

Australian Newspapers Beta Service: Summary of usage in the first 3 months after public release (4 August – 3 November 2008)

Author: Rose Holley (Manager - ANDP)

Version: 1.0

Date: 24 December 2008

Background

This document outlines usage of the Australian Newspapers beta service in the first 3 months after public release. (4 August 2008 – 3 November 2008). The Library has gathered these usage statistics in order to get an indication of how users are using the service. The usage information is being used in conjunction with user feedback to help in further development of the service. The Library has not been aiming for high usage at this stage. Content from the service was not harvested by Google in this period, and no media releases or planned publicity was carried out to publicise the service. Despite this there was an immediate uptake and usage of the service which was higher than anticipated. The usage figures are standalone and no attempt has been made to compare these figures with other services. Figures have been obtained from the newspapers content management system, the search and delivery system and Google Analytics.

Beta Usage Figures

1. Content in Service.

During the first 3 months the number of pages available increased significantly each week. The volume of content in the service is being measured in number of pages and number of articles. Statistics are automatically calculated each week for the number of pages added to the service. In the pages processed to date there is an average of 10 articles per newspaper page. Statistics on the number of volumes or issues delivered was not recorded.

Fig 1: Number of Newspaper Titles available in beta August 2008 – November 2008

Date	Total Number of Titles in service
4 August 2008	12 (Pilot)
3 September 2008	14
4 October 2008	25
1 November 2008	26

Fig 2: Page Content in beta 4 August 2008 – 3 November 2008.

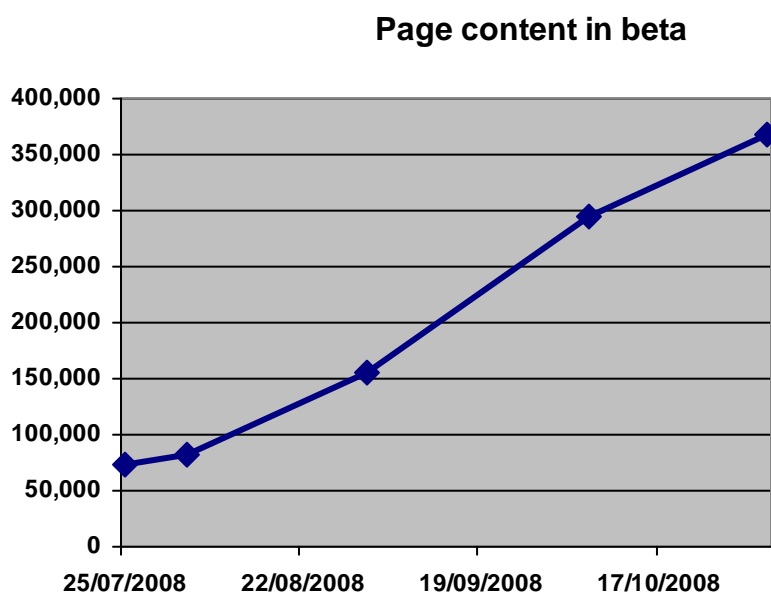


Fig 3: Breakdown of new page/article content added to beta August 2008 – November 2008.

Date	Total Pages added in preceding month	Total Number of Pages in service	Total number of Articles in service
25 July 2008 release date		73,000	
4 August 2008	36,234	82,104	768,852
1 September 2008	73,277	155,381	1651,010 (1.6 million)
6 October 2008	139,045	294,426	2560,814 (2.5 million)
3 November 2008	73,470	367,651 pages	3304,825 (3.3 million)

2. Service Availability

The beta service was available to users continually throughout the 3 months (25 July – 3 November 2008) with some minor temporary outages as below. The system coped well with usage that was significantly higher than expected.

Fig 4: ANDP beta service outages August – November 2008

Date of Outage	Time of outage	Duration of outage	Reason for outage
Mon 1 Sep 2008	9pm	3 hrs	Planned emergency shutdown for all NLA services
Sat 24 Oct 2008	9pm	40 mins	Planned maintenance of the server
Thur 5 Nov 2008	12pm	Intermittent	Intermittent unexplained outages possibly due to back up procedure

3. Number of Users

The total number of individuals using the service is unknown. This figure cannot be obtained since login to use the service is optional not compulsory.

The number of users who have registered through the login process to use the service is 1,488.

The number of unique visitors from Google Analytics (i.e. users logging on at different IP's) is 94,000.

4. Number of users registered to the announce list.

Prior to beta release potential public users who had contacted the ANDP via the 'Contact us' page on the website indicating they were interested in becoming testers of the new service were added to a list. This was 100 people. After beta release an invite to join an announce list was added to the survey feedback form and the 'about' page of the service. Users were told that announcements about significant things in beta, progress and outages would be made via the list. 300 people requested to be added to the list. The announce list now has 400 people on it.

5. Number of fans registered on Face book

When beta was launched the Library set up an identity in Face book for it. 48 individuals have now registered as 'Fans' of beta in Face book.

6. Number of Users who have provided feedback

Feedback for beta could be given in several ways:

- Via the online feedback survey set up in the beta service
- Via the 'Contact us' webpage link in beta and on the ANDP website
- By direct e-mail to ANDP team members
- By direct phone call to ANDP members
- Via other NLA web enquiry forms (reference enquiry, digital collection enquiry)
- By adding 'comments' at article level in beta (Comments was not intended for this purpose, but some users added comments here)

The Library has also been observing user comments about the service in online blogs and forums.

More than 340 people have answered questions in the online web survey and provided positive comments and suggestions for improvements. Approximately 200 people have contacted us via other methods with feedback. The feedback channels will remain open indefinitely and are proving very useful in gathering user feedback. 400 users can also be contacted via the announce list if we want to ask specific questions, do further user testing or have a more refined survey e.g. on OCR correction.

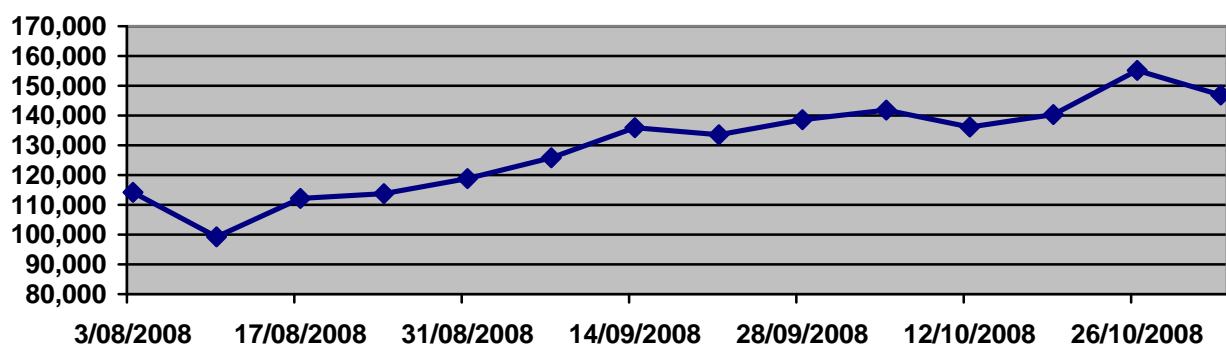
7. Keyword Searching

Measuring keyword searching with any degree of accuracy or meaningfulness has proved very difficult. The system itself does not have a search statistic facility so Google Analytics has been used. It is only significant to know about people searching the site, not robots crawling the site. Google analytics can recognise and exclude most robots that crawl the site. Crawls by robots vastly increase search statistics. Google analytics also does not include all searchers – only those using JavaScript and only those who have not blocked Google analytics. A single search may be counted several times if users go to different pages in the results or refine the results (another page loads each time). The search figures do not include browsing the newspapers (selecting browse page and picking title/date/state). The figures below should therefore be treated as a very approximate figure.

Fig 5: Summary of keyword searches in beta 3 August 2008- 2 November 2008

Average number of keyword searches per day	18,500
Average number of keyword searches per week	129,500
Total approximate number of keyword searches in period	1.8 million

Fig 6: Keyword searches in beta by week 3 August 2008 – 2 November 2008



8. Browsing

The amount of people accessing the newspapers by browsing (rather than keyword searching) cannot be measured accurately since there are several ways into browse mode. When designing the user interface and doing user testing we worked on the basis that at least half the users started their search by looking for the title/date of paper (rather than keyword). Therefore both 'Browse' and keyword 'Search' options are prominently available from the home page. If this is the case the figures above for keyword searching are only showing half of the usage (or possibly less).

9. Tag usage

Tags can be added to individual articles. The ANDP team were initially dubious about the value of adding tags to full-text material. Tagging has only been utilised on images 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 amount of tags (14,270) were added by users over a very short time. The majority of these tags are personal names and are possibly being used so that users can track their family research.

5079 of the tags are unique and have only been used once. The most used tag is LRRSA and has been used 559 times. It is suspected that this tag is being used by a group to track their research. Of the tags created only 18 have been used more than 100 times. The most popular tags are listed in Attachment A at the end of this document and the majority of these are not family names. The majority of users are tagging 1-5 times and often using the same tag (a family name). Only 27 users have tagged more than a hundred times. Of these 5 users have tagged over a 1000 times with the top tagger tagging 5546 times. 9054 individual articles have been tagged. 89 articles have more than 10 different tags attached to them. A summary of tag usage appears in the figure below.

Fig 7: Summary of beta tagging statistics 4 August 2008 – 4 Nov 2008

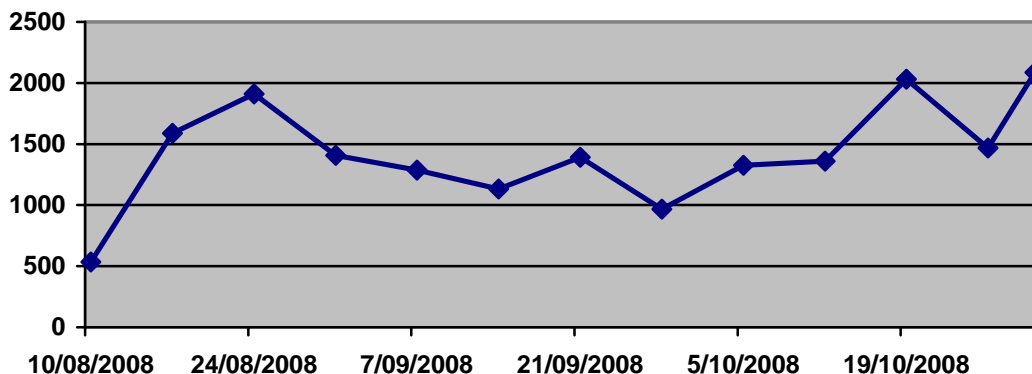
Total amount of tags created	14,270
Number of tags only used once	5079
Number of tags used more than once	9191
Total amount of articles tagged	9054
Amount of articles with more than 10 tags associated with them	89
Number of times most users have tagged	1-5
Number of users who have tagged more than 100 times	27
Number of users who have tagged more than 1000 times	5

Highest number of times a user has tagged	5546
Highest number of usages of a tag	559 (LRRSA) see Attachment A for rest of list
Number of tags used more than 100 times	18
Average amount of tags added in a week	1422

Fig 8: The most tagged articles in the beta service as at 3 November 2008

Number of tags associated with article	Title of article and Newspaper Citation details
50 tags	'The West End, Early Melbourne Memories' The Argus, 6 March, 1920 page 5. http://nla.gov.au/nla.news-article1680039
38 tags	'Early Melbourne, The West End' the Argus, 31 January 1925 page 8, http://nla.gov.au/nla.news-article2034683
37 tags	'Sydney' Sydney Gazette and New South Wales Advertiser 15 July 1820, page 2 http://nla.gov.au/nla.news-article2179614
35 tags	'The West End, Early Melbourne Memories' The Argus, 27 March 1920, page 4 http://nla.gov.au/nla.news-article1686507

Fig 9: Tagging by week 4 August – 4 November 2008



10. OCR Text Corrections

Users have actively been correcting OCR text since day 1 of beta release. There has been no time of the day or night when text correction has stopped since launch of the service. The basic ability to correct text has been given in the beta version, but no advanced power user mode to correct text or gather statistics is available at present.

Text correction is being measured by number of lines, and by number of articles corrected. It is not possible to gather automated statistics on % correctness of articles before and after public text correction. The statistics therefore show number of text edits rather than 'corrections'. We assume that edits are making the text more correct.

In the first 3 months 868 registered users have corrected text and approximately 390 unregistered users (total of 1200 text correctors). This means that 58% of registered users are correcting text. 720,795 lines of text have been corrected within 50,887 articles. The top text corrector has corrected 59,000 lines of text within 1890 articles. Some articles have had corrections added by more than 7 users (e.g. articles in the first Australian newspaper the 1803 Sydney Gazette). This particular issue in its entirety has had several different users working on corrections (because it is difficult to read and is an important paper).

Fig 10: Summary of public OCR correction figures 4 Aug – 4 Nov 2008

Total number of lines corrected by public	720,795 lines
Total number of articles that have had text corrected by public	50,887 articles
Number of articles corrected by anonymous users	15,732
Number of articles corrected by registered users	35,155
Number of registered users doing text correction	868
Number of unregistered users doing text correction (approx)	390
Total approx number of users doing text correction	1200
% of registered users doing text correction	58%
% of correction being done by registered users	73%
% of overall users doing text correction	Unknown
Highest number of articles corrected by a single registered user	1895
Highest number of lines corrected by a single registered user	59,316
Most edited article	410 edits
Number of articles edited by multiple users with logins	1264

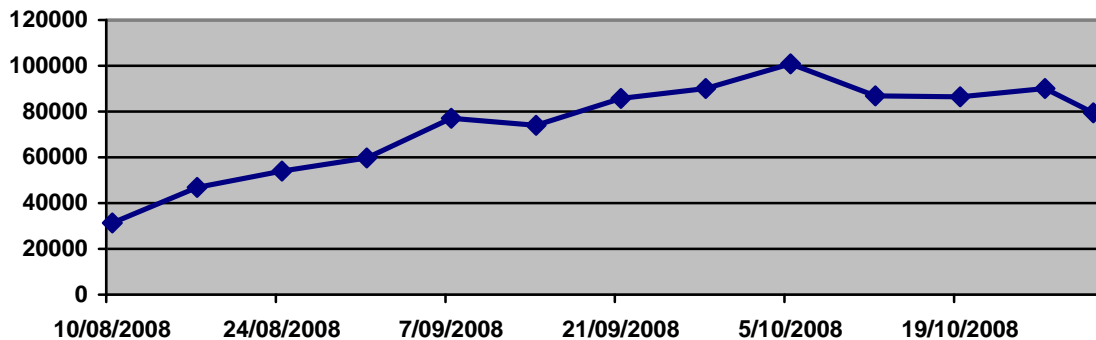
Fig 11: Greatest amount of corrections by month by line (top 5 correctors)

August 2008	Lines corrected
jhempenstall	23,623
Mrbh	9,673
Camerong	6,277
Scmorris	5,623
Maurielyn	5,372
September 2008	Lines corrected
Maurielyn	28,111
Jhempenstall	23,623
Mrbh	23,139
Fwalker13	13,141
Cmdevine	8,748
October 2008	Lines corrected
Maurielyn	24,538
Jhempenstall	23,265
John F Hall	21,886
Fwalker13	19,714
Cmdevine	16,896

Fig 12: Top 3 correctors over 3 month period 4 August – 4 November 2008, by line and by article.

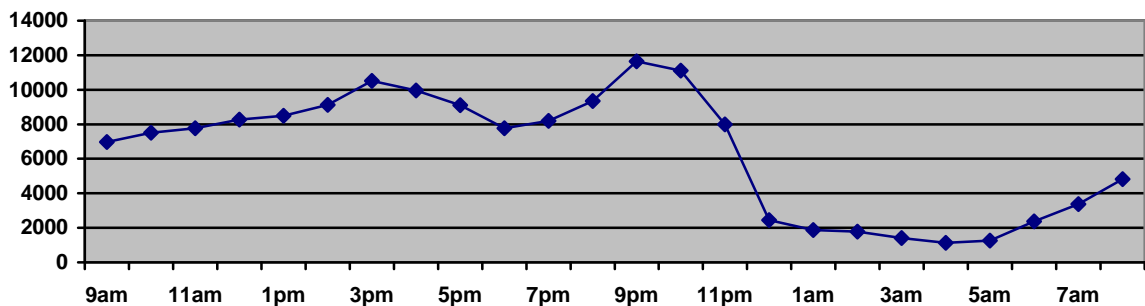
Identity	Total Number of lines corrected	Total Number of articles corrected
Jhempenstall	59,316 lines	1,895 articles
Maurielyn	58,021 lines	1,081 articles
Mrbh	46,488 lines	971 articles

Fig 13: Number of lines corrected by week August – November 2008



An average of 74,045 lines are corrected per week. Highest is 100,772 in a week, lowest is 31,390 in a week.

Fig 14: OCR corrections by time of day 4 August 2008 – 3 November 2008 (number of times 'save OCR corrections' is clicked).



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

Fig 15: Sample of OCR correction activity on an individual article 2 November 2008

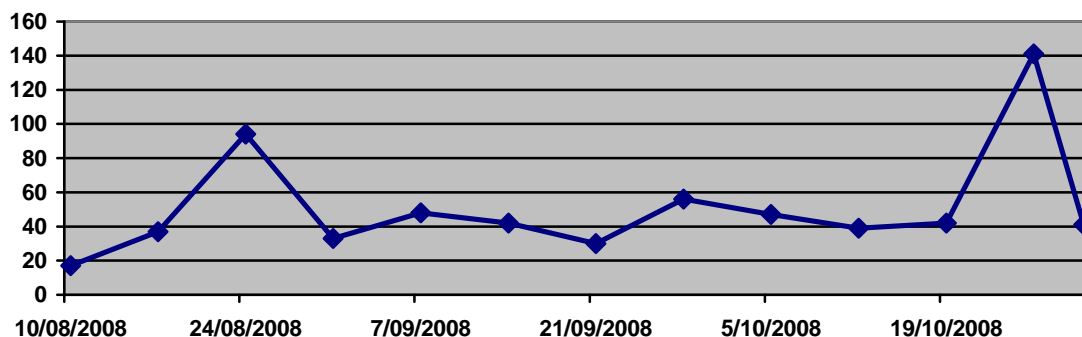
Time	User	Activity
12:23	user 1	Corrects text in the article but doesn't know how to insert the pound symbol
12:24	user 1	Creates a comment on the same article saying that he doesn't know how to enter the pound symbol.
19:13	user 2	Enters the pound symbol in the article.
23:09	user 3	Spots an error in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

11. Comment Usage

The comment functionality has been not been used very much in comparison to the tagging or OCR correction features. Comments were instigated 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 called 'comments'. The lack of use of the 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 articles have been found.

A total of 667 comments have been added over the 3 month duration. This is on average 51 comments per week. The highest amount added in a week is 141 and the lowest is 17.

Fig 16: Comments added by week



On reviewing comments the following 3 things have been noted:

- Users are using comments 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 can't do this at present unless they know html code. 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.

12. Usage of individual newspaper titles

No data available at time of writing

13. Most popular articles

No data available at time of writing

14. Most popular keyword searches

- Winston Churchill
- Cricket

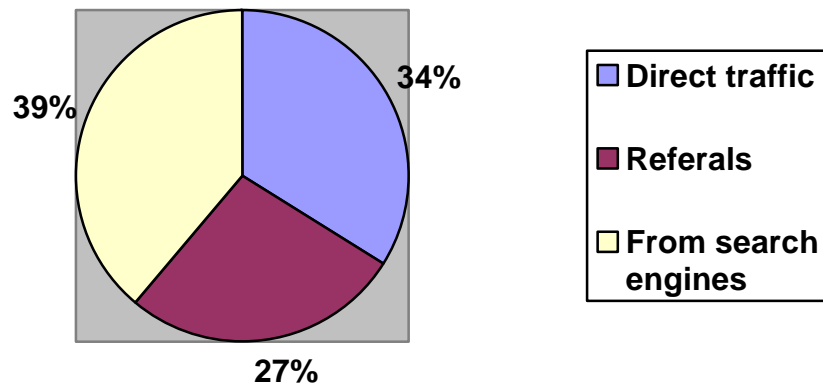
15. Publicity and Peaks

Despite giving no official press releases several newspapers quickly picked up on the release of beta and published articles. This generated surges and peaks in usage. The most notable peak was when the online version of the Brisbane Courier Mail published a front page story with a hyperlink to the newspapers first edition in beta. Another peak occurred when a message was posted on the Australian Wikipedia editors page about the new service. Blogs and forums widely discussed the service and publicised it.

16. Referrals from other sites

The Library expects that in the long-term much of the usage of the service will be generated from directions from other sources. During the beta phase August – November 2008 the Library was in discussion with Google about generating suitable sitemaps so that articles could be harvested and appear in the results list of Google web search and Google News Archive <http://news.google.com/archivesearch>. This was not achieved by 3 November 2008 so there was a limited amount of redirection from Google in the first 3 months. When it is included in Google usage is likely to rapidly increase. Links were added from Wikipedia newspaper title pages to the service and a noticeable amount of redirections was received from Wikipedia.

Fig 17: Traffic to beta 4 August –3 November 2008



The top 5 referring sites are:

1. NLA website
2. Google web
3. Coraweb – Australian Gateway for Family History
4. News.com.au
5. Wikipedia

(Not all State and Territory Libraries or other Libraries had added a link to the site from their websites during this period).

17. Average Time on site

Average time spent on the site by users in Australia is 18 minutes.

18. International Usage

78% of the usage is by people located in Australia. The top 10 countries using beta are:

1. Australia
2. UK
3. USA
4. NZ
5. Canada
6. Turkey
7. Netherlands
8. Germany
9. France
10. Brazil

Attachment A: Most Popular Tags being used in Beta 3 August – 3 November 2008

Name of tag	Usage Count
LRRSA	559
Cakebread	249
suicide	211
Bendigo	207
Murder	167
Cane	160
advertisement	152
milo cigarettes	146
Ticket of Leave	136
Caulfield Grammar School	127
loco	116
Proportional Representation	115
Wollongong NSW	111
STV	109
Hans Irvine Ebeling	109
Ulladulla	105
Animal Accidents	105
dundonians	100
Armistice Day	94
Peterson	91
William Macmahon Ball	90
Norfolk Island 1st Settlement	90
Female Orphan School	89
shooting	79
Cigarette Advertising	79
meningitis	69
Chinese	68
Pastoral Stations WA NorthWest	66
mine	65
poetry	62
timber	61
Fortitude Valley School	60
Walter Murray Buntine	60
poisoning	58
Murder WA North West	57
obituary	56
Old Sydney Burial Ground	56
Dargin	55
brass band	52
John Batman	47
Roebourne Gaol	45
for sale	45
coal	45
Charleville	44
Peter Headland	44
Peter Hedland	44
Racism	44
Leaving the Colony	44
Burkinshaw	43
Drowning	42
early Melbourne	42
sawmilling	42
colliery	41

Aborigine	40
Victorian Railways	40
Shark Attack	39
A J Macinnis	38
Macinnis	38
family	38
Juniper Hall	38
Ned Kelly	38
Retreat Station	37
cerebro-spinal meningitis	37
shipwreck	37
Cricket	36
Cyclones WA North West	36
Hare-Clark system	36
Roebourne Cemetery	35
NSW Corps	35
Robert Cooper	34
Dampier Archipelago	33
Kendall	33
C J Dennis	32
Roebourne	31
Berner	31
Catherine Helen Spence	31
2ft gauge	31
samuel berjew fookes	31
Senate electoral system	31
Dame Nellie Melba	30
Tambo	30
Hay	30
roma villa	29
poem	29
Mining	29
band contest	29
plumber and glazier	28
Gallipoli	28
Wilbow	28
Gold	28
message in a bottle	27
Pardons	26
Pastoralists WA North West	26
Mandeville Hall	26
PR Society of Victoria	26
hall scott	26
Underwood	26
Sir James Barrett	26
James Raworth Kennedy Family	25
anthrax	25
Francis Cadell	24
Francis Oakes	24
Strychnine	24
Pearling WA North West	24
Amelia Earhart	24
Sandhills Cemetery Sydney	24
Cossack	23
Bendigo Hospital	23
opium	23
fire	23
Donald Bradman	23
Edward C O Howard	23
Aboriginal - South Coast NSW	22
Orphan Fund	22

Everingham	22
Jamison Street	22
Victoria Mill	22
Kable	22
Lonely Graves WA North West	22
Women's Suffrage	22
Execution	21
yarra river	21
