human hand. We did not have any sample data like this when we started back in 2007.

Software can be trained by looking for patterns, and repeating these patterns millions of times. Neural networks allow massive processing capability in a timely way. Hardware to support this is becoming cheaper every day.

I want to introduce you to a new concept. That is to use artificial intelligence based software to improve the accuracy of text, which has been trained using a

giant newspaper corpus of manual corrections. This software would be applied *after* standard OCR correction has taken place and *before* manual crowdsourcing activity. It should significantly improve the accuracy of text and therefore searching.

Software research and development in this field has been ongoing for the last 7 years, most notably from the European Funded IMPACT group and also from Project Computing in Australia, which is the company Kent Fitch, the systems architect of Australian Newspapers and Trove, works for.



# OverProof: Post OCR text correction software



[Home](#) [Web demo](#) [Login](#) [About](#)

**OverProof corrects OCR**  
so that you can search, read and reuse your text more effectively

OCR engines are in the business of *character recognition*.  
OverProof is in the business of *word recognition*.

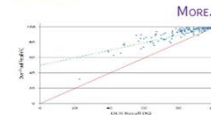
## The problem overProof solves

### OCR is hard and error-prone

OCR engines attempt a difficult task: converting colour and contrast variations on an image into text. Their job is made harder by poor original and image quality: poor and aged paper and inks, poor print quality, bleed-through, page-skew and less than optimal image-capture are common yet impossible or expensive to address constraints.

OCR engines need to balance processing speed and cost against accuracy, and they don't come equipped with the massive statistical and contextual language models required to generate the best possible output.

On a typical scanned newspaper corpus, overProof will reduce the number of articles missed by a keyword search due to OCR errors by over 50%.



I will now talk about OverProof, which is a product on the market that I would use if I was doing the same project again with English language newspapers. OverProof has been trained using many of the manual corrections applied to Australian Newspapers, and uses multiple OCR error models and models of the English Language. It is noted that even the manual corpus still contains errors. The manual corrected corpus is referred to as the gold or ground truth. The software runs through a training sequence where it checks against the entire ground truth for example 100 million lines several times, for example 100 times, before it is ready for business. The difference about OverProof is that it is correcting and measuring at word level rather than character level, because the artificial intelligence enables better understanding of words and their context.

Tests show that when applying OverProof on the Australian newspapers the word accuracy is raised from about 80% to 94% which is significant. Note this accuracy level is not related to OCR character accuracy, and also that word accuracy probably cannot be raised above 96%, since some of the original print is unreadable or missing.

## Try it out - supply your sample OCR text to be corrected

The text you supply below will be processed by overProof using profiles that assume

- English language
- OCR output from recent versions of ABBYY FineReader

Up to 28,000 characters may be entered.

At a later stage it was decided to levy a rabbit tax of 3d in the £, to apply all over the district. Mr. Clarke, in moving for this, said it was necessary, although it might seem hard on the townspeople. The Warden and Mr. Yaxley strove to have it reduced to 2d in the £, but were defeated. The intention of the council is to provide poisoned bait already prepared for use free of charge to ratepayers, and to instruct those applying in the use of it, appointing an assistant inspector for that purpose.

### Rabbit tax: Newspaper image


At a later stage it was decided to levy a rabbit tax of 3d in the £, to apply all over the district. Mr. Clarke, in moving for this, said it was necessary, although it might seem hard on the townspeople. The Warden and Mr. Yaxley strove to have it reduced to 2d in the £, but were defeated. The intention of the council is to provide poisoned bait already prepared for use free of charge to ratepayers, and to instruct those applying in the use of it, appointing an assistant inspector for that purpose.



If you go to the OverProof website you can read a lot more about the development, testing sample and training sets. Also on the site is a demo. I have pasted a sample of poor quality article text I found in Australian newspapers that had many errors into the box and then I click the 'correct' button. You can see on the right the original quality of the paper.

Processing complete: 22 Oct 17 23:32:46 words: 97 corrections: 20

| Received Text<br><input checked="" type="checkbox"/> Show  | Corrected Text<br><input checked="" type="checkbox"/> Show   |
|--|--|
| At a. later stage it <b>'was dlicidedl</b> to levy a <b>rablit</b> tax <b>or</b> 3d in the <b>??</b> , to apply all over the <b>dlistrict</b> . Mr. Clarke, <b>ii</b> moving for this, said it was <b>necos</b> sary. <b>ahloui-h</b> it might seem hard <b>oni</b> the" townspeople. The <b>Warden andl</b> Mr. Yiualey strove to have it <b>reduced</b> to 2d in the 'C, bt; were defeated. The intention of the council is to provide poisoned <b>hbait</b> already prepared for use free of charge to <b>ratgpayers,</b> and to instruct those <b>..?gplying inl</b> the use of it, appointing an <b>assistant in. specter</b> for that purpose. | At a later stage it <b>was decided</b> to levy a <b>rabbit</b> tax <b>of</b> 3d in the <b>way</b> , to apply all over the <b>district</b> . Mr. Clarke, <b>in</b> moving for this, said it was <b>needs</b> sary. <b>although</b> it might seem hard <b>on</b> the" townspeople. The <b>Warden and</b> Mr. Yiualey strove to have it <b>reduced</b> to 2d in the 'C, bt; were defeated. The intention of the council is to provide poisoned <b>bait</b> already prepared for use free of charge to <b>ratepayers,</b> and to instruct those <b>applying in</b> the use of it, appointing an <b>assistant inspector</b> for that purpose. |



Instantly the text is corrected and the words changed are highlighted. You can see the obvious improvement. This is quite a radical breakthrough. Obviously you would not manually use it like this, but run it across the corpus in an automated process. This software is available now.

If you were starting from scratch, then obviously it would be logical to use this software or similar.

[Ending]

Well, that concludes my presentation on crowdsourcing based curation, and user engagement in digital library design.

I hope that you have found this information interesting and relevant to your own project, and that you are now able to apply some of the things we have learned here in Australia. I wish you all the best for the Unesco National Digital Library of India project! Thank you.