[3] Richard Cole, Peter Eklund and Bernd Groh. Dealing with large contexts in formal concept analysis. In *Second International Symposium on Knowledge Retrieval, Use and Storage for Efficiency,* pages 151-164, Vancouver, B.C., Canada, August 1997.

[4] Richard Cole, Peter Eklund and Don Walker. Using conceptual scaling in formal concept analysis for knowledge and data discovery in medical texts. In *Second Pacific Asian Conference on Knowledge Discovery and Data Mining,* 1998.

[5] Christiane Fellbaum (editor). *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[6] Bernhard Ganter and Rudolf Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer-Verlag, 1999.

[7] Stephen Green. *Automatically generating hypertext by computing semantic similarity*. University of Toronto, Canada, 1997.

[8] Stephen Green. Automated link generation: Can we do better than term repitition? In *Proceedings of the Seventh International World Wide Web Conference,* pages 75-84, Brisbane, Australia, April 1998.

[9] Jane Morris and Graeme Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics,* Volume 17, Number 1, pages 21-48, 1991.

[10] Rudolf Wille. Landscapes of knowledge: A pragmatic paradigm for knowledge processing. In *Second International Symposium on Knowledge Retrieval, Use and Storage for Efficiency,* pages 2-13, Vancouver, B.C., Canada, August 1997.

wondershare™
PDF Editor

TML: A Thesaural Markup Language

Maria Lee
Mathematical & Information Sciences
CSIRO
Locked Bag 17, North Ryde 1670
Australia
Maria.Lee@cmis.csiro.au

Stewart Baillie
Mathematical & Information Sciences
CSIRO
723 Swanston Street, Carlton 3053
Australia
Stewart.Baillie@cmis.csiro.au

Jon Dell 'Oro
Mathematical & Information Sciences
CSIRO
723 Swanston Street, Carlton 3053
Australia
Jon.Delloro@cmis.csiro.au

1 Abstract

Thesauri are used to provide controlled vocabularies for resource classification. Their use can greatly assist document discovery because thesauri man date a consistent shared terminology for describing documents. A particular thesaurus classifies documents according to an information community's needs. As a result, there are many different thesaural ***schemas.*** *This has led to a proliferation of schema-specific thesaural systems. In our research, we exploit schematic regularities to design a generic thesaural ontology and specify it as a markup language. The language provides a common representational framework in which to encode the idiosyncrasies of specific thesauri. This approach has several advantages: it offers consistent syntax and semantics in which to express thesauri; it allows general purpose thesaural applications to leverage many thesauri; and it supports a single thesaural user interface by which information communities can consistently organise, store and retrieve electronic documents.*

Keywords: Electronic Documents, Metadata, Ontology, Thesaurus, XML

2 Introduction

Many problems common to electronic document systems are often not new, but well-known problems occurring in a new medium. In a search for solutions to problems in the electronic medium, we can often learn from the experience of traditional media. This is true for resource discovery in large electronic information repositories. The solutions offered by search engines have evolved rapidly to fill a need for resource discovery in the electronic storage medium. But, in a managed information environment, their free text search approach can be a poor substitute to thesaurally organised metadata approaches.

The use of metadata search can complement and enrich the text matching approach of search engines. Metadata is data which describes data. It provides a conceptual description of a resource's content, context, and function. The keywords list at the head of this paper is an example of **metadata—** it describes something about the document content. In document management systems, metadata is often used to index a document by describing what it is about and its catalogue detail. However, this metadata content, when used for search, can run into a similar problem to that of document content: it lacks a consistent shared vocabulary. A traditional solution to this problem is to use a thesaurus to control metadata content.

In the terminology of the record keeping community, a thesaurus is a fixed vocabulary of approved and unapproved terms, their functions and meanings, and their inter-term relationships. A thesaurus can provide *accuracy* of description through explicit classification by approved terms; *consistency* through controlled terminologies; and *efficiency* in retrieval through the use of the right terminologies [Lancaster 1972].

A thesaurus is valuable if its vocabulary acts as a *lingua franca* that reflects the culture of a user community and purposes the information repository schema. This often means that a different thesaurus is necessary for each user community. The result, in the electronic storage medium to date, has been many incompatible thesaural applications each one designed about its particular thesaurus. In our research we have sought a generic ontology in which to represent the idiosyncrasies of these many specific thesauri. This would allow a single application to work with many different thesauri. In this paper, we describe this ontology, a markup language used to express it, and introduce a general purpose Thesaural Explorer application based upon them.

3 Generic Thesaural Ontology

An ontology, in computer science, has come to denote an explicitly specified conceptualisation of part of the world. In software, an ontology is implemented as a data structure. What distinguishes the ontology from the data structure is semantics: that it talks about something in the world. An ontology provides users with a representation which is essential to effective communication and coordination.

Proceedings of the 4th Australasian Document Computing Symposium,
Coffs Harbour, Australia,
December 3, 1999.

Our goal was to design a Generic Thesaural Ontology (GTO) capable of representing many different thesauri. This would allow us to express a specific thesaurus in a common language. The way we approached this goal was to review six major existing thesauri and model their classes and relations at a higher level of abstraction. The six thesauri selected were:

- Keyword AAA Australian Government Thesaurus [Keyword AAA],
- Getty Art and Architecture Thesaurus [AAT],
- Getty Thesaurus of Geographic Names [TGN],
- Library of Congress Subject Headings [LCSH],
- OCLC Dewey Decimal Classification [OCLC],
- Medical Subject Headings [MeSH].

These thesauri were selected because they are well-known, used by different communities in different domains, represented both function-based and subject-based classification schema, and are based on the monolingual thesaural standard [ISO 2788]. In the draft of the GTO presented here, it was not our goal to represent multilingual thesauri or to map inter-thesaural links.

We tried to keep the thesaural ontology as simple as possible. In this case, we found a taxonomic graph sufficient to our purposes. The classes and relations of the GTO are shown in Figure 1.

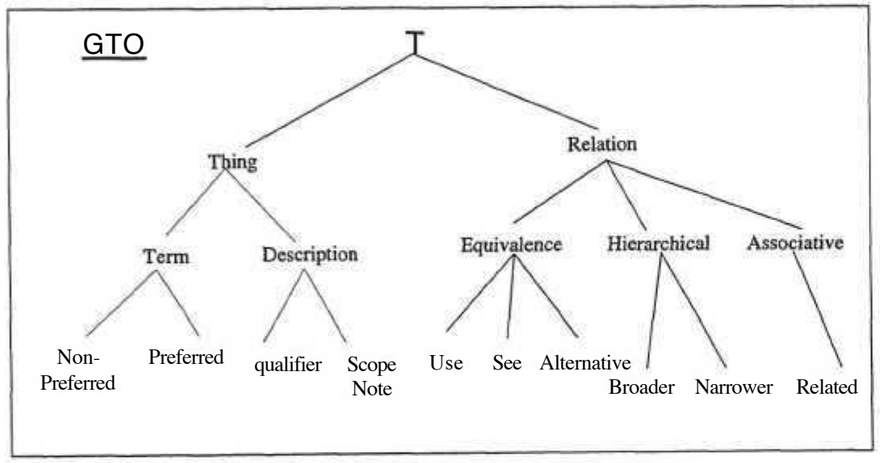

Figure 1. Graph of the Generic Thesaurus Ontology.

The GTO graph illustrates inheritance, where each class on the lower level inherits properties from the preceding level. The root symbol T is a neutral representation for the universal entity. T bifurcates to superordinate the GTO classes of *Thing* and *Relation:*

Thing:
is defined by a monadic predicate p(x) in terms of the form of the entity x (including its inherent parts and properties) and not in terms of anything external to x.

Relation:
is defined by a dyadic predicate (x,y) that relates the entity x to some independent y that is not an inherent part or property of x.

The class Thing has two subtypes:

Term:
represents any word or phrase used to represent a concept. Thesaural terms are divided into preferred terms and non-preferred terms. Preferred terms are authorised terms and the only ones valid for use in resource description. Non-preferred terms (synonyms, spelling variants, inverted form, etc) are designated by a USE relation which links them to the preferred term.

Description:
describes the meaning of a concept. It includes Scope Note description and parenthetical qualifier. A Scope Note is a statement that clarifies the meaning and usage of a descriptor or guide term within the thesaurus. The parenthetical qualifier is used to qualify or specify the context of an entry and so remove ambiguity. It allows users to distinguish among the homographs at a glance, while their scope notes further define them.

The class Relation consists of three subtypes:

Equivalence:
the equivalence relationship exists between or among terms that represents the same concept. Equivalent terms may be synonyms, variant spellings, inverted forms of multi-word terms, etc. Equivalent terms falls into three categories: Use terms, See and Alternative.

- *Use:* When a concept can be expressed by more than one term or more than one spelling, one of the terms is selected as the preferred term, and the other included as non-preferred or use-for terms. In all cases the two terms involved (referred from and referred to) are essentially equivalent. In many thesauri the use reference is also employed to effect one-to-many mapping.
- *See:* Although *use* reference is the usual thesaurus convention for directing from a term that cannot be used in indexing and searching to a term that can be used, some vocabularies prefer *see* to serve the same function.
- *Alternative:* different grammatical forms of the descriptor. Generally they are to allow for variety of indexing practices, such as use of singular instead of plural, and to provide a combination form for use in constructing headings from more than one descriptor.

Hierarchical:
is the most fundamental thesaural relationship, the basic type of links establishing a term's membership in the thesaurus. The relationship generally is restricted to the formal genus-species relation. If a term is a type of, kind of, example of, or manifestation of another term, then a genus-species relation exists. Within the context of the genus-species relationship, the genus or class is called the Broader Term and the species or member is called the Narrower Term. The broader-narrower relations are reciprocals of one another.

Associative:
relates terms that are not hierarchical (broader-narrower) nor equivalent (use) but in some other way linked. Usually they link between terms that belong to different categories, with no siblings; these provide the basis for the most common types of related terms and are the most difficult links to define rationally. Generally speaking the functions of the related terms in thesauri are to clarify the scope of and to define the main term, and to alert the indexer or searcher to other terms or concepts of interest.

4 Thesaurus Markup Language

The general thesaural ontology gives us a conceptual representation of thesauri. A thesaural markup language (TML) manifests this as a grammar in which to express the content and structure of specific thesauri. TML is specified as an XML schema which defines the permitted markup element types and embedding structure. The TML syntax consists of the element names and structure shown in Figure 2.

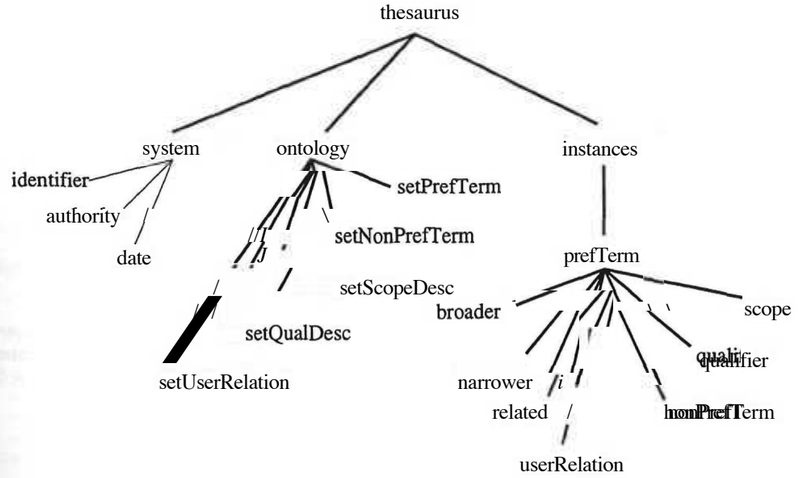

Figure 2. TML Element Graph

The TML graph structure is not isomorphic with the GTO graph structure; ie. it does not classify according to the thesaural ontology, but reorganises the semantic classes and relations of the GTO into a process-oriented data structure which reflects how the data are used. The TML element *thesaurus* subordinates three types of TML elements: system, ontology, and instance. The *system* element represents metadata about the thesaurus. The *ontology* element represents a particular thesaurus' structure (its idiosyncrasies); it extends the generic GTO taxonomy down to a specific thesaurus instance. The *instances* element represents the content needed to populate a thesaurus. Its prefTerm sub-element represents a preferred thesaural term. The prefTerm element is the lynchpin of the TML instance structure.

wondershare™
PDF Editor

The following tables give more detail on the TML elements. In the following tables, the occurrence column indicate the existential status of each element. The meanings of the occurrence symbols are:

Occurrence Symbol	Meaning
1	Required, not repeatable
?	Optional, zero or one occurrence
+	Required, repeatable (one or more occurrence)
*	Optional, repeatable (zero or more occurrences)

The *system* element is composed of the following sub elements:

Name	Occurrence	Description
identifier	1	The name of the thesaurus
version	1	The version number of this thesaurus
language	?	The language used in this thesaurus
authority	?	Organisation authorisation
createdBy	?	The name of the person/organisation who created the record defining the term
approvedBy	?	The name of the person/organisation who approved the record defining the term
date	?	
createdDate	?	The date on which the record defining the term was created
modifiedDate	?	The date on which the record defining the term was last modified

The *ontology* element describes the thesaural GTO extensions (see Figures 4 & 5). It is composed of the following sub elements:

Name	Occurrence	Description
setPrefTerm	+ *	Register the type and name of a class of preferred terms
setNonPrefTerm	*	Register the type and name of a class of non preferred terms
setScopeDesc	*	Register the type and name of a class of scope notes
setQualDesc	*	Register the type and name of a class of qualifiers
SetUserRelation		A user defined relation

The *instances* element is composed of the following sub elements:

Name	Occurrence	Description
prefTerm	1	The instance of the preferred term
scope	7	The instance of the scope note
qualifier	?	The instance of the qualifier
broader	*	The instance of the broader term
narrower	*	The instance of the narrower term
related	*	The instance of the related term
nonPrefTerm	*	The instance of non-preferred term
userRelation	7	The instance of user defined relation

An convenient way to understand how TML works is to look at some worked examples. These are described below.

4.1 TML for Keyword AAA Thesaurus

Keyword AAA [Keyword AAA] is the thesaurus most extensively used by Australian Government agencies. It uses the relationships broader term, narrower term and related term. The broader and narrower term relations are reciprocal and the related and top relations are reflexive. Figure 3 illustrates some of these **Keyword** AAA terms and relations.

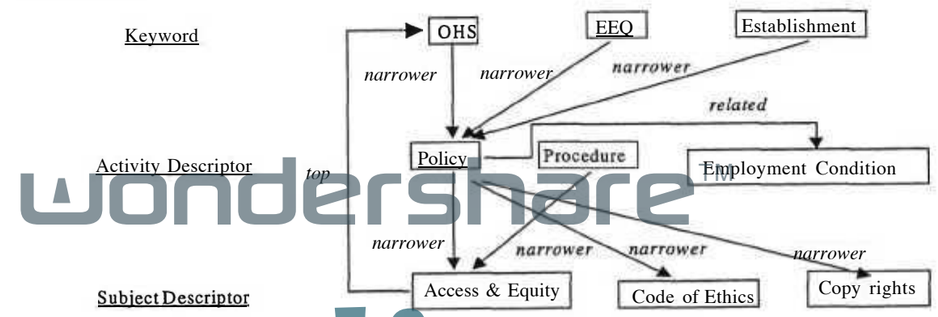

Figure 3. Keyword AAA Thesaurus Graph

The Keyword AAA thesaurus uses three types of preferred terms: Keywords, Activity Descriptor, and Subject Descriptor. Permitted Acronyms and Forbidden Terms are types of non-preferred terms. A Scope Note construct

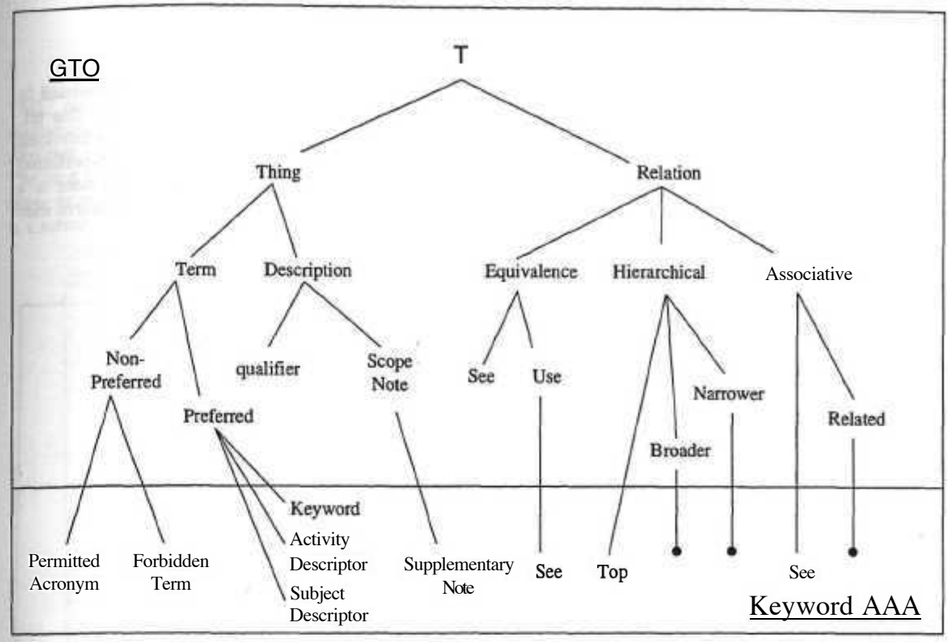

Figure 4. Keyword AAA GTO Graph

called Supplementary Note provides additional free text description. Keyword AAA employs standard hierarchical (broader & narrower) and associative (related) relations. However, the see relation in Keyword AAA is overloaded: one sense represents the equivalence use relation, and the other sense represents the associative relation. The associative relationship of the see reference is rather vague in the Keyword AAA. In some sense, it is similar to the related relation, but it may also include the near synonym and other conceptually close relationships.

Keyword AAA extends the GTO taxonomy as shown in Figure 4.

In what follows, we give examples of how **Keyword** AAA can be marked up in TML. The system markup element is used to record thesaural metadata.

```
<system>
<identifier version="CSIRO KeywordAAA 0.1" language="English"/>
      <authority createdBy="CMIS OMT Project"/>
      <date modifiedDate="981022"/>
</system>
```

The ontology markup element is used to define the **Keyword** AAA specific types extension to the general ontology (those that fall below the horizontal line in Figure 4).

```
<ontology>
      <setPrefTerm        type= "KW"    name="Keyword"/>
      <setPrefTerm        type= "AD"    name="Activity Descriptor"/>
      <setPrefTerm        type=" SD "   name="Subject Descriptor"/>
      <setNonPrefTerm     type="PA"     name="Permitted Acronym"/>
      <setNonPrefTerm     type= »FB"    name="Forbidden Term"/>
      <setScopeDesc       type= "SN"    name="Supplementary Note"/>
      <setUserRel         type="TOP"    name="Top"/>
</ontology>
```

The instance markup element is used to populate the ontology structure with instances:

```
<instances>
      <prefTerm           type="KW"     value="Occupational Health &amp; Safety">
            <narrower     type="AD"     value="Policy"/>
            <nonPrefTerm  type="PA"     value="OHS"/>
            <nonPrefTerm  type="FB"     value="OHS&amp;S/>
      </prefTerm>
      <prefTerm           type="AD"     value="Policy">
            <broader      type="KW"     value="Occupational Health &amp Safety"/>
            <broader      type="KW"     value="EEQ"/>
            <broader      type="KW"     value="Establishment"/>
            <narrower     type="SD"     value="Access &amp; Equity"/>
            <narrower     type="SD"     value="Code of Ethics"/>
            <narrower     type="SD"     value="Copyright"/>
            <related      type="AD"     value="Employment Condition"/'>
      </prefTerm>
</instances>
```


wondershare
PDF Editor

4.2 TML for Getty Art and Architecture Thesaurus

The Getty Art and Architecture Thesaurus [AAT] is another example of a thesaurus which can be represented in TML. AAT is a controlled vocabulary for describing and accessing cultural heritage information, e.g. fine art, architecture, decorative art, and material culture. It is structured by facets (categories), which are further subdivided into sub-facets or hierarchies. The thesaurus uses a preferred term to represent a single concept, while non-preferred terms (synonyms, spelling variants, inverted forms) are designated using a use relation and a scope note is a statement that clarifies the meaning and usage of a descriptor or guide term within the thesaurus. The thesaurus uses hierarchical relationships and equivalence relationships.

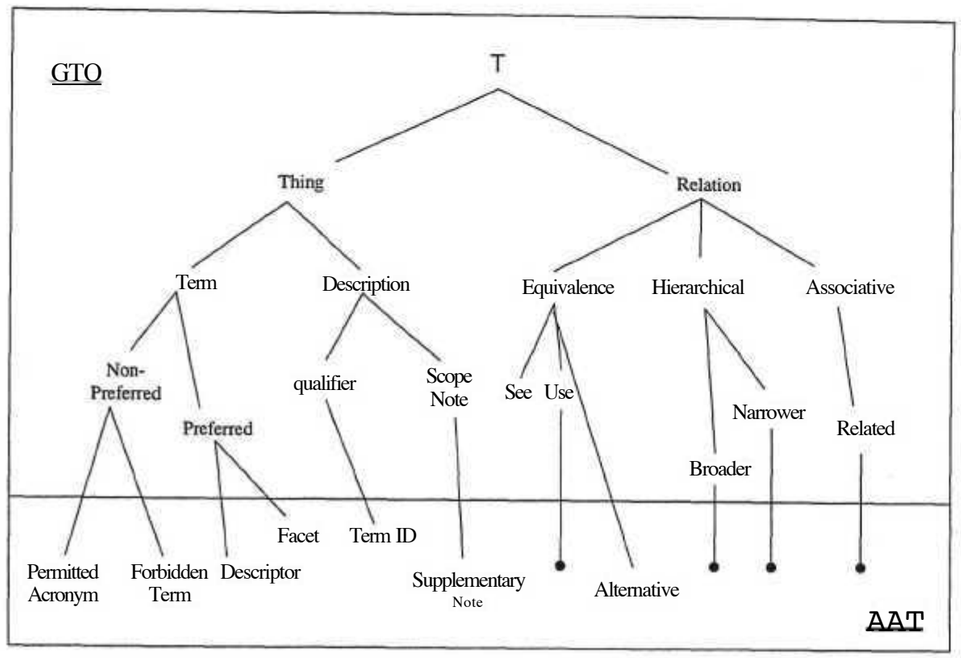

Figure 5. Getty AAT GTO Graph

AAT extends the GTO taxonomy as shown in Figure 5.
AAT can be marked up in TML using an ontology element such as:

```
<ontology>
      <setPrefTerm          type="FC"    name="Facet"/>
      <setPrefTerm          type="DS"    name="Descriptor"/>
      <setNonPrefTerm       type="PA"    name="Preferred Term"/>
      <setNonPrefTerm       type="FB"    name="Forbidden Term"/>
      <setQualDesc          type="ID"    name="Term ID"/>
      <setScopeDesc         type="SN"    name="Supplementary Note"/>
      <setUserRelation      type="ALT"   name="Alternative"/>
</ontology>
```

The ontology structure can be populated as:

```
<instances>
      <prefTerm             type="FC"    value="Associate Concepts">
            <narrower       type="DS"    value="concepts in the arts"/>
            <narrower       type="DS"    value="culture and related concepts"/>
            <narrower       type="DS"    value="environmental concepts"/>
            <qualifier      type="ID">4 7309</qualifier>
      </prefTerm>
      <prefTerm             type="DS"    value="concept in the arts">
            <broader        type="FC"    value="Associated Concepts"/>
            <narrower       type="DS"    value="artistic concepts"/>
            <narrower       type="DS"    value="genres in the arts"/>
            <qualifier      type="ID">56107</qualifier>
      </prefTerm>
</instances>
```

Wondershare™
PDF Editor

5 Application

TML provides a way to represent task-domain specific thesauri and make them available to a document management systems. In order to demonstrate this generality, we developed a *Thesaural Explorer* application. The Explorer reads a thesaurus from its TML file, presents it graphically, and supports browser style term navigation. The user selects a thesaurus to explore and then can navigate the structure along inter-term relations by clicking on terms or using various look up tables such as ordered lists by class, term alphabetic, and browsing history. Figure 6 is a screen image of the Explorer in action browsing the Keyword AAA thesaurus.

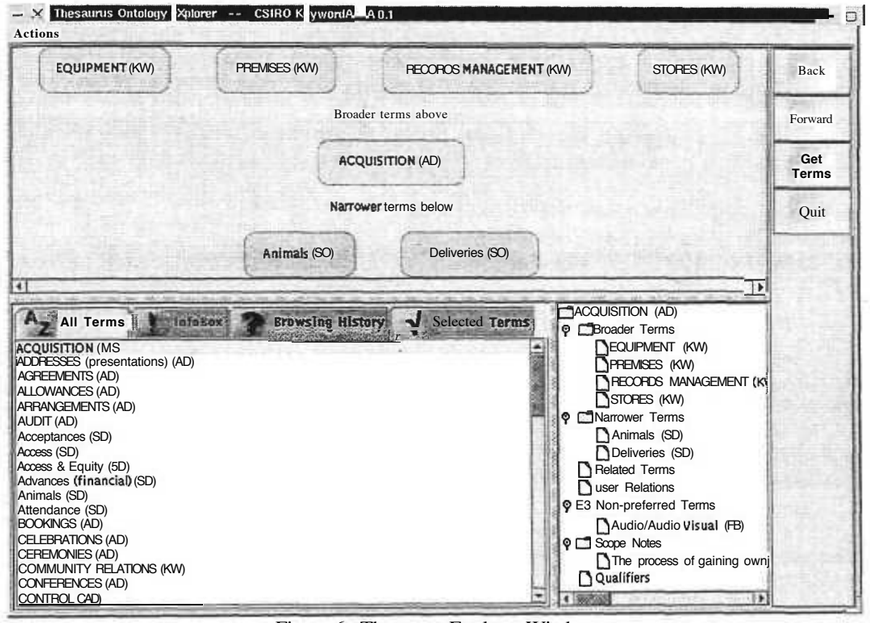

Figure 6. Thesaurus Explorer Window

The Explorer consists of three windows and a tab pane:

1. the *Terms Browser* (the window on the top) focuses on a particular term *(ACQUISITION),* and displays its type *(AD* - in a unique hue), the hierarchical relationships between terms, and indicates if a term is selected (using intensity).
2. the *Term Viewer* (the tree window at the bottom right) shows all of the information for a particular term in a single place. This allows the user to see the full structure of a term itself.
3. the *Tab Pane* (the tabbed window at bottom left) allows the user to select from a range of information views. The default tab is "All Terms" view.

The Tab Pane shows the textual information of the Term Viewer in various presentations designed to help the user maintain browsing context:

- The *All Terms* tab has every preferred term in the thesaurus sorted alphabetically, regardless of term type. This enables examination of all preferred terms in a single list and facilities navigation to known terms.
- The *InfoBox* tab gives a textual rendition of the information graphically presented in the Term Viewer.
- The *Browsing History* tab logs all terms that have been chosen for browsing in this window so far, with the capability to move back in history and return to each term in the order they were viewed.
- The *Selected Terms* tab lists the terms that the user has chosen so far, and also has the capability of previewing them all.

6 Discussion

Our goal was to provide generic thesaural support for resource description. The GTO is not the only attempt to model thesauri. Zthes [Zthes 1999] describes an abstract model for representing and implementing thesauri. Zthes proposes the Z39.50 attribute architecture, but so far no complete implementation exists. In comparison to the TML, Zthes supports a quite restricted set of relations. It is not possible to extend Zthes in the many instances where thesauri use a wider set of relations. A limitation of the Zthes' semantics is that it does not distinguish inter term relations from term to attribute relations.

Research in ontological modelling suggests that first-order logic and other formal languages enable more precise specification of messages [Fensel *et al* 1998, Farquhar, Fikes and Rice 1996, Finin et **al** 1994]. However, the simplicity of the GTO does not require such powerfully expressive languages. TML comes with minimal

semantics. Semantics are modelled only at the GTO level. Beyond that, at the level of individual thesauri, semantics are user defined in syntax extensions to relation types or pushed down into the application layer. We claim this as a strength, although it could be considered as a weakness from a theoretical language stance. But it is by this simplicity that we gain our generality — by concentrating on the high-level regularities and leaving low-level peculiarities to the syntax. This gives TML a tremendous advantage over languages understood only by computer scientists. We believe that the users and maintainers of document management systems should not need to have strong technical backgrounds to do their jobs.

Our aim with TML is to represent thesauri in a practical implementable way. The TML syntax is constrained through the use of a schema, but the schema does not fully specify the language; ie., we did not attempt to include all possible thesaural semantics or to prevent all representational errors within the syntax. The verification of compliance of a thesaurus instance to the model of the GTO requires some data validation to be carried out at the application level. This is also necessarily true for the class type and relation type extensions of a particular thesaurus. This is less of a danger than it might appear, because we expect such processing validation to occur in the TML authoring tools.

The choice of XML as the TML syntax is overwhelmingly pragmatic. XML is sufficient to our requirements, an open international standard, and is emerging as a software modelling standard. It is ubiquitous and can be understood and authored without great training. It allows TML maintenance and parsing tools to leverage the power of many off the shelf authoring products.

7 Conclusions

We have demonstrated that general thesaural support is feasible by designing a generic thesaural ontology and markup language that amalgamates different thesauri structure and allows us to represent the idiosyncrasies of specific thesauri in a common language. This permits general purpose thesaural tools such as our Thesaural Explorer to be built. These tools can work with many thesauri thereby leveraging development costs, providing a common user interface, and supporting flexible thesaural maintenance and evolution. Such tools permit document management systems to better organise and access repository content.

8 Acknowledgments

The work reported in this paper has been funded in part by the Research Data Networks (RDN) Co-operative Research Centre (CRC) program, Australia.

9 References

[AAT] Getty Art and Architecture Thesaurus, URL: http://www.gii.getty.edu/aat_browser

[Bradley 1998] Bradley, N. The XML Companion, Addison-Wesley, 1998.

[Dublin Core] An International recognised core set of metadata elements. URL: http://metadata.net.de/rdf/DC/

[Farquhar, Fikes and Rice 1996] Farquhar, A., Fikes, R., and Rice, J. The Ontolingua Server: a Tool for Collaborative Ontology Construction, Knowledge Acquisition Workshop, 1996.

[Fensel *et al* 1998] Fensel, D., Erdmann, M., Studer, R. Ontobroker: The Very High Idea, *Proceedings of the 11th International Flairs Conference, 1998*.

[Finin et al 1994] Finin, T., Fritzson, R., McKay, D. and McEntire, R. KQML as an agent communication language, CIKM'94 in the proceedings of the third international conference on information and knowledge management, 1994, 456-463.

[Glushko, Tenenbaum, and Meltzer 1999] Glushko, R., Tenenbaum, J., and Meltzer, B. An XML Framework for Agent-based E-Commerce, Communication of the ACM, March 1999, Vol 42, No 3, 106, 114.

[Harold 1998] Harold, E. XML: Extensible Markup Language, IDG Books Worldwide, 1998.

[ISO 2788] International Organisation for Standardisation (ISO) 2788, Documentation Guidelines for the establishment and development of monolingual thesauri, 1986.

[Keyword AAA] *Keyword AAA: A Thesaurus of General Terms*, Archives Authority of New South Wales, Sydney, 1995.

[Lancaster 1972] Lancaster, F., Vocabulary Control for Information Retrieval, Information Resources Press, 1972.

[LCSH] Library of Congress Subject Headings, URL at: http://www.loc.gov

[MeSH] Medical Subject Headings, URL at: http://www.nlm.nih.gov/mesh/filelist.html

[OCLC] OCLC Dewey Decimal Classification, URL at: http://www.oclc.org/oclc/fp/index.htm

[RDF] The Resource Description Framework for metadata syntax and interoperability, URL: http://www.w3.org/Metadata/RDF/

[TGN] Getty Thesaurus of Geographic Names, URL at: http://www.ahip.getty.edu/tgn_brower

[XML] Extensible Markup Language, URL: http://www.w3.org/XML/

[Zthes 1999] Z39.50 Profile for Thesaurus Navigation, URL: http://www.n-four.demon.co.uk/mirk/zthes-02.html

4shared
PDF Editor

Building rich metadata from critical reviews for a scrutable filtering system

Sacha Groves

Basser Dept of Computer Science
University of Sydney
AUSTRALIA 2006

sacha@cs.usyd.edu.au

Judy Kay

Basser Dept of Computer Science
University of Sydney
AUSTRALIA 2006

judy@cs.usyd.edu.au

Abstract

We describe the Review Coder system for creating rich metadata for a scrutable filtering system. A scrutable system maintains explanations of the data and processes that drove the system operation. In the current paper we use Review Coder as part of a filtering systems for movies: the scrutability of the system means that a user can determine why the system recommended a particular movie or not.

The filtering process is based upon movie reviews and metadata built in association with them. These provide high quality information about the movie objects. From these, the filtering system is intended to build stereotypic models of reviewer's preferences for movies. These can drive the filtering process and the user can scrutinise both these models and the actual reviews which were used to construct them.

Keywords: Multimedia Resource Discovery, Multimedia Filtering, Scrutable Filtering, Extraction of Metadata

1. Introduction

As electronic objects become increasingly accessible, there is a growing need for tools that assist users in filtering large collections of objects so that the user can find objects of interest. The effectiveness of a filter depends significantly upon the quality of the metadata describing the objects. Accordingly, this establishes a critical role for tools which assist in the creation of high quality metadata. The importance of such tools is indicated by the vigorous efforts to create a range of tools. These tools support a range of tasks, including, for example: creation of metadata templates with tools such as Dublin Core Metadata Template (Koch, Borell, and Berggren, 1998) or the discipline specific tools like Medical Metadata Creator (, 1999); tools such as Mantis (Shafer, 1998) which manage templates and assist in production of metadata.

Filtering for large video objects such as movies is important since there is a large cost for 'browsing' such objects compared with simple text objects. This cost is both in terms of the user's time and in the bandwidth required to deliver a segment of the object suitable for browsing.

Because the effectiveness of filtering depends so heavily upon the quality of the metadata, there seems to be promise in developing collections of very rich metadata for movie objects. An indication of the interest in this area is the number of online resources about movies, such as the Internet Movies Database (Database, 1999, Guide, 1999, Finder, 1999).

Such resources can be regarded as metadata suitable for human analysis as a basis for manually selecting or filtering movies. Equally, there is considerable activity in automation of the processes, for example, development of schemas for representing movie metadata. (Hunter and Iannella, 1998, Hunter and Armstrong, 1999).

An important model for assisting in filtering involves a three stage process: define a metadata structure; allow various providers to create metadata; allow various providers to create filtering tools which operate by using the metadata. A good example of such a model is PICS (Resnick, 1998) which provides a specification for metadata intended for rating objects so parents and teachers can filter out objects which are unsuitable for their children. It

Proceedings of the Fourth Australasian Document Computing Symposium, Coffs Harbour, Australia, December 3, 1999.