

Concept-based Mining to Enhance the Scope and Speed of Archival Qualitative Research

Andrew E. Smith

Key Centre for Human Factors and Applied Cognitive Psychology
The University of Queensland
Queensland, Australia, 4072.

Email address of corresponding author: asmith@humanfactors.uq.edu.au

Abstract. The development of a pilot system for archiving and retrieval of qualitative data in Australia is described. This system, called the Australian Qualitative Archive (AQuA), is not only intended as a valuable historical record, but as a key element in allowing qualitative research to achieve greater scale and impact through data sharing and secondary analysis. Limitations in existing retrieval techniques such as static classification and keyword search are discussed, and some solutions are suggested involving the Leximancer text analysis system. The functional architecture of the archive is described, including: ways to store the context and method descriptions alongside the data; ways to store coding and annotation analysis product as well as original data; ways for summary data to be retrieved when access to raw data is restricted; ways for qualitative data to be browsed and analysed on-line using emergent classification systems, or concept maps, extracted using Leximancer.

Introduction - the Problem

The employment of qualitative archival data for research is not new. Content Analysts have been performing longitudinal studies for many years, for example (Pratt, 2005). Indeed, it is one of the strengths of qualitative data that its message and style is still able to be interpreted to some degree after the passage of time. In a sense, it is part of the function of a document to be a self-explanatory archival record. Of course much context can be lost, particularly from fragments of discourse, but the situation can be much worse for quantitative and closed form data. Tables of numbers or Likert scale values are largely meaningless in the absence of explanatory text, or some shared convention which is still well understood by the researcher.

Archival stores of primary qualitative data for Social Science research, such as interview and survey data, are currently being established in several countries, for example ESDS

Qualidata in the UK, or the ARC funded AQUA project in Australia. In addition, there are of course large archival holdings of secondary and other qualitative data, such as newspaper articles, parliamentary transcripts, court judgements, etc. There are also holdings of image, video, and audio qualitative data. However, this paper will assume that textual metadata, captions, and transcripts must be available for these other media. Semantic retrieval directly from image or acoustic data is still a difficult research problem.

So the question arises: How can the researcher quickly locate all and only the archived textual resources which are relevant to their question? A second question, which is important but not so obvious, is: Can the researcher know in advance how to find everything that is relevant?

Strategies to Alleviate the Problem

At first inspection, it might seem that the retrieval of pertinent textual data by the researcher simply requires a good search engine. However, search engines have well-known problems. The main issue is the changing and diverse vocabulary used to express any idea. Allied to this is the context-dependent interpretation of meaning for most specific words. The recent work by (Pratt, 2005) is a good example of this. Her study examined the changing semantics around the word reconciliation over nine years of Australian parliamentary debate. These two factors result in (a) the searcher not being aware of all the relevant words to search for, and (b) some retrieved texts which contain a target term being nevertheless irrelevant.

Another issue which can pose problems is the granularity of the retrieval. A long interview or text based survey may cover many topics, and there may be one or two responses which were not the main subject of the document, but these responses may be very relevant to a later researcher. Normal document retrieval systems may not rank this document as being very relevant in its entirety, because only a couple of fragments qualify.

A common approach to these problems is to manually classify each document at deposition time. This almost always assumes a fixed taxonomy (aka thesaurus, or classification system, or ontology), which divides the subject domain of the archive into a finite number of categories. However, this approach has serious drawbacks. Human classification decision making is error-prone and uncertain on psychological grounds (Nisbett & Wilson, 1977), and this is certainly acknowledged in the Content Analysis community (Krippendorff, 2004), where much effort is expended in normalising human coding given a particular research question. Even given consistent classification within a certain community with a certain question, it is certain that researchers at later times and with different questions in mind would classify the material differently. In very simplistic terms, the answer to whether an orange is the same as a lemon depends on why you are asking. Psychological experimentation has established several important ways in which the framing of a decision can manipulate the choice made (Tversky & Kahneman, 1981). Library classification schemes show these deficiencies — most have been expanded in rather uneven ways to accommodate rich new fields such as IT, and most subject term hierarchies are cross linked extensively to try to support overlapping and interdisciplinary items and user communities. In on-line library catalogues as well as web classification systems such as

Yahoo, a great many users resort to the search engines rather than the classification systems. The W3C consortium has invested much time, effort, and strategic commitment to the Semantic Web, which is a single hierarchical classification system for everything, and which relies on human classification of documents on the web. This author sees limited merit in such an enterprise, except in highly controlled domains, for the reasons mentioned above: inconsistent human classification performance, dependence of category selection on situation, and the changing nature of knowledge.

So what options remain? Keyword searching is not bad, so long as the search process is iterated and refined by the researcher as they learn a more accurate query for their information need. The main hazard is that a significant unknown collection of data exists with few explicit links to the known information set. What is wanted may be a system which allows the user to construct an ad hoc and incomplete definition of their research topic, and which then proceeds to find other documents and vocabulary that match the idea, rather than the literal key words. These can be called ‘indirect’ retrievals. To locate relevant fragments contained in less relevant documents, the system should be able to index and retrieve single responses from the data sets.

Quoting heavily from an unpublished paper (Rowse & Holloway, 2003) on historical use of archived data:

Following indirect pathways to data can allow the historian to collect data over more points of time, making any analysis and conclusions more robust (though the researcher would have to comment on the significance of any differences in the wording of questions in different studies). While the historian who follows the direct path through the archive would find data covering a little over twenty years the historian following indirect paths could have data covering a 100 year period.

The most adventurous etc. historical researcher would be willing to put in many hours trawling the indirect studies, believing that that is where you find the nuggets of data that lift one’s study above the average, challenge the paradigm, etc. However, accessing the indirect corpus is relatively difficult because it requires a variable-by-variable inspection of the corpus of indirect studies. Is there a way that this could be automated?

A Proposed Solution

As part of the ARC funded AQuA project (Australian Qualitative Archive), a concept mining system is being implemented to help address some of these issues. The concept mining system being employed is called Leximancer¹. This is a relatively new system which employs machine learning algorithms from Corpus Linguistics to generalise from an incomplete concept definition (called a concept seed set, which essentially looks like a keyword set), into a broader definition of the concept, and which collects other synonyms, component terms, and strongly diagnostic terms. This broader classifier is learned automatically from the text corpus under consideration, and is used to retrieve indirectly related text fragments. The Leximancer system has been quite rigorously evaluated using validity criteria from the Content Analysis discipline (Smith & Humphreys, To appear). That paper also describes the operation of the system in some detail. In consequence, only a short description of the technique will be repeated here. Leximancer has also been used as a tool to support qualitative analysis in several research articles, for example (Davies, Green, Rosemann, & Gallo, 2005), (Rooney, To appear), (Scott & Smith, 2005), (Watson, Smith, & Watter, 2005), (Varre, Ellaway, & Dewhurst, 2005).

¹ <http://www.leximancer.com>

Summary of Leximancer Technique

The essential features are as follows: a unified body of text is examined to select a ranked list of important lexical terms based on word frequency and co-occurrence usage. These terms then seed a bootstrapping thesaurus builder which learns a set of classifiers from the text by iteratively extending the seed word definitions. The resulting weighted term classifiers are then referred to as *concepts*. Next, the text is classified using these concepts at a high resolution, which is normally every three sentences. This produces a concept index into the text and a concept co-occurrence matrix. By calculating the relative co-occurrence frequencies of the concepts, an asymmetric co-occurrence matrix is obtained. This matrix is used to produce a two-dimensional concept map via a novel emergent clustering algorithm. The connectedness of each concept in this semantic network is employed to generate a third hierarchical dimension which displays the more general parent concepts at the higher levels.

Leximancer concept seed sets can be specified as required by the researcher as primary topics of interest. The system can take such a primary topic, and can extract a hierarchical concept network of related and dependent concepts. This process is called *profiling*. This allows any researcher to dynamically reclassify a set of text data in a manner which is focused on their current interest. It is planned that the AQuA archive will allow researchers to profile a topic of their choosing across multiple data sets to retrieve indirectly related records and to identify aspects of the topic which may not have occurred to them. This can also help to identify data contamination, in which case the researcher can refine their concept seed set, and possibly exclude certain data sets, for example (Scott & Smith, 2005).

Leximancer can also generate a concept network which characterises a given set of text data in a completely automatic and unsupervised manner. It is planned that concept maps of each deposited data set, and of important collections of multiple data sets, will be generated at deposition time for researchers interested in trends across, and content of, the particular sets.

Several graphical examples of actual Leximancer text analyses can be seen at the Leximancer site and also in (Smith & Humphreys, To appear).

Architecture of AQuA

The pilot of the Australian Qualitative Archive (AQuA) functions as follows:

The Gateway

The researcher goes to the AQuA front page. Here they will find an interface to a database search engine. The database contains metadata describing the deposited *data bundles*. The database will also enforce user authentication and access control. The archive user will search the database by the usual metadata, and also by topical concepts which have been automatically discovered from the archive and assigned to each data bundle by the Leximancer engine. Having found one or more interesting data bundles, the researcher has two or three options. They can download, browse, or perform on-line analysis.

The Data Bundle

A data bundle is a set of qualitative **data** files plus supporting documents describing **context** and **method**. Note that a study may be made up of several data bundles: (1) the original data and supporting documents describing context and method for the data collection; (2) one or more derived data bundles containing coding/annotation data derived from the original data plus supporting documents describing the re-contextualisation and method that led from the original data to this derived data bundle. The coding/annotation data will likely be generated by a CAQDAS² system. The archiving of derived data bundles should promote data sharing and secondary analysis where this is appropriate. The author is involved in a separate project, also funded by the Australian Research Council (grant SR0567263), which is developing an open standard and a library for storage and interchange of qualitative analysis product between CAQDAS systems and archives. The draft standard is called QDIF, the Qualitative Data Interchange Format³. It should be noted that though the standard should allow export of most of the important information from most CAQDAS systems, not all projects will necessarily be compatible with all systems without some loss of information. For example, one CAQDAS system may allow coding of text segments shorter than a sentence, and another may not.

The Metadata

The metadata for each data bundle is assigned at deposition time. It will be stored using the DDI metadata schema (Data Description Initiative⁴). Some metadata, such as title, author, and method, will be assigned by the archivist in consultation with the researchers. Topical metadata will be assigned by classifying the textual content of the data bundle using a Leximancer thesaurus. The archivist will from time to time use Leximancer to extract a topical thesaurus from the whole archive, and also from designated partitions of the archived data. The thesaurus can either be extracted unsupervised, or additional controlled concepts can be added into the concept seed set. This thesaurus will then be used to classify the text content of deposited data sets. The classification will be entered into the metadata as a *concept co-occurrence matrix*. A matrix has a significant advantage over a simple list of concepts in that the matrix stores the strength of relationships between concepts within the data bundle, thus allowing for more precise searching.

For example, if someone deposited the set of publicly released Enron e-mail⁵, the data bundle might be classified with the concept co-occurrence matrix shown in Figure 1.

The Leximancer Data bundle

Most original data bundles which contain text will have a Leximancer map attached as a derived data bundle when they are first deposited. This will be an unsupervised topical and relational analysis of the deposited data and supporting documents. Should an archive

² Computer Assisted Qualitative Data Analysis Software

³ <http://delacruz.csse.unimelb.edu.au/portcode/>

⁴ <http://www.icpsr.umich.edu/DDI/>

⁵ <http://www.cs.cmu.edu/enron/>

Entity	Enron	power	California	market	should	time	energy	electricity	information
Enron	221719	11111	7189	9973	10820	11981	13826	7086	5541
power	11111	96591	21900	13661	4250	4684	10199	19440	1208
California	7189	21900	81953	11166	3700	3985	11560	16620	900
market	9973	13661	11166	76968	5298	5588	5970	9833	1943
should	10820	4250	3700	5298	143538	9293	2144	2380	6169
time	11981	4684	3985	5588	9293	144796	2679	2887	2632
energy	13826	10199	11560	5970	2144	2679	54623	6681	1407
electricity	7086	19440	16620	9833	2380	2887	6681	62612	614
information	5541	1208	900	1943	6169	2632	1407	614	65587
gas	6052	6554	3963	4561	2317	2069	3699	5365	500
state	4454	21995	14045	4829	2345	2361	7267	12116	541
message	5435	51	80	40	2873	425	59	9	10597
e-mail	6683	115	95	172	899	928	143	75	5667
natural	3593	4192	2485	2798	1159	1158	2681	3385	276
price	3906	6169	5998	7297	2558	2504	2998	4860	669
business	8706	1718	1011	2131	1866	1993	2369	1380	2484
year	6081	2863	2841	2189	1440	2159	2924	1962	470
million	8371	2703	2241	1144	728	804	2531	1458	256
Houston	10574	1552	1121	1168	1850	5265	2636	1208	1121
review	5969	194	110	219	902	745	135	72	3430

Figure 1. Unsupervised Leximancer Classification Matrix of Enron E-mail Set

user wish to browse an original data bundle to assess how much relevant material is contained within, they can simply load the associated Leximancer map into the on-line Leximancer engine.

When the Enron data bundle is added to the archive, the archivist would add an unsupervised Leximancer map of the e-mail collection as derived data bundle. The concept map part of this is shown in Figure 2.

The Leximancer map also allows the user to drill down by concept or relationship to browse relevant text segments.

Access Control

Multiple access control levels will allow depositors to control whether archive users can see the full data of a data bundle, or just the supporting documents, or neither. A Leximancer map bundle allows the user to see a summary map of the data, plus cross tabulations of discovered concepts, without necessarily allowing the user to drill down to the text data itself.

On-line Analysis

Data bundles which contain text can be analysed further on-line using Leximancer. The researcher can construct a new exploratory map of original data, or load an existing secondary analysis to perform further work. Leximancer will allow the researcher to seed concepts of interest, and perform profiled discovery around these customised concepts.

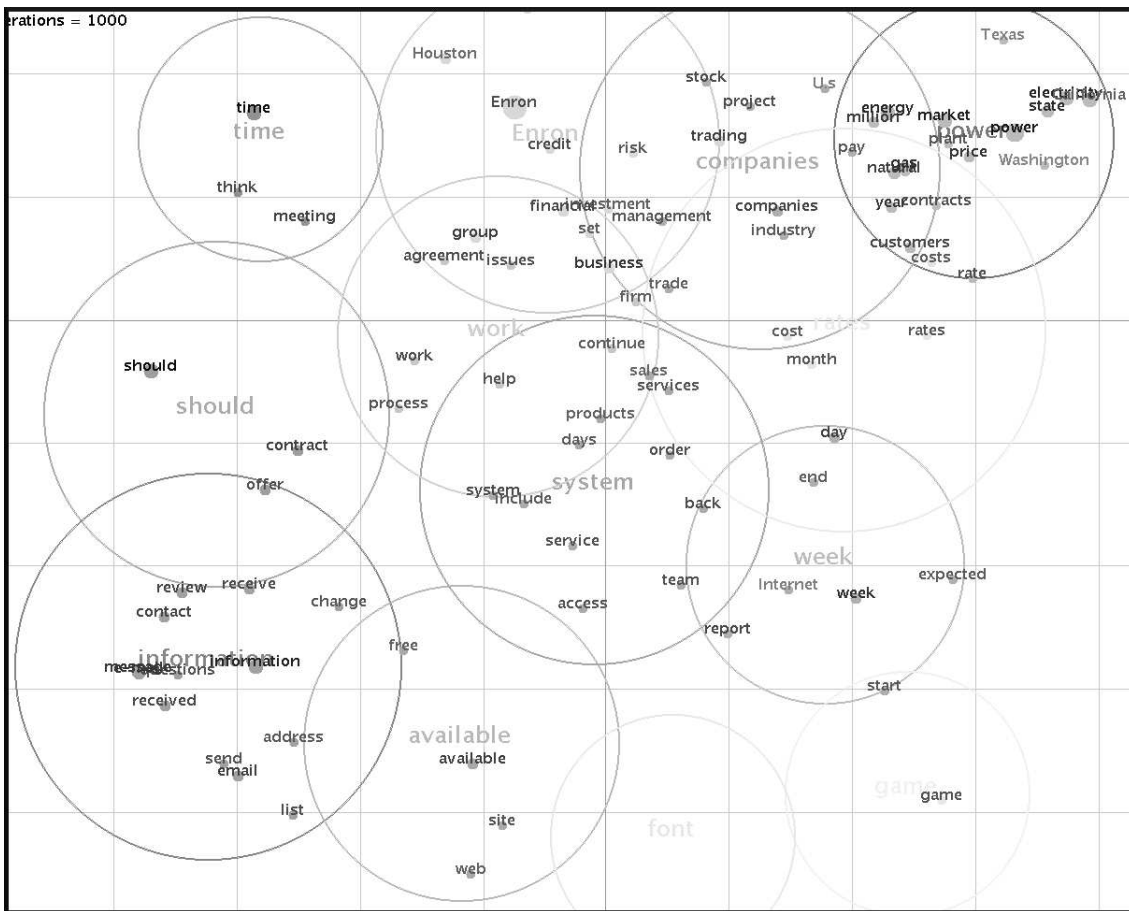


Figure 2. Unsupervised Leximancer Map of Enron E-mail Set

For example, say that a collection of Hansard speeches from the Australian Federal Parliament⁶ was stored in the archive. In this example data bundle, the speeches cover the debate on the prohibition of human cloning and stem-cell research, conducted from 20 August 2002 to 29 August 2002.

Figure 3 shows a Leximancer analysis of the text content mapped alongside metadata tags for the two main political parties involved in the debate — Labor and Liberal. The concepts in this map were extracted from the text in an unsupervised manner, while the political party tags are assigned based on the affiliation of the nominal speaker for each Hansard speech file.

Next, the researcher might analyse the very same data in a different way. If their interest was in the framing of illness and disease in the debate, the researcher would seed the concept *illnesses* with just three strong terms — ‘illnesses’, ‘disease’, and ‘diseases’. The system presents a list of frequent names and words from the data to make this process easier. The researcher then asks Leximancer to learn a broader concept around *illness* from the text data. Figure 4 shows the most relevant terms from the resulting thesaurus. The numbers indicate relevancy to the concept and the double square brackets denote proper names.

⁶ <http://www.aph.gov.au/hansard/index.htm>

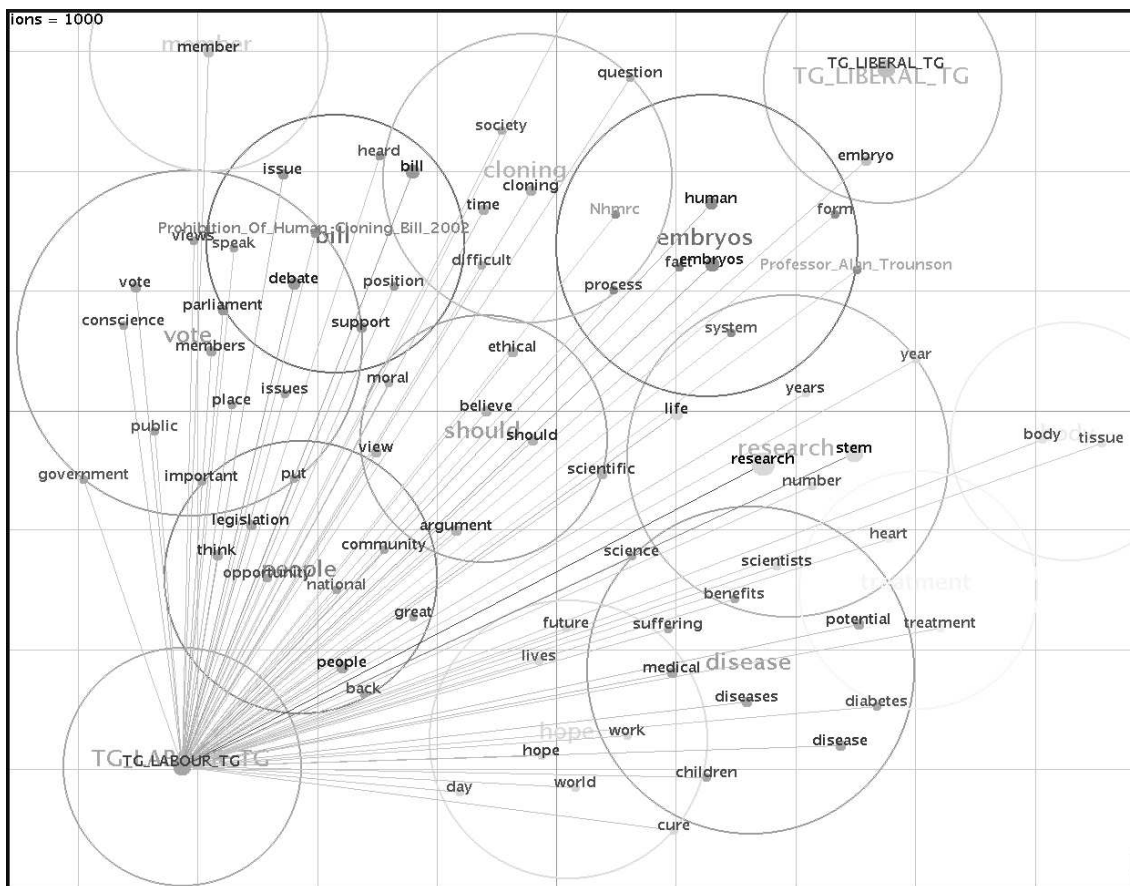


Figure 3. Map of Stem Cell Debate from the Australian Parliament

The researcher can also ask the system to *Profile* the concept. This process extracts and learns additional concepts from the text which are relevant to the target concepts, in this case *illnesses*. Figure 5 shows the resulting map, showing the themes which frame this concept in the data set.

Future Plans

The next stage of this project will scale up the pilot archive into an operational system. There are four main aspects to this:

- Upgrade hardware and place under appropriate housing and administration.
- Assess and enhance usability of the interface in consultation with community.
- Check, revise, and document metadata schema after surveying available qualitative data sets.
- Establish and document data deposition, access control, and management procedures after extensive consultation with existing qualitative archives, particularly ESDS Qualidata, and the Australian qualitative research community.

It is freely admitted that this project is mainly focused on textual qualitative data at present. However, it is certainly envisaged that data bundles which include other media

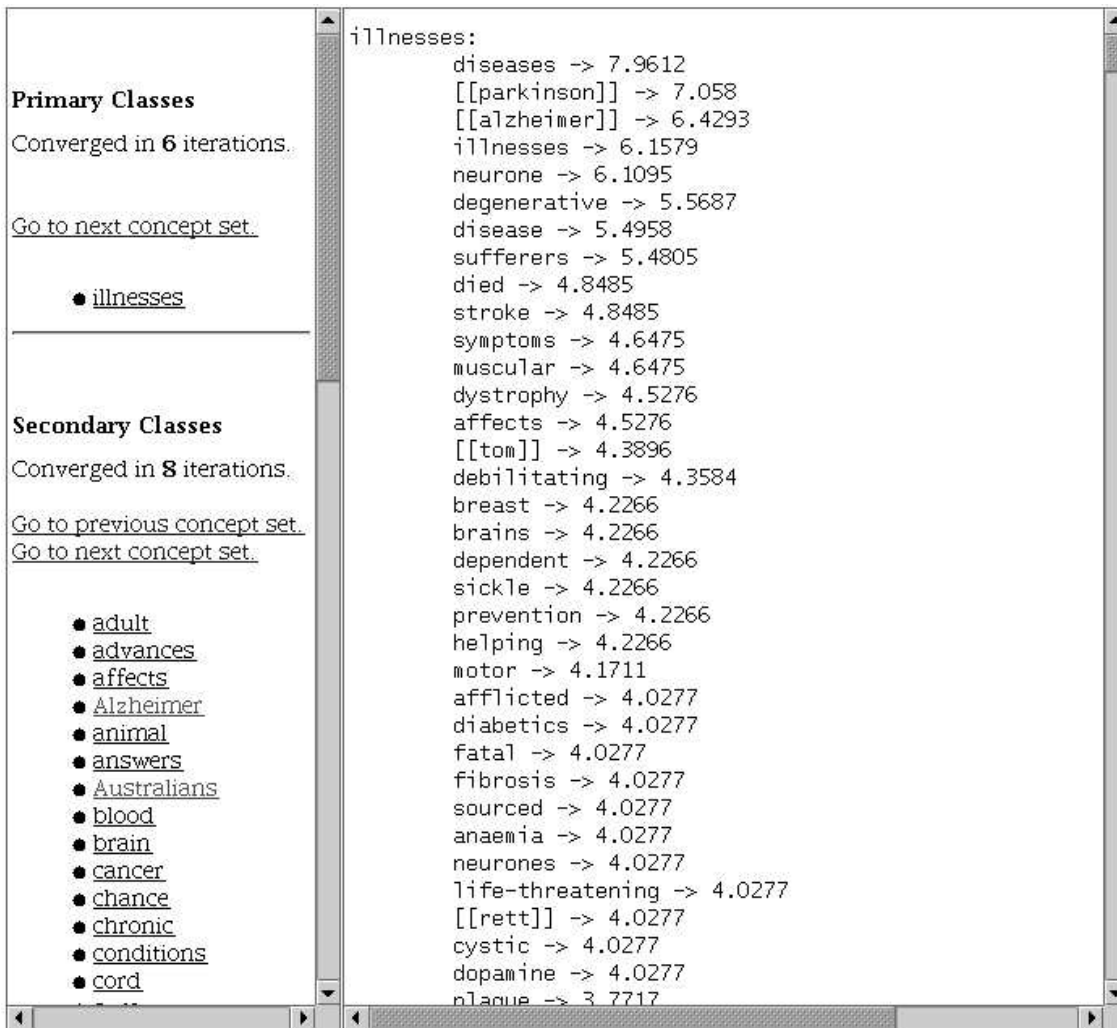


Figure 4. Learned Thesaurus Entry for concept *illnesses* (abridged)

types will be stored, along with supporting documents and metadata. The main limitation is that Leximancer analysis and classification will only operate on the supporting documents, and not on the binary data objects.

Acknowledgements

This research is supported under the Australian Research Council LIEF granting program, grant number LE0560677: Australian Social Science Data Archive: Facility Enhancement & Network Development.

References

- Davies, I., Green, P., Rosemann, M., & Gallo, S. (2005). Conceptual modelling - what and why in current practice. *Lecture Notes in Computer Science*, 3288, 30–42.
- Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (second ed.). Sage Publications.
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know. verbal reports on mental processes. *Psychological Review*, 84(3), 231–259.
- Pratt, A. (2005). *Practising reconciliation? the politics of reconciliation in the Australian parliament, 1991-2000*. Canberra, Australia: Commonwealth of Australia. (ISBN 0-9752015-2-2)
- Rooney, D. (To appear). Knowledge, economy, technology and society: The politics of discourse. *Telematics and Informatics*.
- Rowse, T., & Holloway, S. (2003). *How to interest historians in the assda*. Unpublished.
- Scott, N., & Smith, A. E. (2005). Use of automated content analysis techniques for event image assessment. *Tourism Recreation Research*, 30(2), 87–91.
- Smith, A. E., & Humphreys, M. S. (To appear). Evaluation of unsupervised semantic mapping of natural language with leximancer concept mapping. *Behavior Research Methods*. (Accepted for publication 29 March 2005)
- Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481), 453–458. (New Series)
- Varre, C. de la, Ellaway, R., & Dewhurst, D. (2005). An analysis of the large-scale use of online discussion in an undergraduate medical course. In J. Cook & D. Whitelock (Eds.), *Exploring the frontiers of e-learning: borders, outposts and migration; alt-c 2005 12th international conference research proceedings*. ALT Oxford.
- Watson, M., Smith, A. E., & Watter, S. (2005). Leximancer concept mapping of patient case studies. *Lecture Notes in Computer Science*, 3683, 1232-1238.