

Tagging Full Text Searchable Articles: An Overview of Social Tagging Activity in Historic Australian Newspapers August 2008 – August 2009

Rose Holley

Manager - Australian Newspapers Digitisation Program (ANDP)

National Library of Australia

rholley@nla.gov.au

Abstract

In August 2008, tagging was implemented on articles that were full text searchable within the National Library of Australia's historic Australian Newspapers service (<http://newspapers.nla.gov.au>).. During the first year, 500 users created over 100,000 tags, 38,000 of which were distinct. The tagging was very successful and the National Library will be extending the tagging functionality to all of its other collections before the end of 2009. In this article, the tagging activity, behaviors and outcomes are analyzed and compared with other research on image tagging.

Keywords:

tagging, tags, social engagement, social metadata, user engagement, folksonomies, historic newspapers, Australian newspapers, full-text resources, web 2.0, user generated content, social tagging.

1. Introduction

Non-profit making and commercial organizations like Flickr, Youtube, LibraryThing, and Amazon have responded to user needs by enabling tagging, commenting, rating and other social metadata engagement (web 2.0) tools across books, videos, images and websites. Despite proving very popular with users, the library, archive, museum and gallery sectors have been slow to follow suit with their own collections. In the cultural heritage sector a fair amount of research and pilot testing has been carried out on social tagging as the precursor to perhaps implementing it more widely. The most notable and extensive research is undoubtedly that around steve.museum¹ an open source tagging tool. Two excellent reports have recently been published on two years of research (2006-2008) undertaken by steve.museum².

In August 2008, the National Library of Australia (NLA) implemented a tagging application for the first time on one of its own collections: the newly released Australian Newspapers beta service (<http://newspapers.nla.gov.au>),³ which contained 1 million full-text searchable articles of historic Australian Newspapers from 1803-1954. The number of articles increased to 5 million by the end of the year. The tagging was not done as an experiment and therefore was not controlled. The taggers were real users who wanted to tag for reasons of their own. Nevertheless, over 500 users created a tag pool of over 102,000 total tags in the first year of which 38,000 were different (distinct) tags.

Users could be registered or anonymous. During this time, all web 2.0 user activity, such as tagging, commenting and text correction, was monitored (but not moderated) by NLA through the gathering of statistics and communicating with users.

The Australian Newspapers service is the only online service from the National Library of Australia that has utilized web 2.0 features. It is also one of only two known large cultural heritage institutions in Australia that have enabled tagging across their collections, the other being the Powerhouse Museum. A survey carried out on Australian cultural heritage institutions in 2008⁴ revealed that though many institutions were thinking about tagging only two were actually doing it. It also reported that "institutions who have not implemented user tagging generally perceive many potential problems that institutions who have implemented user tagging do not report".

I have undertaken my own research into the tagging activity that occurred in the Australian Newspapers service over the first year. This article gives an overview of the public reaction to and utilization of the tagging facility in a full-text searchable collection, and provides statistics over a year's duration, observations on the use of tagging and suggestions for future developments. These may be relevant for other libraries and collections who are considering implementation of tagging. I was also interested in finding out if tagging activity and behavior may be any different on full-text resources as compared to image collections, so I have compared the NLA findings with other recent research on tagging in image collections. This includes the *steve.museum* project research where 1621 users added 36,981 tags to 1,784 images of museum and gallery works between March 2007- March 2008, and the Library of Congress Flickr pilot project ⁵ where 2,518 users added 67,176 tags to 4,615 photographs from their collection between January 2008 – October 2008. These can be considered a fair comparison to the Australian Newspapers service.

Tagging of resources in the Australian Newspapers collection was implemented for the primary purpose of improving the data quality of the resource. The success of the tagging was measured on three things:

- 1) the public would utilize the tagging feature;
- 2) they would not abuse the feature; and
- 3) the tags created would enhance the data and be useful for other users.

By these measures, tagging of the resources was considered successful.

A secondary but very significant outcome was that the Library harnessed a high level of social engagement from its users. The Library is now extending the tagging functionality across all of its analogue and digital collections.

This article does not cover the public text correction feature in Australian Newspapers that was introduced along with the tagging and commenting features and was even more successful than the tagging. Text correction is fully covered in the 'Many Hands Make Light Work' report ⁶.

2. Tagging functionality and implementation in Australian Newspapers

The functionality given to users for tagging in Australian Newspapers include the following:

- Tags can be added to articles either by registered users or anonymous users. (Anonymous users must complete a "captcha" challenge once per session.)
- Registered users can edit or delete their own tags.
- Anonymous users can edit or delete their own tags within the same user session only.
- All tags can be viewed in the tag cloud.
- The 10 most used tags can be viewed on the browse page.
- The latest tags added can be viewed on the home page.
- A user's own tags can be viewed on his or her personal profile page.
- Tags added to an article can be viewed on the article view page.
- Tags can contain only a-z, A-Z, 0-9, underscore, hyphen, and apostrophe.
- Tags can contain more than one word, e.g. Animal Accidents.
- Tags are limited to 60 characters.
- An unlimited amount of tags can be added to an article.
- An individual user can add up to 50 tags to the same article.
- There is no limit on the amount of tags an individual can create.

Although it was always intended that users would be able to search across tags, users have not, however, been given that ability so far. . Unfortunately, project priorities were diverted from implementing the agreed interface and functionality enhancements, including the searching of tags, to other more critical priorities. Therefore, during the entire period of this research tags were not searchable by the public, though they were browsable and could be viewed at article level. In addition, no guidelines for creation or management of tags were provided, due to lack of staff resource to develop them and to the team's thinking that guidelines were not essential.

The data enhancements from tagging, commenting and text correction are stored in layers within the Lucene database (they do not overwrite existing data, even for text corrections). The data layers can, in theory, be searched separately as user layers or in combination with library provided metadata layers. At present a searching facility across tags and comments has not been enabled but the search across text corrections has been (both user and library layers). Most other library collections that have tagging do so across either images or at item level, e.g. book. The Australian Newspapers service is perhaps unique in that it has enabled tagging of searchable text at article level. When tagging was implemented, it was not anticipated that it would be used very much since all the articles are full-text searchable. In this sense it is quite different from tagging an image collection. Nonetheless, article tagging has been utilized a great deal and has proved very popular with users.

Implementation of tagging was a relatively easy task and took little time. A challenging question and one not yet answered is this: if you enabled searching across the different user-generated layers and library layers together, how would the presence of a tag that matches your search term affect the relevancy ranking?

3. Staff resource required to support tagging

After initial development and implementation of tagging, no staff resource was required to support it, because the public tagging was not moderated. However, in the first year the Australian Newspapers Digitisation Program (ANDP) team decided to monitor tagging by gathering statistics and communicating with users, since this was the first implementation of tagging on a collection at the National Library of Australia. This was a task within the main project plan, since other specific things were also being monitored at that stage, such as text correction. In the first year around 20 hours were spent on monitoring and statistics gathering related to the tagging. Once the service was officially launched and out of beta phase, it was not a requirement to gather tagging statistics for reporting purposes or to monitor tagging. Social engagement is not measured at the National Library of Australia, nor is improvement to data quality by addition of user generated content layers. I did feel, however, that it would have been desirable to have the staff resource to discuss, establish and write tagging guidelines, and on a weekly basis to manually scan new tags created to ensure that no abusive terms had been created. Neither of these things has yet been done. Should tagging guidelines be established and if, for example, it were decided to 'tidy up' tags, then existing tags would need to be retrospectively converted. This could be done largely with user volunteers, rather than library staff. The tagging feature has been a big crowd-pleaser for public users, and It was a quick win for the Library that required very little work to implement and little to no support.

4. Usage of the Australian Newspapers service and tagging feature

The beta service was not publicized or promoted by the National Library of Australia. It was not originally intended to be in 'beta' version for a year, only for 3 months. Originally it was anticipated that relatively few users would become aware of the service and that they would agree to become 'testers' and give feedback in specific areas. As it turned out, the service was in 'beta' version for a year and thousands of users became aware of the service via viral marketing (mainly genealogy blogs)

resulting in half a million users by the end of the year. Hundreds of users gave feedback multiple times and responded to specific queries, and the data in the service expanded considerably during beta phase. As users and data increased, tagging also increased. The tables below give an overview of service usage and activity.

Table 1: Australian Newspapers Service Usage August 4 2008 – November 4 2009.

Statistic type	4 Nov 2008 (3 months after release)	4 Feb 2009 (6 months after release)	4 May 2009 (9 months after release)	20 Aug 2009 (1 year after release)	4 Nov 2009 (15 months after release)
Number of pages in service	367,000	367,000	367,000	538,334	832,665
Number of articles in service	3.5 million	3.5 million	3.5 million	5.8 million	8.4 million
Unique visitors to site	94,000	205,000	347,000	492,000	787,000
Number of registered users	1,488	2,994	3,796	4,762	6,006
Lines of text corrected	1 million	2.2 million	3.4 million	4.7 million	7 million
Number of articles corrected	60,000	104,000	154,000	216,093	318,169
Number of comments added	800	1,806	2,582	3,441	4,618
Number of tags added	18,000	43,000	73,733	105,028	197,579
Total keyword searches since 4 August 2008 release	2 million	2.5 million	2.8 million	3.3 million	3.9 million

Table 2: Tagging Activity in Australian Newspapers - August 4 2008 - August 4 2009.

Activity Type	Number	% of Total
Total amount of tags added (tag pool)	102,929	
Number of different (distinct) tags in the pool	38,259	
Number of public tags in the pool	100,681	98% of tag pool
Number of private tags in the pool	2,248	2% of tag pool
Number of registered users who are taggers	549	11% of total registered users
Number of anonymous users who are taggers	unknown	
Number of tags added by registered users	95,013	92% of tag pool
Number of tags added by anonymous users	7,916	8% of tag pool
Number of different (distinct) tags used only once	28,348	74% of distinct tags
Number of different (distinct) tags used 2-5 times	7,414	19% of distinct tags
Number of different (distinct) tags used more than 5 times	2,497	7% of distinct tags
Number of different (distinct) tags used 100 times or more	66	0.2% of distinct tags
Total amount of articles tagged	38,874	Less than 1% of articles in database
(Total amount of articles corrected)	(216,093)	

Activity Type	Number	% of Total
Amount of articles with 10 or more tags associated with them	1,022	3% of tagged articles.
Number of times registered users usually tag.	No pattern, no predominant number, varies from 1 - 19,431 15 users: more than 1000 71 users: 101 - 1000 176 users: 10 - 100 213 users: 2 - 9 74 users: 1	
Highest number of times a user has tagged	19,431 (next highest is 9,872)	
Average amount of tags added in a month	10,000	
Most tagged article	Tagged 52 times.	

Table 3: Top 10 Tags August 2008 – August 2009.

Top 10 Tags	By number of times assigned
LRRSA	2,312
Murder	846
Bendigo	620
Lady Jane Franklin	515
Maryborough Qld BDMs	491
Gold mining	425
Suicide	400
Sir John Franklin	365
Cane	347
Sawmilling	331
Top 10 Tags	By number of different registered users who assigned the same term (+ unknown amount of anonymous users)
Murder	39 + anonymous
Death	27 + anonymous
Cricket	23 + anonymous
Suicide	22 + anonymous
Marriage	22 + anonymous
Melbourne	23 + anonymous
Canberra	20 + anonymous
Accident	20 + anonymous
Adelaide	17 + anonymous
Drowning	17 + anonymous

Table 4: Most common tags (based on number of times assigned and number of different users who assigned) grouped by type.

Subjects			
Murder	railway	Insolvency	Death notice
death	Shipping	Aboriginal	Deaths
cricket	Inquest	divorce	Kelly Gang
suicide	Gold	Victorian Railways	racism
Marriage	Bushrangers	Aborigines	railway accident
Accident	birth	Mining	Police
drowning	shipwreck	immigration	Gold mining
execution	Bushranger	fashion	Sawmilling
shooting	obituary	Flood	cane
hanging	poetry	Football	timber
fire	Ticket of Leave	funeral	LRRSA

Places			
Melbourne	New Zealand	China	Collingwood
Canberra	Williamstown	Ballarat	Brunswick
Adelaide	Darwin	Bendigo	Dalby
Sydney	Brisbane	London	Kyneton
Geelong	Rockhampton	Ipswich	England
Hobart	Toowoomba	Echuca	Fiji
Maryborough	St Kilda	Fremantle	North Brisbane
Newcastle	Tasmania	Brighton	South Brisbane

Events			
world war 1	WW2	Bubonic plague Brisbane	Windsor murder
world war 2	1891 Shearers strike	Norfolk Island 1 st settlement	Gun alley murder

People			
Burke and Wills	Lady Jane Franklin	John Blaxland	Donald George Bradman
Smith	Sir John Franklin	Henry McGuigan	Frederick Wright Unwin

Table 5: The most tagged articles in the service as at 4 August 2009.

Number of tags associated with article	Title of article and newspaper citation details
52 tags	The Moreton Bay Courier, Sat 20 June 1846 page 1. Classified advertising tagged entirely with personal names. http://nla.gov.au/nla.news-article3710379
51 tags	The Argus, Fri 11 June 1915 page 6. Australian casualties of war. Personal particulars. Tagged entirely with personal names. http://nla.gov.au/nla.news-page381071
51 tags	The Argus, Sat 6 March, 1920 page 5. 'The West End, Early Melbourne Memories'. Tagged with personal names and building names. http://nla.gov.au/nla.news-article1680039

Table 6: Top 20 Taggers by number of tags created from August 4 2008 – August 6 2009.

Username	Number of tags created	Top 20 text corrector also
User:1	19,588	√
User:2	9,872	√
<i>[Anonymous users all together]</i>	<i>[8,079]</i>	-
User:3	7,968	X
User:4	5,042	√
User:5	4,305	√
User:6	2,683	X
User:7	2,230	X
User:8	2,172	X
User:9	2,121	X
User:10	1,917	X
User:11	1,762	X
User:12	1,602	X
User:13	1,582	X
User:14	1,293	X
User:15	1,230	X
User:16	1,150	X
User:17	990	X
User:18	851	X
User:19	843	X
User:20	807	X
Total tags top 20 created = 70,008 % of all tags in the pool = 69%		4 users

5. Tagging guidelines

Tagging took off from day one of release. After the first 12 weeks, around 14,000 tags had been added and quite a few e-mails were received saying that there was tagging chaos. There was a strong expectation from users that, since this was a service run by a library, there would be some tagging rules and that librarians would be monitoring and editing tags that did not adhere to the rules. The ANDP team took no action at this time other than telling users that there were no rules or guidelines for tagging. As time went by and users successfully used the other web 2.0 features (commenting and text correction) and understood that a certain level of control and monitoring was in their own hands, they began to suggest that they themselves should be able to monitor and edit other people's tags to help make them conform. The large majority of tags added were for people's names, and taggers mainly wanted to know how the names should be entered. There was an expectation that the library would want them in some kind of library-authorized format, for example surname first, and taggers worried that they were doing it wrong. Taggers could edit their own tags (for example to correct spelling mistakes or change the order of words in personal names). After about 6 months when the ANDP team again confirmed it would not create guidelines, the taggers themselves bought order to the perceived chaos. Through common sense and their observation of other users' tagging activity, they clearly developed their own unwritten rules. Amazingly, they achieved this without being able to communicate with each other using the system. The unwritten rules they developed for tagging can be described as follows:

- Use natural language order for names, subjects, places
 - e.g. John James Clark
 - e.g. Caulfield Grammar School
 - e.g. Japanese war crimes
- Don't join up phrases; keep them separate
 - e.g. Sydney Opera House
- It is okay to use apostrophes
 - e.g. St Helen's Orphanage
- It is okay to use hyphens – usually to convey subject hierarchy
 - e.g. socio-economics
 - e.g. Tramways – horse-powered – 1856
 - e.g. Tramways – wooden – proposed
- If there are a lot of tags on the same topic try and be specific and use hierarchy with the main topic word appearing first
 - e.g. Soccer injuries 1894
 - e.g. Soccer injuries 1910
 - e.g. soccer players
 - e.g. soccer in victoria
 - e.g. soccer in WA
 - e.g. Tramways – horse-powered -1856
 - e.g. Tramways – wooden – proposed
- Upper/lower case does not really matter – no agreed rule here
- Try and use a term that will be logical and meaningful to others
 - e.g. Ticket of Leave
 - e.g. Jetties and piers
 - e.g. Turon River Goldfields
- With people's names it is preferable to put in full all the known first names, or else use initials
 - e.g. John Patient Smith
 - e.g. J S Smith
- If the name is common or duplicated, it can be further qualified by a hyphen at the end and by adding dates of life/birth/death and/or occupation
 - e.g. Edward James Clark – Architect
 - e.g. Rev A M Henderson 1820-1876
 - e.g. Thomas – Wright- Artist – 1830-1880
- Try not to use abbreviations unless they are well known
 - e.g. WW2
 - e.g. RAAF
- It is okay to use numbers and dates in tags
 - e.g. 14th Battalion
 - e.g. 11 Manor Place
 - e.g. 18 October 1843

e.g. soccer injuries 1894

- It is okay to use tags to track your own or group research or text correction
 - e.g. done
 - e.g. not done
 - e.g. check
 - e.g. checked
 - e.g. completely corrected
 - e.g. mine
 - e.g. xx

6. Observations on tagging activity

The total number of tags added in each three-month period was recorded during the first year of service availability. We were unsure if the pattern and amount of tagging that occurred in the first three months would be different than that in other months once the service and a tag cloud had been established. The result was that in the first three months there was a lot of confusion on the part of users who were unsure what to put in their tags since there were no guidelines and no examples. Users also appeared to be unclear about the purpose of tags, where they would be able to view them and whether it was possible to search or browse for tags. On reflection, if establishing a new tagging service, it would be preferable to seed a sample subject area with tags so that users could see the tags in action and have some examples to which to refer, and also to provide guidelines for those who wanted them. All the taggers were real users who had discovered the service themselves and decided on their own to start tagging. They were not directed or encouraged in any way.

As a result of user requests, the tag length was increased from 30 to 60 characters, and the limit of 50 tags per article was removed. Some articles, for example family notices, were tagged with more than 50 names.

Once the tagging community had established its own unwritten commonsense guidelines, the tagging settled down. During that period, the number of taggers did not increase much; it remained around 500+, and users consistently added about 10,000 tags a month. In the first three months, most of the tags created were distinct tags, and these were mostly used only one time. This may be a normal pattern when a tag pool is being established. By the end of the year users were duplicating tag terms, so new tags being created were not always unique. 74% of the distinct tags were used only once, and most of these tags were personal names. This is noted because some information professionals are of the opinion that tags are only useful if used more than once; however, the taggers do not seem to share that opinion. Tonkin's⁷ research on sample data from Flickr showed single use tags comprised 10 -15% of the tags (and may be due to misspellings), so the incidence of single use tags in Australian Newspapers is higher. Less than 1% of the distinct tags had been used 100 times or more. This is why the tag cloud looked more like 'tag fog' and was not useful. No words jumped out; the tag cloud was mostly just a solid mass of names.

At 12 weeks the tag fog had already developed, and there were 18,000 tags in the tag cloud, most of which were distinct (used only once). It was becoming impossible to easily browse the cloud or find items within it. Due to the lack of tag search functionality, people were using the internet browser 'find' function to try to find items in the cloud; however, a few weeks later this was taking on average 10 minutes because the page took so long to load, and using the 'find' function became a very unsatisfactory option. Unfortunately, this could not be addressed during the first year. Despite the unsatisfactory nature of the tag cloud and the lack of guidelines, users continued to create and use tags at a far greater rate than was ever anticipated.

As expected, there was the usual range of spelling mistakes, inconsistencies in upper and lower case, variation in description of dates, mixed use of singular and plural, and creation of non-dictionary-word tags, e.g. xx1. Tonkin's research of tagging inconsistencies shows that in Flickr and Deli.ci.ous spelling mistakes (or terms not found in a range of dictionaries) appear in around a third of tags. We were not able to confirm this rate of spelling mistakes in the Australian Newspapers tags.

98% of the tags created were given the status of 'public' because users stated that they wanted to feel they may help the wider community. 2% were private. There was no discernable difference in the type of tags created as public vs. private. 14% of the taggers utilized the private tag feature. It appeared that there were two reasons for creating private tags: 1) either the users thought their tag would not be helpful to anyone else, or 2) they did not want anyone else to add tags to 'their tag' because they were using their own tags to track their research progress.

92% of the tags were added by registered users and 8% were added by anonymous (unregistered) users. The research by the Library of Congress and by the *steve.museum* also showed higher use of tagging by registered users than anonymous users. 57% of the tag pool was created by the top 10 'super taggers'. Super taggers create a significantly higher number of tags than other users (usually thousands). The presence of super taggers is not unusual. This correlates with the findings in the Library of Congress Flickr project where 40% of the tags were added by a group of 10 super-taggers. The top super-tagger entered more tags than all the anonymous users put together.

The overwhelming majority (estimated to be 80%) of distinct tags created were for personal names and were being used by genealogy researchers. This was clearly a different tagging pattern to that seen in museum and image collections, where subjects and geotags dominate. 37% of the tag pool was comprised of distinct tags. This was slightly higher than the findings of *steve.museum*, which had 32% distinct tags, and Library of Congress, which had 21% distinct tags.

It was observed that far more users (approximately 10 times more) opted to correct text than added a tag, and five times more articles were corrected than were tagged. This was perhaps because users understood that correcting the text had a more radical effect on search results than adding a tag did. Two of the four super-taggers, who were also super text correctors, said they added tags to articles at the same time as correcting text, because they thought it might help other people find things in a different way. They both said they were not using the tags for their own purposes, instead finding articles by keyword searching, but they hoped the tagging would help other people, and they found it easy to do as they went along. Other text correctors said they saw no point in tagging once they had corrected the words in which they were interested. A survey of the text correctors and user testing of the system had revealed that many users were confused by the three interaction options available (tagging, commenting and text correction). They were sometimes unsure which one to choose or "which one was best". Many users had never used features like tagging or rating or reviewing before and did not understand the purpose of tagging. This certainly implied that the majority of users would do one or another but would rarely use all three features together.

Users wanted to be able to see in the keyword search results list if articles had been corrected or tagged. This was not implemented until the end of the year. Although no moderation took place, as far as the ANDP team were aware no abuse of tags took place. Users were quick to report errors and inconsistencies, and since no users reported abuse, it was assumed there was none. The fear of abuse is probably unjustified since both the *steve.museum* research and the Library of Congress Flickr project research found a tiny percentage of inappropriate tagging.

Our understanding of what the 'top tags' were (viewable from the browse page) was open to interpretation. At the end of the year, it was apparent that the *most created tags* were quite different from the *tags used by the most users*. We were displaying the top most created tags – some of which had been used hundreds or thousands of times, but if Clay Shirky's ⁸ hypothesis in his article 'Ontology is overrated' is correct that users want to know: "is anyone tagging it the way I do?" then they would find the second type more useful. Interestingly, there was a direct correlation between the second type (*tags used by the most users*) and the most frequently used search terms, i.e. the way people think when they are looking for things is the same as the way they think when they are describing things.

Users want the tags to be of benefit to everyone, and they think consistency, guidelines and moderation is the key to this. Whether they are right or not is hard to tell. Clay Shirky says that "Tagging gets better with scale". Perhaps we should not get too hung up on guidelines and just do it. Shirky also says "If there is no shelf, then even imagining that there is one right way to organise things is an error". In the digital, shared space everything is different from the library with shelves.

A summary of the observations made during the first year of tagging are below:

- Tagging is a very popular activity and one that users want to do. It is seen as a good thing by users.
- Tagging appears primarily to benefit individual users as a way of tracking their own useful articles.
- The large majority of users choose to tag items as 'public' rather than 'private' in case doing so helps other people.
- Users have expressed the desire that tags should benefit everyone as a way of data enhancement.
- The majority of tags are for personal names.
- The most heavily tagged articles contain lists of personal names such as casualties of war, electoral rolls, classified advertising, shipping lists, births deaths and marriages (family notices).
- The users had an urgent need for an agreed form of description for personal names and wanted to be directed on how personal names should be entered. When the need was not met, the user community established its own unwritten rules.
- The use of multiple words in a single tag was very common – particularly in the top 200 tags (by number of times created).
- Natural language order emerged as the preferred way to create tags for names (not reverse order as per usual library rules).
- Hyphens, apostrophes and numbers were commonly used in tags.
- Some users want a retrospective conversion of tags of personal names into an agreed consistent form, and they are prepared to help doing this.
- After a year, 37% of the tag pool was comprised of different (distinct) tags.
- 74% of distinct tags were assigned to an article only once.
- The majority of users urgently want to be able to search across tags to utilise them more fully, e.g. to find all the tags that contain the same surname.
- Users want the tag layer to be utilised in an advanced article search, e.g. to be able to specify to search across text and tags in an article.
- Users understand that tags are a user layer created by other users and that the tags may not be correct or accurate.
- Private tags make up only 2% of the tags for the Australian Newspapers service.
- Text correctors do not always tag; in fact, in most cases more tagging is done by searchers than by text correctors.
- Only 4 of the top 20 taggers were also top 20 text correctors.
- At least 4 of the top taggers were serious professional researchers and made themselves known to the ANDP team.

- During the same period, there were around 5,000 users correcting text and an estimated 500 users tagging. More users were correcting text than tagging. 216,093 articles had been text corrected whilst only 38,874 articles were tagged.
- Serious researchers found the tagging feature essential. At present there are two very significant examples of this – one is the tagging of soccer articles where the researcher has created an extensive hierarchy for soccer, and the other is the group tagging by members of the Light Railway Research Society for Australia using the tag LRRSA and additional subject tags. Both have thousands of tags.
- Tagging enables group research to happen effectively with users who are geographically distributed but who share the same interests.
- Library staff were surprised by the immediate uptake on tagging and the large volume of tags created (especially when no publicity had taken place).
- The ANDP team had not thought that tagging of full-text searchable text would be so popular.
- No moderation of tags took place, and yet no user reported abuse of tags during the one-year period.
- The most commonly used tags (by number of people used) almost exactly match the most common search terms – the most common by a long way being ‘murder’.
- The tagging community wanted to be able to socially engage with each other via a communication channel.
- The tags with the highest usage by creation are often used by a single person only and do not match the tags with highest usage by number of times used by different users.
- The most common tags (a combination of the two top tag lists, by number of times assigned and number of users assigning) reflect Australian history and culture as well as the convict past and are: Murder, Death, Marriage, Drowning, Suicide, Shooting, Shipwreck, Shark attack, Fire, hanging, horse accident, cricket, gold, mining, ticket of leave, Sydney, Melbourne, Bendigo, Canberra.
- 57% of the tag pool has been created by the top 10 ‘super-taggers’.

7. Tagging enhancements suggested by public users

Within the first 12 weeks many of the public testers/users e-mailed the ANDP team saying that they urgently wanted the following:

1. Guidelines for tagging (how to enter, what rules to use especially for family names)
2. The ability to search tags (especially for names)
3. The ability to edit other peoples tags to make them conform
4. Something better than the tag cloud.

The user’s perceptions of the priority of these items did not change throughout the year. Other suggestions were also made. A complete list of all suggested enhancements for tagging is below. Those marked in green were implemented before the year ended. Enhancement requests for a single feature received from many users were given a high priority. The team were all in agreement that the ability to search tags and improvements to browsing the cloud were needed, but unfortunately they could not take action on these items since there were other more pressing priorities to be addressed first.

Tagging Feature- suggestions for enhancements	Suggested by	Priority
1. Guidelines for tagging (how to enter, what rules to use, especially for family names)	Public	
2. The ability to search across just tags (especially for names). The tag cloud is too big to browse.	Public ANDP team	High
3. The ability to search across tags, especially for names	Public ANDP	High
4. The ability to edit other people's tags to make them conform or to remove typos	ANDP team Public	Medium
5. The ability of public users to be 'moderators' of tags – tidying up inconsistent tags	Public	
6. Something better than the tag cloud	Public ANDP team	High
7. In advance search the ability to choose the search layers, e.g. text and tags and comments or combinations of these	Public	
8. Ability to define if you want to see your tags in a list or a cloud	ANDP team	
9. Give tags property types, e.g. personal names, place names, events so that you can search on tags, e.g. 'Scotland' as a person not a place (if you could search tags), or to be able to browse through the types	Public	
10. When typing in a new tag, for a suggestion to be given of similar tags already applied	ANDP team	
11. To have a spellchecker working when creating new tags	ANDP team Public	
12. Make suggestions when people try to add tags for synonyms, e.g. for WWI use first world war, etc.	ANDP team	
13. A 'related' tag feature for synonyms or similar tags, e.g. there are several thousand with SOCCER – subheading.	ANDP team	
14. Ability to add tags at page level and at issue level as well as article level	ANDP team	
15. Ability to follow through to other articles/pages/issues with the same tag when you are at the article level tag within the article	Public ANDP team	
16. Standardise personal names in the tags and guidelines for how to enter names	Public	
17. Add tags/comments to a specific place in the article (for long articles), e.g. when a user wants to pinpoint a name in a classified ad, family announcement, etc. (At present all tags are associated with the first line of the article.)	ANDP Public	

Tagging Feature- suggestions for enhancements	Suggested by	Priority
18. Ability to print/save own tags	Public	
19. Ability to print/save the full text of the article, including tags	Public	
20. Length of tag (characters) to be increased. 30 characters is not enough. (This was done and the limit is now 60 characters.)	Public	High
21. Number of tags an article can have to be increased/unlimited. 50 is not enough, especially for long articles with names. (This has been completed and the number of tags per article is now unlimited.)	Public	High
22. Show which articles have been tagged in results list. Show the last five tags added plus the total number. (Completed.)	Public	High
23. Tag where events happened on map (linked to geospatial visual searching)	ANDP	
24. Ability to Geotag using co-ordinates in tag	Public	
25. How can users keep track of their research? (Users appear to be using tags for this purpose and there could be much better ways to do this.)	ANDP	High
26. Ability to keep track of everything done (e.g. tagging, commenting, text correction) indefinitely in the user profile, not just to show the last 10 things done. Option to view by month/week/year the number of corrections/tags or ALL, in a list or a cloud	Public	
27. Ability to manage one's own tags page	ANDP team	
28. Ability to make bulk changes to one's own tags	ANDP team	
29. View a list of top taggers (like text correctors hall of fame)	ANDP team	
30. View top tags – by number added and/or by number of different users who have used (very different lists)	ANDP team	
31. Tagging multiple items at once	ANDP team	

8. Future development of tagging at the National Library of Australia

The National Library of Australia has decided that:

- Tagging will continue in the Australian Newspapers service.
- Tagging will be implemented across all other library collections before the end of 2009 if possible.

I have also suggested that the following activities take place:

- User activity with tags in Australian Newspapers is continually monitored.
- Tagging and other web 2.0 features are actively promoted to users for a) improving data quality for all users by the adding of layers, and b) social engagement with the organisation and collections.
- Searching tags is implemented as a priority.
- Searching user generated content (tags, comments, corrections) in combination mixes with library generated content is enabled.
- Strategies to minimise spelling errors and tag inconsistencies at tag creation point are implemented in preference to strategies to enable users to 'tidy up' tags.

- The rest of the public suggestions are evaluated and implemented if feasible. (As far as the ANDP team is aware, all suggestions are technically feasible.)
- A survey of users is carried out to find out if private tags are really needed and understood, or if it is an unnecessary level of complexity.
- The 'unwritten' guidelines developed by taggers are provided in written form to users on the site.
- Retrospective clean up of personal names tags, if necessary, is done by digital volunteers.
- The positive outcomes, lessons learned, and issues to be solved for social metadata in the wider library and archive context are evaluated. For example, can tags be shared between organisations, and should there be an international tag consortium? This may enable libraries to pre-populate their catalogues with social metadata from LibraryThing, Amazon, the National Library of Australia or other organisations.
- The ANDP team should participate in the RLG Social Metadata Working Group⁹, and other appropriate international forums.
- The National Library of Australia's tagging data is shared with any institution that wants to undertake further research on tagging.

9. Conclusion

The observations show that there were both similarities and differences in tagging activity and behaviours across a full text collection as compared to the research done on tagging in image collections. Similarities included that registered users tag more than anonymous users, that distinct tags form 21-37% of the tag pool, that 40% or more of the tag pool is created by 'super-taggers' (top 10 tag creators), that abuse of tags occurs rarely if at all, and that spelling mistakes occur fairly frequently if spell-check or other mechanisms are not implemented at the tag creation point. Notable differences were the higher percentage of distinct tags used only once (74% at NLA) and the predominant use of personal names in these tags. This is perhaps related to the type of resource (historic newspaper) rather than its format (full-text). It is likely that this difference may be duplicated if tagging were enabled across archive and manuscript collections. There was an expectation from users that since this was a library service offering tagging, there would be some 'strict library rules' for creating tags, and users were surprised there were none. The users quickly developed their own unwritten guidelines. Clay Shirky suggests "Tagging gets better with scale" and libraries have lots of scale – both in content and users. We shouldn't get too hung up on guidelines and quality. I agree with Shirky that "If there is no shelf, then even imagining that there is one right way to organise things is an error".

The experience of the National Library of Australia shows that tagging is a good thing, users want it, and it adds more information to data. It costs little to nothing and is relatively easy to implement; therefore, more libraries and archives should just implement it across their entire collections. This is what the National Library of Australia will have done by the end of 2009.

10. References

¹ Steve.museum <http://steve.museum>.

² Trant, J. (2009). Tagging, Folksonomy and Art Museums: Results of steve.museum's research. <http://verne.steve.museum/SteveResearchReport2008.pdf>.

Institute of Museum and Library Services. (2009). Researching Social Tagging and Folksonomy in Art Museums: Final Report LG-25-06-0131-06 October 1, 2006- January 1, 2009. (not online).

³ Australian Newspapers service <http://newspapers.nla.gov.au>.
Australian Newspapers Digitisation Program project website <http://nla.gov.au/ndp>

⁴ Clayton, S; Morris, S; Venkatesha, A; Whitton, H. (2008) *User Tagging of Online Cultural Heritage Items: A project report for the 2008 Cultural Management Development Program prepared by the Australian War Memorial, the National Library of Australia, the Royal Australian Mint and the National Archives of Australia.*

<http://www.nla.gov.au/openpublish/index.php/nlasp/article/view/930/1205>.

⁵ Springer, M., Dulabahn, B. Michel, P. Natanson, B., Reser, D., Woodward, D., et al (2008). *For the Common Good: the Library of Congress Flickr Pilot Project.*

http://www.loc.gov/rr/print/flickr_report_final.pdf.

⁶ Holley, R. (2009) *Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers*, National Library of Australia, ISBN 9780642276940

http://www.nla.gov.au/ndp/project_details/documents/ANDP_ManyHands.pdf.

⁷ Guy, M., Tonkin, E. (2006) Folksonomies, Tidying up Tags? *D-Lib Magazine*, January 2006, volume 12 number 1. <http://www.dlib.org/dlib/january06/guy/01guy.html>.

⁸ Shirky, C. (2005) *Ontology is overrated: Categories, Links and Tags.*

http://www.shirky.com/writings/ontology_overrated.html.

⁹ RLG Social Metadata Working Group

Background: <http://www.oclc.org/programs/ourwork/renovating/changingmetadata/aggregating.htm>.

Overview of progress June 2009: <http://www.oclc.org/programs/events/2009-06-02j.pdf>.