

Write | Edit | Index April 2015

Grisoula Giopoulos
Assistant Director, Database and Media Services
Australian Parliamentary Library, Canberra.

Automatic indexing in the Commonwealth Parliamentary Library – Grisoula Giopoulos

Today I'm going to talk about how the Commonwealth Parliamentary Library in Canberra uses an automated system to select and index Newspaper clippings for the parliament. Why it was introduced, how it gets done and how effective it is. I won't go into the technical side of things but give you an overview of what happens to get the newspaper clippings into LAST and indexed. But first a run down on Parlinfo, which is the search interface our clients and the public use.

1. First off, Parlinfo is the search interface used by our clients, the members and senators, and their staff, as well as the parliamentary departments. It is also accessible to the public. **PP1_ Parlinfo and the datasets.** Most of the datasets in Parlinfo are freely available to the public but some of the library databases are restricted because of copyright reasons. For example, as a member of the public you can search the newspaper clippings database in Parlinfo and you will get the metadata, the title, source, date, journalist's name/s and subject headings but you won't be able to view a pdf of the newspaper clipping because of copyright restrictions.
2. **Contributors to the Parlinfo databases** are the chamber departments: Hansard, Notice papers, Votes and proceedings, Bills, etc. Within the library, DMS adds the material to all Library databases and indexes them. This is done manually by creating a record for each item and then indexing it. However, only the newspaper clippings are fully automated: from the ingestion of the newspaper clippings into the Newspaper clippings database from the supplier, to the selection and the indexing of those clippings.
3. The Parliamentary Library receives on average 1800 newspaper clippings every day but only about 250 – 300 of those clippings will be available in Parlinfo. **PP2 - List of newspapers titles covered.** The clips are prepared for us by a contracted supplier and delivered seven days a week. They provide all articles from the newspapers excluding the classifieds. They make no selections for us. They supply us with a text searchable pdf of each article and a separate xml file that contains all the metadata, that is, the title, author, source, date, section for all those clippings. We get an email notification by about 6.30am to tell us that this data is available on their web site and an automatic ftp process begins the download from their web site into our system, **LAST (Library Authoring System and Thesaurus)**. This material is then

ingested into LAST and is available to our clients by 7am from Parlinfo. They can search by the automatically assigned subject headings, title and source.

4. When these 1800 odd clips are ingested into LAST they go through a couple of classifiers. The first one is the selection classifier which actually does the first lot of selection for us. The 1800 or so clips are reduced to about 250 or so. The second classifier is the subject classifier and it assigns subject headings from the Library's thesaurus to most of the clips.
5. So how do these classifiers know how to do these things? Both the selection and subject classifiers were trained by being fed training data in 2010 when we went live with LAST. The training data was 3 months worth of newspaper clippings that had already been selected by the Library Databases team and the accompanying subject headings assigned to those articles by them. So the classifiers built their statistical models from these documents and the words they contained. This training continues daily. Overnight the classifiers analyse what actions have taken place on the newspaper clippings of that day, such as what was selected and what subject headings were assigned and new model files are built overnight and applied the following day to the new clippings that have been ingested. So it learns from the actions of the indexers and also from its own actions.
6. The first classifier **selects** the newspaper clippings by deciding whether a clipping is "interesting" or "uninteresting". It does this by using a statistical model of words it expects to find in "interesting" documents and it has a separate statistical model for words it expects to find in "uninteresting" documents. When it analyses documents it calculates a score for "interesting" and "uninteresting" words within the document. The "uninteresting" words are subtracted from the "interesting" word score and will show as a negative number. If the value is a **positive** number then the document is regarded as being of interest and will get a **Yes decision** while if the score is **negative** it will be regarded as "uninteresting" and be classed as **No – not interesting**. If the classifier finds that a document has no interesting or uninteresting words it classes it as **Unsure**. This is because it hasn't come across this content before. **PP3 & PP4 INTERESTING - Positive vs Negative words;**
PP 5 & PP6 UNINTERESTING -Positive vs negative words.
7. The **subject classifier** works in a similar way but it has a statistical model for every subject heading that it has seen in the training data. The subject headings come from an in-house thesaurus. The classifier analyses a document by looking at the words in that document and calculates a weighting for every possible word and then suggests and assigns the subject heading associated with those words. The human indexers role is to check these headings and make sure they apply to the article.

8. Within the subject classifier's suggestions are two categories of suggestions, **Recommended** headings and **Others**. Subject headings with weightings above a threshold figure are displayed as Recommended headings while those with weightings below this threshold are displayed as **Others**. The significance of this is that the headings that it has Recommended are automatically assigned to that article and are attached to it first thing in the morning. The article is published to Parlinfo. Our clients can then search Parlinfo for these articles by subject heading, title, journalist's name, etc. The 'Others' headings are not assigned but they are there as a suggestion for the human indexer to assign if relevant. If the human indexer uses this subject heading then this forms part of the classifier's overnight training and will learn to apply that subject heading to articles with those words in the content.
9. About 25% of newsclips each day do not get assigned any subject headings by the classifier and the indexers index these before checking the classifier indexing. The reason the classifier doesn't provide subject headings is that none of the words in the article have enough weighting to be associated with a subject heading.
10. A bit more about the **threshold** between the 'Recommended' terms and the 'Others'. The threshold is numerical and can be adjusted by us manually. This was initially set by the software developers. What now happens is that staff notice and give feedback that suitable headings may be suggested in the 'Others' section and so sit **below** the threshold and are not applied to the article. This has been happening every 18 months or so, so the system is quite consistent. **PP7, 8 & 9 Bali Nine – January; February; March comparisons.**

PP10 Change in weighting for 'DRUG TRAFFICKING' and words associated with the subject heading.

11. So with these efficiencies what do the staff do?

In the past they arrived at work from 7.45 - 8am onwards. They would start selecting from the newspapers and they would finish this task by about 10am. Then they would start indexing those clippings. Even when the newspaper clippings were outsourced and prepared for us as pdfs, they still had to be selected and any problems with the provided titles or journalists' names had to be corrected. As this selection was taking place the clips would start dribbling into Parlinfo and hence become available in Parlinfo at a trickle.

Now when the staff arrive the newspaper clippings have automatically been selected by LAST and are already in Parlinfo and accessible by our clients so the staff's first task of the day is more of a quality check. They review the selection classifier's decisions to make sure they fit our selection guidelines. They go through the newspapers and compare what the classifier has selected and what is in the

newspaper. If any classifier selected clips don't fit our guidelines they will remove them and they will add any that do fit the guidelines but the classifier has not selected. They correct badly transcribed titles and add or correct journalists' names when incorrectly spelled or missing. This review process is usually finished by 9am. Once this is done they start checking the classifier assigned subject headings.

12. What are the advantages and disadvantages of the system:

- a) The newspaper clippings are available to our clients by 7 am. Most of them will have subject headings and can be searched in Parlinfo by title or source, etc., as well as by subject.
- b) The selection classifier does a good job. There are always some articles that have been missed and some that may be removed by the indexing team but this is minor and doesn't affect the main delivery of the clippings.
- c) The automatic indexing can be improved. One way we hope to achieve this is to give more emphasis to the indexing of the human indexers than the classifier's indexing. At the moment they are weighted the same but we would prefer it to learn more from our indexers than itself. We have already started discussing this with the software developers but to date the result isn't satisfactory so more development needs to happen before we see any improvement.

Thank you for listening to my presentation; Questions

Automatic indexing in the Commonwealth Parliamentary Library

Grisoula Giopoulos

Grisoula Giopoulos@aph.gov.au

<http://aph.gov.au>

PARLIAMENTARY
LIBRARY
INFORMATION ANALYSIS ADVICE



Parliament of Australia

Department of Parliamentary Services

Parlinfo and the indexed databases

<input type="checkbox"/> House of Representatives	<input type="checkbox"/> Senate
<input type="checkbox"/> Committees	<input type="checkbox"/> Bills and Legislation
	<input type="checkbox"/> Bills of the Current Parliament
	<input type="checkbox"/> Bills Before Parliament Q
	<input type="checkbox"/> Other Bills Q
	<input type="checkbox"/> Bills of Previous Parliaments Q
	<input checked="" type="checkbox"/> Bills Digests Q
	<input type="checkbox"/> Bills Lists Q
	<input type="checkbox"/> House Disallowable Instruments Lists Q
	<input type="checkbox"/> Senate Disallowable Instruments Lists Q
	<input type="checkbox"/> Tariff Proposals Q
	<input type="checkbox"/> Links
<input type="checkbox"/> Media	<input type="checkbox"/> Constitution
<input type="checkbox"/> Press Releases Q	<input type="checkbox"/> Australia Constitution Q
<input checked="" type="checkbox"/> Newspaper Clippings Q	<input type="checkbox"/> 1890s Federal Conventions Q
<input type="checkbox"/> Radio and TV Q	<input type="checkbox"/> Australia Act 1986 Q
<input type="checkbox"/> ParView (audio-visual records) Q	<input type="checkbox"/> Statute of Westminster Q
<input type="checkbox"/> Library	<input type="checkbox"/> Publications
<input checked="" type="checkbox"/> Catalogue Q	<input type="checkbox"/> Senate Publications Q
<input checked="" type="checkbox"/> Articles Q	<input type="checkbox"/> House of Representatives Publications Q
<input checked="" type="checkbox"/> Political Party Documents Q	<input type="checkbox"/> Tabled Papers Register Q
<input type="checkbox"/> History of the Federal Capital and Parliament House Q	<input type="checkbox"/> Parliamentary Papers Series Q
	<input checked="" type="checkbox"/> Library Publications, Seminars and Lectures
	<input type="checkbox"/> Parliamentary Handbook

PARLIAMENTARY
LIBRARY

INFORMATION ANALYSIS ADVICE



Parliament of Australia

Department of Parliamentary Services

Newspaper titles supplied by iSentia

Advertiser
Age
Australian
Australian Financial Review
Canberra Times
Courier-mail
Daily Telegraph
Herald Sun
Mercury
Northern Territory News
Sydney Morning Herald
West Australian



What the indexers see in an “Interesting” article

ons	ID	Document Date	Source	Section	Page	Title	Author	Decision	Score	Major Subjects
	3712678	14/03/2015	Sydney Morning Herald	General News	1	Bali nine pair's new hope	ALLARD, Tom	Y	4346	Indonesia; Capital punishment; Drug trafficking; Australians overseas



Selection weighting showing positive and negative terms for “interesting” clipping

Feedback: Bali nine pair's new hope

Sydney Morning Herald - General News - p.1
Bali nine pair's new hope
ALLARD, Tom
3712678

Positive [4346]		Negative [0]	
government	13321	squad	1431
cabinet	7366	cool	1370
mps	4971	pair	747
the government	3443	saturday	724
of asylum	2757	the track	550
minister	2627	mates	227
policy	2327	wasnât	89
asylum seekers	2258		
government has	2214		
the policy	2114		
mr	1848		
asylum	1846		
governments	1696		
seekers	1444		
non government	668		
secretary	464		
attorney general	456		
tony	361		
spokesman	142		
foreign	43		



What the indexers see: Decision for “uninteresting” article and its score

Actions	ID	Document Date	Source	Section	Page	Title	Author	Decision	Score
	3713020	14/03/2015	Age	Business News	5	\$35,000 fridge becomes coolest status symbol	EVANS, Simon	N	-4537



Selection weighting of “Not interesting” article

Feedback: \$35,000 fridge becomes coolest status symbol

Age - Business News - p.5
\$35,000 fridge becomes coolest status symbol
EVANS, Simon
3713020

Positive [0]		Negative [4537]	
spending on	1402	fridge	15506
mr	320	appliances	7352
the billion	247	chefs	4625
		fi	3475
		saturday	3015
		winning	2948
		kitchen	2651
		the kitchen	2582
		cooking	2450
		crisp	2224
		fashion	2195
		hawthorn	2097
		screen	1938
		celebrity	1860
		luxury	1425
		style	1222
		s page	1173
		at and	1139
		frequency s	1120

Bali Nine 10/1/15


Citation

General

Subjects

Attachment

Publish

Document Info 

Jakarta's call: PM

Citation Id: 3602776
Document Date: 10/01/2015
Published Date: Fri Jan 16 15:30:06 2015
Source: Advertiser
Document Type: pressclp

Auto Suggestion:

Recommended

- [S] ABBOTT, Tony, MP
- [S] Indonesia
- [S] South Australia

Others

- [S] Visits from Australia
- [S] Diplomatic relations
- [S] Espionage

Recent:

[Add to MJ >](#)

- [S] Intergenerational reports
- [S] Retirement
- [S] Retirement housing
- [S] Aged
- [S] Retirement income

Major:

- [S] ABBOTT, Tony, MP
- [S] Australians overseas
- [S] Capital punishment
- [S] Drug trafficking
- [S] Indonesia



10/2/15 – starting to assign headings

The screenshot displays a web interface for a citation. On the left, a navigation menu includes 'Citation', 'General', 'Subjects', 'Attachment', and 'Publish'. The 'Subjects' menu item is highlighted. Below the menu, the document title is "'Godfather' tag hurt the pair's chances of a reprieve". Metadata includes Citation Id: 3649567, Document Date: 10/02/2015, Published Date: Tue Feb 10 08:23:48 2015, Source: Age, and Document Type: pressclp.

The main content area is titled 'Auto Suggestion:' and is divided into two sections: 'Recommended' and 'Others'. The 'Recommended' section lists: [S] Drug trafficking, [S] Australian Federal Police, [S] Capital punishment, and [S] Indonesia. The 'Others' section lists: [S] ABBOTT, Tony, MP, [S] Travel advisories, [S] Arrest, and [S] Thailand.

At the bottom, there are two columns: 'Recent:' and 'Major:'. The 'Recent:' column contains a button 'Add to MJ >' and a list of subjects: [S] Intergenerational reports, [S] Retirement, and [S] Retirement housing. The 'Major:' column contains a list of subjects: [S] Australian Federal Police, [S] Capital punishment, [S] Drug trafficking, and [S] Indonesia.



Bali Nine 13/3/15

Document Info

Legal hitch holds up Bali Nine duo appeal

Citation Id: 3710843
Document Date: 13/03/2015
Published Date: Fri Mar 13 11:05:43 2015
Source: West Australian
Document Type: pressclp

Auto Suggestion:

Recommended

- [S] Drug trafficking
- [S] Capital punishment
- [S] Indonesia
- [S] Australians overseas
- [S] ABBOTT, Tony, MP
- [S] Western Australia

Others

- [S] BISHOP, Julie, MP
- [S] WIDODO, Joko
- [S] YUDHOYONO, Susilo Bambang

Recent:

[Add to MJ >](#)

- [S] Intergenerational reports
- [S] Retirement
- [S] Retirement housing
- [S] Aged
- [S] Retirement income

Major:

- [S] ABBOTT, Tony, MP
- [S] Australians overseas
- [S] BISHOP, Julie, MP
- [S] Capital punishment
- [S] Drug trafficking
- [S] Indonesia



Change in weighting for 'Drug trafficking'

10/2/15 weight:2,188,543

13/3/15 weight: 14,167,665

Weight	Term
7910	chan
6742	sukumaran
5088	bali nine
3023	bali
1833	myuran
1678	chan and
1200	and sukumaran
791	myuran sukumaran
735	the bali
595	andrew chan

Weight	Term
34059	chan and
33940	chan
19125	bali nine
18378	sukumaran
17330	bali
10077	myuran
7608	andrew chan
6473	and myuran
6082	myuran sukumaran
4822	executions

