Harvesting and semantically tagging media releases from political websites using web services

Peter Neish
Systems Officer
Victorian Parliamentary Library
peter.neish@parliament.vic.gov.au

Abstract

The Victorian Parliamentary Library built an application to automatically locate and download media releases using the RSS feeds of political parties. Using the OpenCalais web service to tag items with relevant semantic metadata has improved the Library's workflow and increased the usefulness of the database, by helping users discover directly linked items. The benefits of using these tools and problems encountered are discussed.

Introduction

Media releases are an important part of the political process. They are used by government and non-government parties and independents to publicise policy initiatives, announce funding for particular projects as well as providing advance notice or background information on events. Importantly, they place on the public record a party's position on an issue at a particular point in time. This makes them a valuable source of information when tracking issues or commitments made by politicians in the past.

The Victorian Parliamentary Library has maintained an archive of political media releases since late 1992. These media releases were stored in various isolated systems depending on where they originated. Sometimes the Library was able to obtain copies of the media release database from government, but in many instances the media releases were added to the database and indexed by the Parliamentary Library. In addition, many media releases were readily available on the web and a decision had been made not to add to the database items that were already available in a publicly-searchable database. However, whenever there is a change of government, there is no guarantee that previous media releases will continue to be made available in a publicly searchable archive, and on a number of occasions the incoming government has decided not to retain the previous government's media releases. An aim of this project was to capture all media releases into a central database so that the Library can ensure that they are not lost from the public record. A single database would also help Library users and staff search more easily across different parties and years, and it was hoped that it would provide a faster and more consistent user experience.

The indexing of media releases across these databases was incomplete and variable. The Library indexed items as they were added to the database but government-supplied media release databases did not come with subject terms, so these records were not indexed due the excessive staff time this would have demanded. It was becoming apparent that the number of media releases was increasing over time (Figure 1), so the project team wanted to look at whether an automatic system could be used to add tags without human intervention.

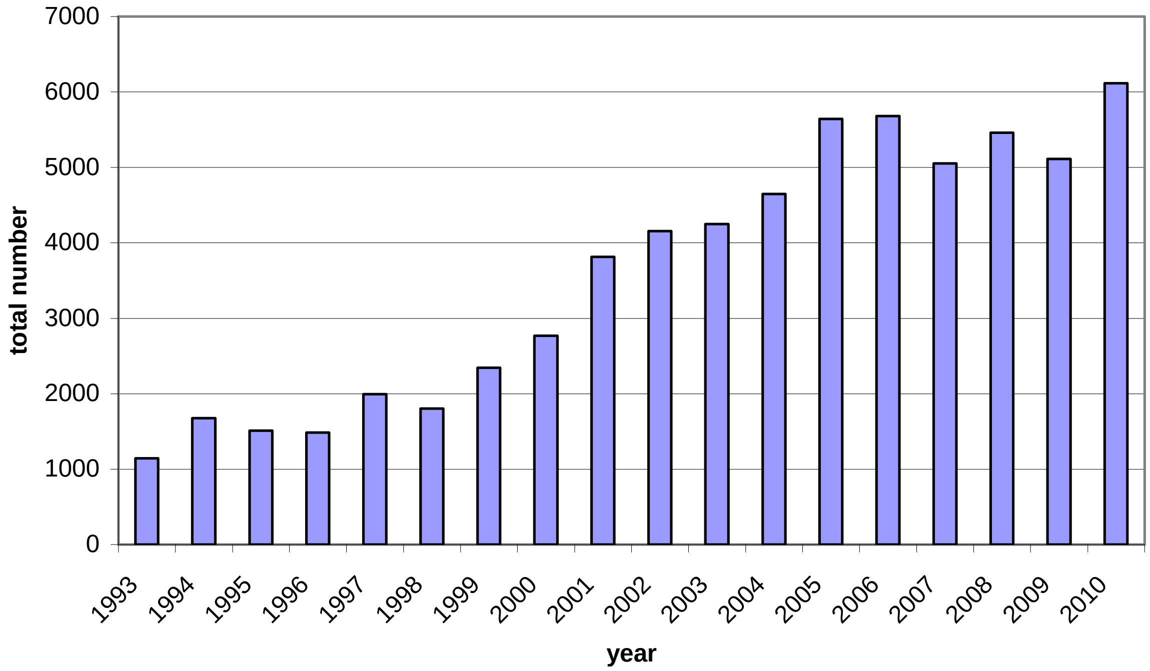


Figure 1 - Total Media Releases per year added to the Victorian Parliamentary Library Media Release database.

The Library is concerned with archiving the media releases from political parties that have members currently sitting in the Victorian Parliament. Without exception, all of these media releases are available on the internet, either on party websites or on the Department of Premier and Cabinet website. Importantly, all of these websites have an RSS feed that can be monitored to see when new media releases had been issued. The project team decided to investigate whether the standard RSS interface could be used to automate some of the workflow involved in capturing media releases.

Project Methodology

The project was run using an agile methodology (Martin Fowler & Jim Highsmith 2001) and made use of rapid prototyping. A team of two people ran the project: a project manager (the Library's intranet librarian, Robin Gallagher) and a developer (Peter Neish). There were also regular meetings with a wider group representing users and management.

There was significant user testing for each iteration of the project, by both Library staff as well as non-expert clients. This was crucial in ironing out bugs in the user interface between different versions of browser software.

The project consisted of two distinct parts. The first was building the system to automatically harvest the media releases, extract the metadata into a database, generate a PDF of the media release, and build the associated user interface.

The second part, which arose directly from the first, was to develop a process to automatically extract relevant terms and subject headings from the text of the media release to aid in searching and browsing the media release collection.

Where possible, the team would use open source software and open standards in an effort to make the systems more easily customisable, flexible and interoperable. This was also in keeping with the requirement to consider open standards and open source software in the Victorian Government (Victoria 2009).

Part 1: Harvesting Media Releases

RSS (originally RDF Site Summary, but commonly dubbed Really Simple Syndication) is a standard web format used to publish recently updated works. Most Content Management Systems (CMSs) come with a built in RSS feed and this is the case for the websites of current members of the Victorian Parliament. As a minimum, an RSS feed contains a title, a short description of the new item, the date published and a link to the item, which was used as the source data for the media release database.

A standard Java servlet application was built to monitor the RSS feeds of all Victorian political party websites for new content. Each RSS feed was checked once an hour and if new content was found, the feed was processed. Due to the standard nature of the RSS format, it was a possible to leverage an existing Java library (Rome (http://java.net/projects/rome/) to connect to and parse the RSS feed. The Rome library was able to easily connect to and parse different versions of the RSS protocol.

Despite being a standard protocol, the actual content of the RSS fields can be implemented slightly differently by different sources. For example, the description field sometimes contains the entire text of the media release, or just a one-line summary. In addition, dates can sometimes be configured to deliver local time or Greenwich Mean Time. To address some of these concerns a number of filters were developed using Yahoo! Pipes (http://pipes.yahoo.com) to manipulate the RSS feed before it was accessed by the Library's application. Yahoo! Pipes is a free, web-based service that allows the user to aggregate, manipulate, and mashup content from around the web. RSS feeds can be 'piped' through various commands to combine, sort, modify or translate into a new feed. The pipes created for this project were fairly simple and involved manipulating only one or two fields; however, the Yahoo! Pipes environment allows the creation of quite complex and innovative data feeds.

When new media releases were detected, the page containing the media release was converted into a PDF file using an open-source tool called wkhtml2pdf (http://code.google.com/p/wkhtmltopdf/). This tool allowed the system to create the PDF exactly as if a user had printed the media release from the website. After trialling a number of PDF conversion products, the project team decided to use wkhtml2pdf. Non-valid HTML and poorly styled web pages meant that many PDF generators either did not work, or produced PDFs of very poor quality. Wkhtml2pdf makes use of the webkit rendering engine (http://www.webkit.org/) to produce the PDF and can be set to make use of the print style sheet for a web page. It was by far the best tool for creating PDFs from the wide variety of HTML found in the wild.

The harvesting application captured metadata from the RSS feed including title and date of release, the author, the site that it came from, the party, the URL of the original item as well as the full text. All of this was then stored in a DB/Textworks database. Although DB/Textworks is a proprietary database, it provides an API (application programming interface) in which XML can be sent to it via HTTP. Data could therefore be loaded directly into the database using a custom-built Java wrapper around the DB/Textworks XML web service. This meant that databases could easily be changed in the future, if needed, by modifying the wrapper to work with a different database.

A complete check of all sites and the conversion of the media releases take between one and three minutes depending on the number of new releases detected. This compares with the many hours it would have taken to manually type in records into a database and generate or upload PDF files to the server. Records can still be edited or deleted if there are duplicates or additional information needs to be included, but in effect the management role for this database has changed from data entry to data curation and many hours per day of staff time have been saved.

Storing all media releases in a single database has lead to an improvement in usability for Library users. Library staff and clients no longer have to search multiple databases for the information they require and can now search or browse a single repository to find the media releases they are after.

Another advantage in storing media releases in a single database is that for the first time it is possible to get a detailed breakdown of how many media releases are being added over time and from which party. Over the last 18 years, there has been a six-fold increase in the issuing of media releases. There are now over 6000 media releases issued each year, a figure that equates to an average of about 16 per day. Government parties release on average twice as many media releases as the opposition parties, although this varies depending on the election cycle (figure 2).

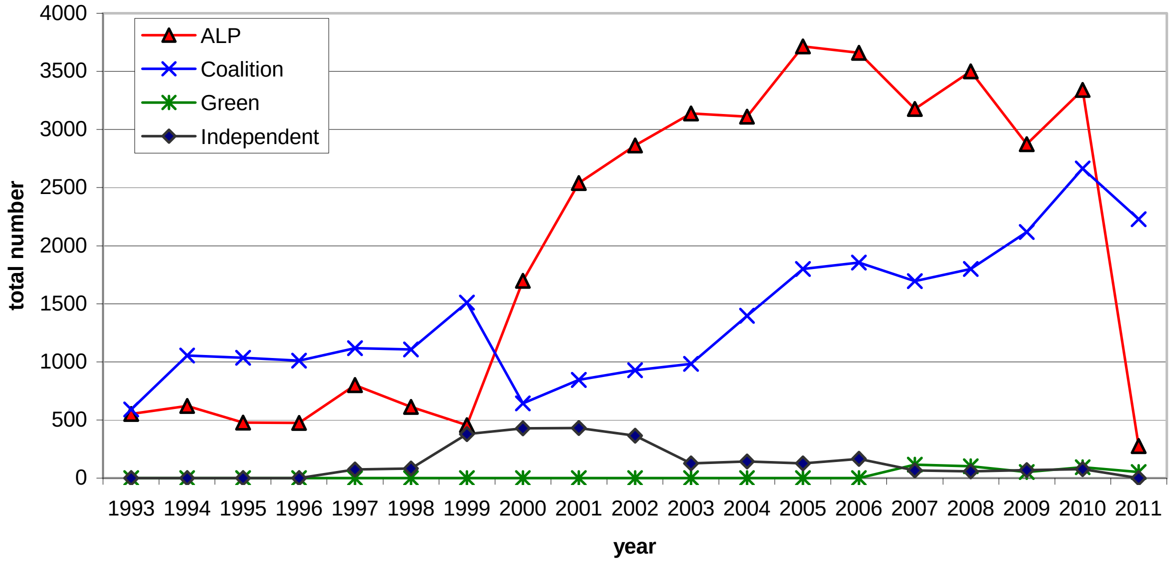


Figure 2 - Media Releases for each party by year added to the Victorian Parliamentary Library Media Release database.

The number of media releases being harvested meant that it had become too time-consuming to manually index each item. The second part of the project was to look at ways to automate tagging of content.

Part 2: Semantic Tagging

The concepts of the Semantic Web and Linked data have emerged as the next development of the World Wide Web, and comprise a set of best practices for publishing and connecting structured data on the Web (Berners-Lee 2009). Within the field of Information Extraction, much effort has gone into developing tools to automatically extract "named entities" (e.g. people, organizations, places and events) from natural language text. In some cases automatic systems have been developed that can tag records with fewer errors than human indexers, albeit within a narrow context (Garrido et al. 2011). Systems that use thesauri and ontologies allow named entities to be linked unambiguously to entities in the semantic web using semantic links.

A number of web services and stand-alone applications were investigated to see how well the process of tagging media releases with named entities could be automated.

The primary criterion for evaluating the web services was their ability to extract an appropriate number of relevant terms for the text provided. Too few results meant that the service was not as useful as it could be, while too many results possibly indicated irrelevant or incorrect results. Another criterion was how many false-positives or false-negatives were occurring, and also whether the results were weighted or gave an indication of relevance or confidence.

Media releases are good candidates for this kind of entity extraction as the web services have usually been built to work with news articles. A news article is very similar to a media release (and sometimes contains large extracts from a corresponding media release), so it was not surprising that trials showed good results with many of these web services.

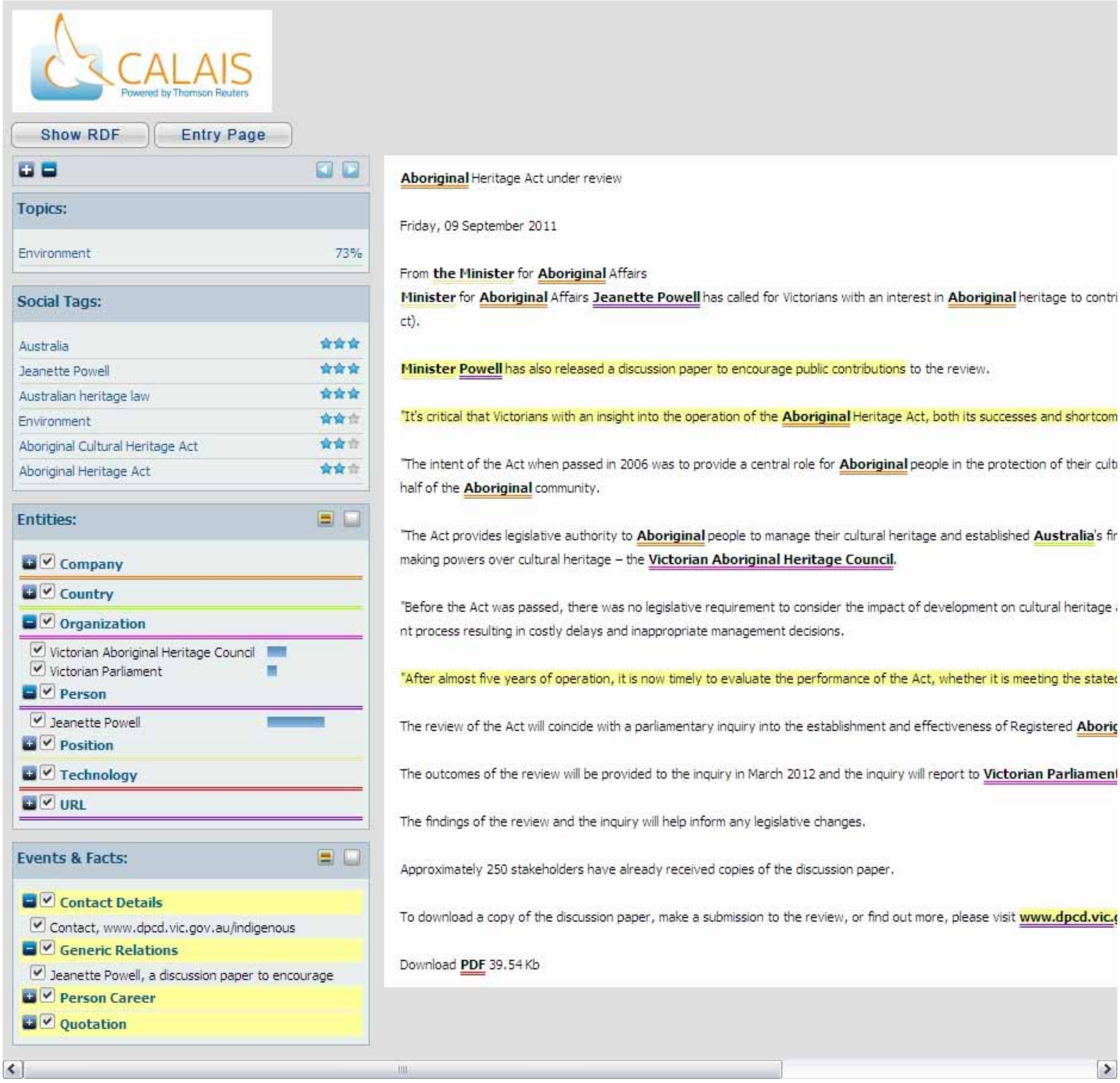
It is not the purpose of this paper to provide a detailed comparison, however the candidates examined were:

- AlchemyAPI (http://www.alchemyapi.com/)
- Evri (http://www.evri.com/)
- OpenAmplify (http://www.openamplify.com/)
- OpenCalais (http://www.opencalais.com/)
- Yahoo Term Extraction (http://developer.yahoo.com/search/content/V1/termExtraction.html)
- Zemanta (http://www.zemanta.com/)

Testing showed that the OpenCalais web service provided by Thomson Reuters gave the best results for the Library's data and returned a reasonable level of tags per item (usually 10 - 30 per item) and had minimal false matches.

A scoping study undertaken by the eResearch Lab at the University of Queensland, also chose OpenCalais as a candidate for more detailed investigation because of its "relative stability, widespread adoption, previous applications, ease-of-use, flexibility and comprehensive documentation and open source community support" (Gerber et al. 2010). However, despite its integration into open source products, OpenCalais itself is a closed-source product. Entities retrieved by OpenCalais are defined in a proprietary ontology hosted by them and there are no public reports in existence that explain details of the algorithms used (Adrian & Schwarz 2011).

OpenCalais is accessed through an application programming interface (API) and like other services imposes limits on usage. However these limits are quite generous in comparison: the current limits are 50,000 calls per day and 4 calls per second (OpenCalais 2011). As well as entities such as people, places, organisations, events, topics and tags it can also extract relationships between entities (e.g. Person X resigned from position Y). Their document viewer (http://viewer.opencalais.com/) gives an easy to grasp overview of what information can be extracted from a block of text (figure 3).



Topics:

Environment	73%

Social Tags:

Australia	★★★
Jeanette Powell	★★★
Australian heritage law	★★★
Environment	★★☆
Aboriginal Cultural Heritage Act	★★☆
Aboriginal Heritage Act	★★☆

Entities:

☑ Company
☑ Country
☑ Organization
☑ Victorian Aboriginal Heritage Council
☑ Victorian Parliament
☑ Person
☑ Jeanette Powell
☑ Position
☑ Technology
☑ URL

Events & Facts:

☑ Contact Details
☑ Contact, www.dpcd.vic.gov.au/indigenous
☑ Generic Relations
☑ Jeanette Powell, a discussion paper to encourage
☑ Person Career
☑ Quotation

Aboriginal Heritage Act under review

Friday, 09 September 2011

From **the Minister** for **Aboriginal** Affairs
Minister for **Aboriginal** Affairs **Jeanette Powell** has called for Victorians with an interest in **Aboriginal** heritage to contri
ct).

Minister Powell has also released a discussion paper to encourage public contributions to the review.

"It's critical that Victorians with an insight into the operation of the **Aboriginal** Heritage Act, both its successes and shortcom

"The intent of the Act when passed in 2006 was to provide a central role for **Aboriginal** people in the protection of their cult
half of the **Aboriginal** community.

"The Act provides legislative authority to **Aboriginal** people to manage their cultural heritage and established **Australia**'s fir
making powers over cultural heritage – the **Victorian Aboriginal Heritage Council**.

"Before the Act was passed, there was no legislative requirement to consider the impact of development on cultural heritage a
nt process resulting in costly delays and inappropriate management decisions.

"After almost five years of operation, it is now timely to evaluate the performance of the Act, whether it is meeting the stated

The review of the Act will coincide with a parliamentary inquiry into the establishment and effectiveness of Registered **Aboric**

The outcomes of the review will be provided to the inquiry in March 2012 and the inquiry will report to **Victorian Parliament**

The findings of the review and the inquiry will help inform any legislative changes.

Approximately 250 stakeholders have already received copies of the discussion paper.

To download a copy of the discussion paper, make a submission to the review, or find out more, please visit **www.dpcd.vic.**

Download **PDF** 39.54 Kb

Figure 3 - OpenCalais document Viewer

In the above example, a media release concerning a review of the Aboriginal Heritage Act has been run through the document viewer. In this case the media release has been assigned to the 'Environment' topic (OpenCalais has 18 top-level topics). OpenCalais also assigns 'Social Tags', which attempt to simulate how a user would have tagged the item. In this sense it is not true semantic analysis, but rather provides an additional way for the content to be categorised. In the above example it did quite a good job applying the social tags: Australia, Jeanette Powell, Australian Heritage Law, Environment, Aboriginal Cultural Heritage Act and Aboriginal Heritage Act. Next come the entities, which are grouped by Company, Country, Organisation, Person, Position, Technology and URL. Then finally are events and facts which include contact details, generic relations (e.g. Jeanette Powell released a discussion paper), person career (e.g. Jeanette Powell is a Minister) and quotations from the release.

OpenCalais can return results in numerous standard formats: RDF (serialised in XML or N3), simple XML (extensible markup language), JSON (JavaScript Object Notification) or microformats. Although RDF (Resource description Framework) provides the most comprehensive and powerful data format, it is the most difficult to work with due to its verbose nature and the fact that OpenCalais have introduced additional classes into the RDF to represent the many N-ary (i.e. between the subject and two or more values) relationships in the data (W3C 2011). For example, an article might be assigned to the category 'Environment', but this category also comes with a category ID, category type and a relevance score. The number of RDF statements grows very quickly as the complexity of the data increases and most media releases needed between 500 and 1000 RDF statements to describe the item. Therefore, for a fairly modest database of 150,000 records, the system would be required to store over a billion triples, which while possible, has major performance and storage overheads.

In the end, the XML format was chosen, which returned about 10 - 30 tags per article (figure 4) and these tags were stored in a MySQL database. While this loses some of the information present in the RDF output, it was consistent with existing systems already running at the Library on the Linux and MySQL platforms. The Media Release application does capture the URIs (uniform resource identifiers, of which locator or URL is a form) for many of the attributes, so that unambiguous links can be made to existing concepts in the linked data ecosystem at a later stage if needed.

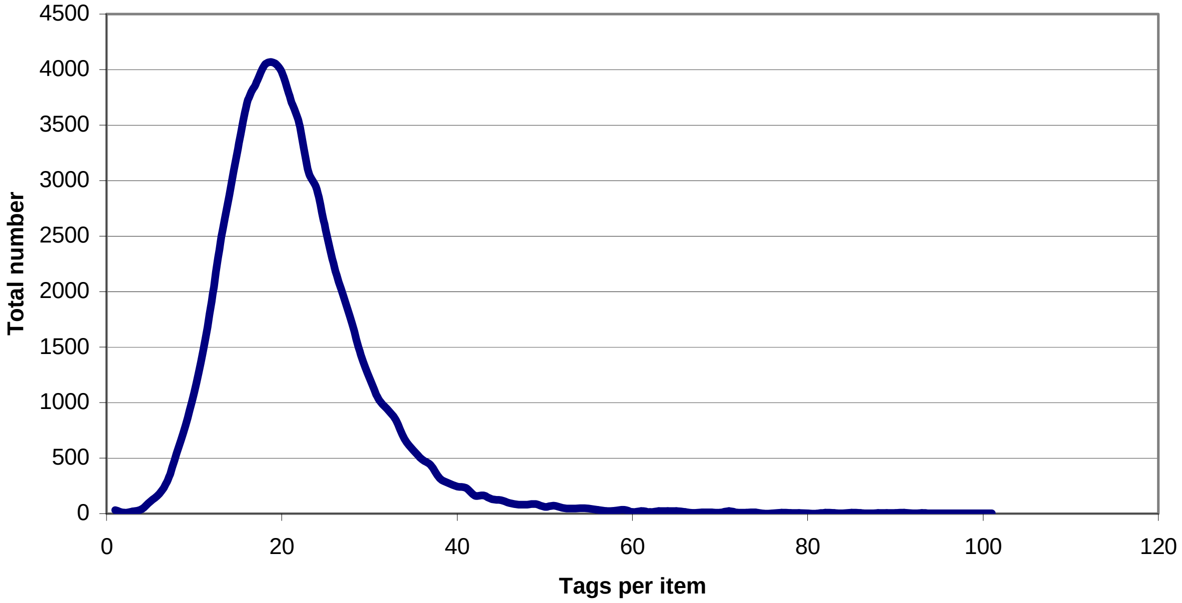


Figure 4 - Number of tags assigned to each media release

An analysis of a sample of tags (figure 5) shows that on average only 4% of tags have been assigned incorrectly, i.e. where a subject term has been incorrectly applied to an article. This rate of false matches has a very minimal impact on the user experience searching the database. Most of these incorrect matches are due to organisations or place names being incorrectly determined to be in Europe or the United States rather than Australia. For example, mention of the Austin Hospital almost always results in the release being tagged with Austin, Texas, while Victoria often gets tagged as Victoria, Seychelles. OpenCalais does try to prevent this by analysing the other words and phrases in the text to try and disambiguate which country is being referred to, but it doesn't always work (even when we append the text 'Australia' to the content being parsed).

The Library has not implemented a formal process for identifying errors, however if it is noticed that an error is occurring repeatedly then it is possible to build in a filter to ignore that particular tag. However, for many tags it is not possible to apply a global filter because there may be cases where tag is actually correct. For example, many media releases do in fact mention Austin, Texas.

An additional 6% of tags are repetitions or synonyms (for example, Victorian Government, Coalition Government and Victorian Coalition Government), and while it is apparent in this context they are probably synonyms, there are cases where similar terms might be subtly different, so it is probably better that they appear in the results where their relevance can be assessed by the user rather the information be lost.

A further 5% are general words that are uninformative in this context. For example, Premier or Minister or Government appear in almost every media release, so they are not really adding any useful information. This is addressed in the user interface by filtering out some of the most common redundant terms.

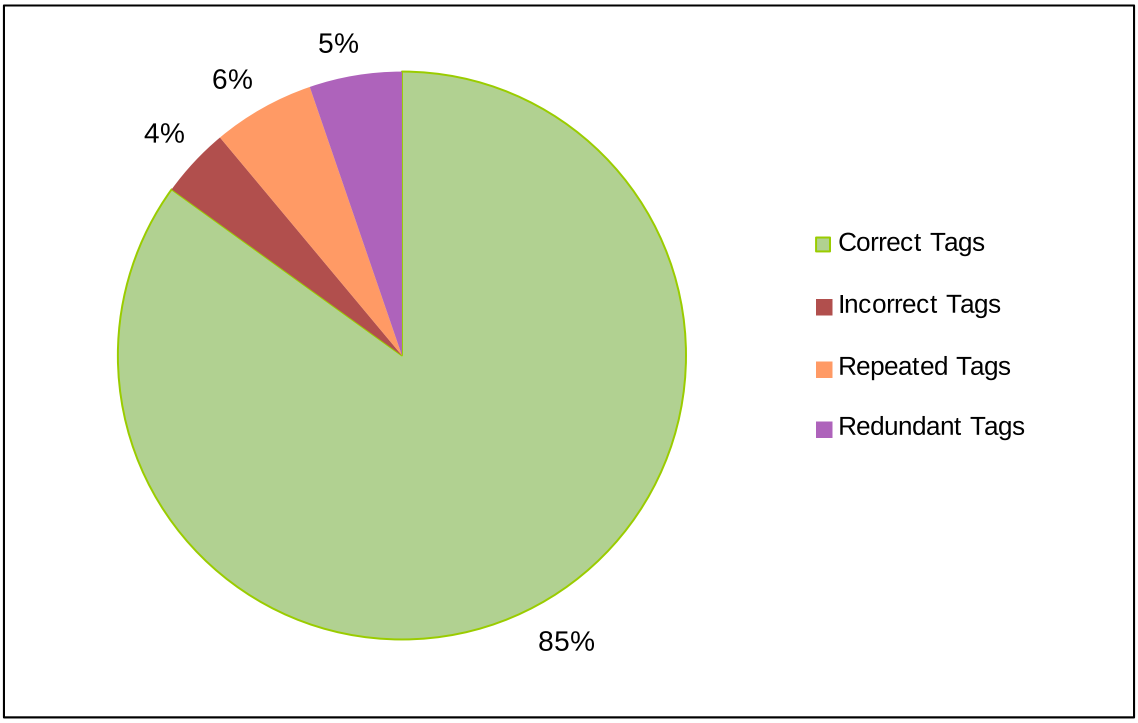
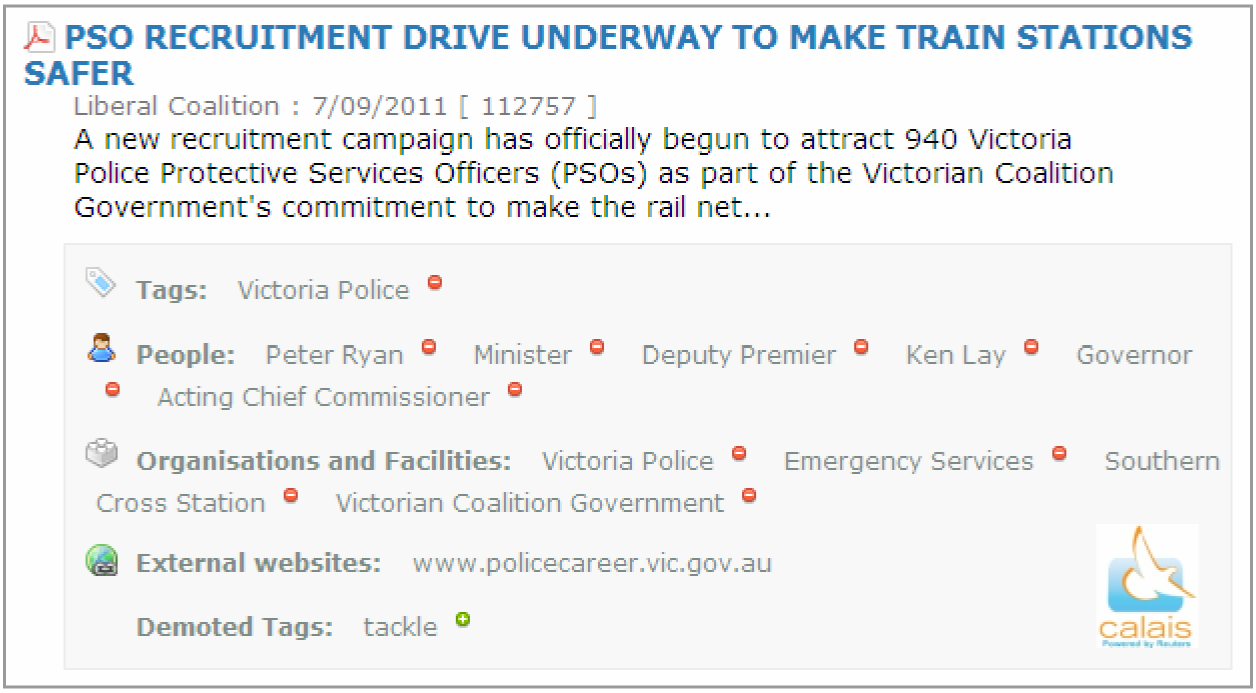


Figure 5 - Quality of tags generated by OpenCalais

User Interface

The tags generated for the media releases are used to give users additional ways to browse related media releases, and appear when viewing the search results in the media release database (figure 6).



PSO RECRUITMENT DRIVE UNDERWAY TO MAKE TRAIN STATIONS SAFER

Liberal Coalition : 7/09/2011 [112757]

A new recruitment campaign has officially begun to attract 940 Victoria Police Protective Services Officers (PSOs) as part of the Victorian Coalition Government's commitment to make the rail net...

Tags: Victoria Police ⊖

People: Peter Ryan ⊖ Minister ⊖ Deputy Premier ⊖ Ken Lay ⊖ Governor ⊖ Acting Chief Commissioner ⊖

Organisations and Facilities: Victoria Police ⊖ Emergency Services ⊖ Southern Cross Station ⊖ Victorian Coalition Government ⊖

External websites: www.policecareer.vic.gov.au

Demoted Tags: tackle ⊕

Figure 6 - screen shot of Media release database

Each tag is a clickable link that will run a search for that term in the database. False matches can be easily removed by any user clicking on the red minus symbol and the tag will be moved down to the 'Demoted Tags' section. In the above example, OpenCalais incorrectly identified the word 'tackle' to mean an entity of type Position (which are displayed under the heading People) and a user has demoted this tag. Users are able to restore any of these tags if they have been mistakenly demoted and the administrator can remove them from display all together.

OpenCalais has about 40 different entity types and many of these reflect the news and entertainment focus of the service. The project team chose to reduce the number of types of tags and group tags under eight main headings to make this easier for the user to navigate (table 1).

Heading in Media Release Database	OpenCalais types
Tags	Document Category, IndustryTerm, Social Tags, MedicalCondition, MedicalTreatment, Product
People	Person, Position
Organisations and Facilities	Company, Facility, Organization
External websites	URL
Events	Anniversary, Holiday, PoliticalEvent
Places	City, Continent, Country, NaturalFeature, Region, ProvinceOrState
Media and Television	EntertainmentAwardEvent, Movie, PublishedMedium, RadioStation, TVShow, TVStation
Sport	SportsEvent, SportsGame, SportsLeague,

Table 1: Media Release database headings and equivalent OpenCalais types

The following OpenCalais types are not used by the system: *Currency, EmailAddress, FaxNumber, MarketIndex, MusicAlbum, MusicGroup, OperatingSystem, PhoneNumber, ProgrammingLanguage, Technology,* as they do not provide relevant tags in this context.

Conclusion

Using freely available tools, the Library was able to build a system to archive the ever-increasing number of media releases produced by political parties. By automating the capture of these media releases and streamlining the workflow related to Media Releases, the Library has saved approximately two staff days per week in the management of this data.

Whereas previously there was a backlog of items to be added to the database, now the media release database is almost completely up to date (within 1 hour of being released). This allows for near real time alerting of relevant media releases to Library users. It also makes it simple to search across parties for media releases on the same subject.

The combination of open source tools and freely available web services has meant that the Library has been able to build a system without incurring additional software costs.

By using a web service, the Library is free of the hassle of keeping the software up to date. Improvements in the OpenCalais algorithm become available to the system immediately. A potential down side is that the organisation now has a reliance on a third party keeping the service available. However, the scale of the Thomson

Reuters operation and their open attitude to the user community indicates that there is minimal risk in this regard. If the worst happened, the system is modular enough that the Library could switch to a different tagging service with minimal code changes.

Library users can now make use of the tags to browse related items of interest and link unambiguously to information held in external databases about the entities tagged in the media releases.

The Library is now in the process of investigating the possibility of automatically tagging other content such as newspaper clippings and media files to help make it more discoverable and to link as much of the Library's content together as possible.

References

Adrian, B. & Schwarz, S., 2011. Using Suffix Arrays for Efficiently [sic] Recognition of Named Entities in Large Scale. In *Knowlege-Based and Intelligent Information and Engineering Systems*. 15th International Conference, Kes 2011. Kaiserslautern, Germany: Springer, pp. 420–429.

Berners-Lee, T., 2009. Linked Data-The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3), pp.1–22.

Fowler, Martin & Highsmith, Jim, 2001. The Agile Manifesto. *Software Development*, 9(August), pp.28–35. Available at: http://andrey.hristov.com/fht-stuttgart/The_Agile_Manifesto_SDMagazine.pdf [Accessed December 1, 2011].

Fowler, Martin & Highsmith, Jim, 2001. The Agile Manifesto. *Software Development*, 9(August), pp.28–35. Available at: http://andrey.hristov.com/fht-stuttgart/The_Agile_Manifesto_SDMagazine.pdf [Accessed December 1, 2011].

Garrido, A.L. et al., 2011. NASS: News Annotation Semantic System. In *23rd IEEE International Conference on Tools with Artificial Intelligence (ICTAIA 2011)*. Boca Raton, Florida (USA).

Gerber, A., Gao, L. & Hunter, J., 2010. *A Scoping Study of (Who, What, When, Where) Semantic Tagging Services*, eResearch Lab, The University of Queensland. Available at: http://itee.uq.edu.au/~eresearch/projects/ands/W4SemanticTagging-report-2011-02.pdf [Accessed September 22, 2011].

OpenCalais, 2011. API Usage Quotas. Available at: http://www.opencalais.com/documentation/calais-web-service-api/usage-quotas [Accessed September 22, 2011].

Victoria, 2009. *Inquiry into Improving Access to Victorian Public Sector Information and Data: Report of the Economic Development and Infrastructure Committee on the Inquiry into Improving Access to Victorian Public Sector Information and Data*, Melbourne: Victorian Government Printer.

W3C, 2011. Defining N-ary Relations on the Semantic Web. Available at: http://www.w3.org/TR/swbp-n-aryRelations/ [Accessed September 22, 2011].

Acknowledgements

The author wishes to acknowledge the support and advice of the project manager, Robin Gallagher and the other members of the project team: Margaret Galbally, Julie Gardner and Jon Breukel. Libby Hill analysed the tags generated by OpenCalais and produced figure 5 while on a student placement as part of the Post-Graduate Diploma in Information Management at RMIT.