

A Web Based General Thesaurus Browser to Support Indexing of Television and Radio Programs

Hennie Brugman¹, Véronique Malaisé², Luit Gazendam³

¹Max Planck Institute for Psycholinguistics
P.O. Box 310, 6500 AH Nijmegen, Netherlands
Hennie.Brugman@mpi.nl

²Free University
De Boelelaan 1105, 1081 HV Amsterdam, Netherlands
vmalaise@few.vu.nl

³Telematics Institute
P.O. Box 589, 7500 AN Enschede, Netherlands
Luit.Gazendam@telin.nl

Abstract

Documentation and retrieval processes at the Netherlands Institute for Sound and Vision are organized around a common thesaurus. To help improve the quality of these processes the thesaurus was transformed into an RDF/OWL ontology and extended on basis of implicit information and external resources. A thesaurus browser web application was designed, implemented and tested on future users.

1. Introduction

The Netherlands Institute for Sound and Vision¹ is both the business archive for many public broadcast corporations in the Netherlands and the official keeper of the Dutch audio-visual heritage. It looks after more than 700.000 hours of television, radio, music and film recordings. Sound and Vision is facing the challenge of having to deal with a constant flow of thousands of hours of incoming broadcast quality digital video and audio per year. To archive all of this digital material in a way that keeps it accessible for future usage new methods and tools for the generation and enrichment of metadata and new retrieval techniques have to be developed and implemented. This challenge has led to a major upgrade of the Sound and Vision information infrastructure, including a complete redesign of its metadata set, the design and development of a matching software tool set, and the merging of legacy terminology sets into Sound and Vision's GTAA thesaurus (Gemeenschappelijke Thesaurus Audiovisuele Archieven – Common Thesaurus for Audiovisual Archives).

The CATCH research program² aims at developing techniques and software tools to support collection managers in Netherlands' cultural heritage institutions in the creation and exploitation of digital collections. Its 10 research projects are housed in participating cultural heritage institutions, thus reflecting the demand driven nature of the research program.

The CHOICE@CATCH³ sub-project that is housed at Sound and Vision primarily focuses on improving the quality and efficiency of the metadata creation process as performed by human cataloguers. It does so by combining Semantic Web and Natural Language Processing techniques. The latter are applied to contextual text

documents that are associated with television and radio programs (TV guide online, broadcasters' web pages, etc.). The texts are analyzed to automatically derive applicable thesaurus terms to be used for the catalog descriptions of the television or radio programs. The derived terms are presented to cataloguers, sorted by relevance order.

The thesaurus itself is turned into a true ontology by conversion to ontology standards. It is enriched by making implicit semantics explicit (e.g. some information present in the Scope Notes) and adding structure and content to the ontology with the help of information present in the thesaurus and external resources. This enables new ways of browsing and searching for terms. Also, the quality of retrieval of audiovisual documents will be substantially increased by enabling semantic search and reasoning over catalog data.

In this paper we focus on design and implementation of a web application for browsing and searching for terms in the GTAA. It allows users to exploit all structures that are present in the (old and new) GTAA. This GTAA Browser application will be used by Sound and Vision cataloguers, by broadcast corporations (to optimize the delivery of content to Sound and Vision), and by researchers in other CATCH projects (e.g. to explore automatic procedures for mapping different ontologies).

Section 2 discusses the thesaurus, its conversion to a suitable Semantic Web representation and thesaurus enrichment. Section 3 presents the design, user interface and architecture of the Browser. A discussion of our work and plans for the future can be found in sections 4 and 5.

2. The GTAA

The GTAA thesaurus plays a central role in the CHOICE project as the primary source for vocabulary used in the Sound and Vision documentation process.

It contains approximately 160.000 terms. The GTAA terms are divided in 6 disjoint facets: Keywords (~3800 terms), Locations (~14.000), Person Names (~97.000),

¹ <http://www.beeldengeluid.nl>

² <http://www.nwo.nl/catch>

³ <http://www.nwo.nl/catch/choice>

Organization-Group-Other Names (~27.000), Maker Names (~18.000) and Genres (113 terms).

The thesaurus mainly uses constructs as presented in the ISO 2788 standard (ISO, 1986) and commonly used in companies or institutions: amongst others, Broader Term, Narrower Term, Related Term, Scope Notes.

Terms from all facets of the GTAA may have Related Terms and Scope Notes, but only Keywords and Genres can also have Use/Use for and Broader Term/Narrower Term relations, the latter organizing them into a set of hierarchies. Additionally, Keyword terms are thematically classified in 88 subcategories of 16 top Categories. Although the data model that is used for the thesaurus allows links between terms across facets, no instances of these links currently exist.

The hierarchical structure of a thesaurus can be the basis of simple reasoning, such as query expansion based on the nearest parent of a term in a hierarchy, but for more sophisticated inferences, a richer representation of the data is required. Semantic Web technology enables these sophisticated inferences, so one of the first aims of the project was to convert the thesaurus into a Semantic Web compliant format. These formats, based on RDF⁴ or OWL⁵ also enable us to state explicitly some knowledge related to the thesaurus (for example, the fact that the facets are disjoint, that some relations are symmetric, etc.). Having converted the thesaurus to RDF/OWL also helps the information exchange on the web, by making its content unambiguous, and makes it possible to enrich it with other existing resources, for example the Getty Thesaurus of Geographical Names⁶.

According to (van Assem et al., 2006) “Thesauri can be converted to RDF/OWL in different ways.[...]. This can introduce structural differences between the conversions of two thesauri which have the same semantics. Using a common framework for the RDF/OWL representation of thesauri either enables, or greatly reduces the cost of (a) sharing thesauri; (b) using different thesauri in conjunction within one application; (c) development of standard software to process them (because there is no need to bridge structural differences between mappings).” SKOS⁷ (Simple Knowledge Organization Scheme) is therefore proposed as such a common framework.

SKOS is an extensible model that enables the RDF/OWL representation of thesauri. It provides a standard definition for the most common ISO 2788 constructs, which become RDF/OWL properties between Concepts. The Concepts have an ID, a Preferred label and may have one or more Non-Preferred labels: this is how the Use/Use for thesaurus relationship is modeled. SKOS proposes a set of additional constructs, beside the most common ones, and can be extended for specific purposes. We tried to stay as close as possible to the SKOS Core representation, but had to use some of the additional constructs for converting the GTAA.

The only particularities of the GTAA are: an alternative non-hierarchical grouping of terms into Categories, and a set of cross-facet links between terms.

As the Categories are meant to show a grouping of terms when browsing the thesaurus, and not to use while indexing, we modeled them with the <skos:Collection> construct. This construct provides a semantic label to a grouping of Concepts. We modeled the cross-axis links as extensions of the <skos:related> property. Thus, our OWL representation of the GTAA is bounded to SKOS constructs and permitted extensions.

Additionally, we interpreted the information contained in some Scope Notes of Person Names. As we mentioned before, no instances of the cross-facet relation were provided in the original GTAA; but we used these Scope Notes to create links between a Person Name and his profession when this profession occurs in the Keywords. We also manually created a nested hierarchy of these professions as a browsable access to the Person Names. As future work, we intend to bring more structure to the GTAA by using information contained in external resources, as for example an ontology of Dutch politics provided by (Van Atteveldt, 2005) and the hierarchical structuring of location names in the Getty Thesaurus of Geographical Names.

3. The Thesaurus Browser

Cataloguers at Sound and Vision currently only have access to terms from the thesaurus in the form of alphabetically sorted flat lists. Although the GTAA has substantial internal structure this is not exploited by the current generation of software tools. Therefore, as a first step to improve the cataloguing process a thesaurus browser application was designed, implemented and tested on potential future users from both Sound and Vision and Dutch broadcasting corporations.

3.1. Requirements

A number of requirements for the browser's user interface and architecture can be formulated. Because the GTAA Browser will be used by both incidental and regular users, and because these users are located both inside and outside of Sound and Vision, a web based application is strongly preferred. Furthermore, within the CATCH project it is desirable that an RDF/OWL version of the thesaurus is accessible over the web, not only for interactive access by humans but also for software clients.

Parts of the thesaurus content are regularly updated, for example Person Names and Locations. There is one authoritative resource for the GTAA, which is a relational database system maintained at Sound and Vision. Since the browser has to be usable on an every day basis for real cataloguing purposes it should at all times show the actual state of the thesaurus. It should therefore be able to directly use the database system as its data source.

The browser should be able to display and exploit all structures that are present in the thesaurus in appropriate and intuitive ways. The same is true for structures and information that we add to the thesaurus.

3.2. User Interface

A user interface was designed, implemented, tested with users and revised. A representative screen shot of the current GTAA Browser can be seen in figure 1.

⁴ Resource Description Framework – www.w3.org

⁵ Web Ontology Language – www.w3.org

⁶ http://www.getty.edu/research/conducting_research/vocabularies/tgn/

⁷ SKOS - <http://www.w3.org/2004/02/skos>

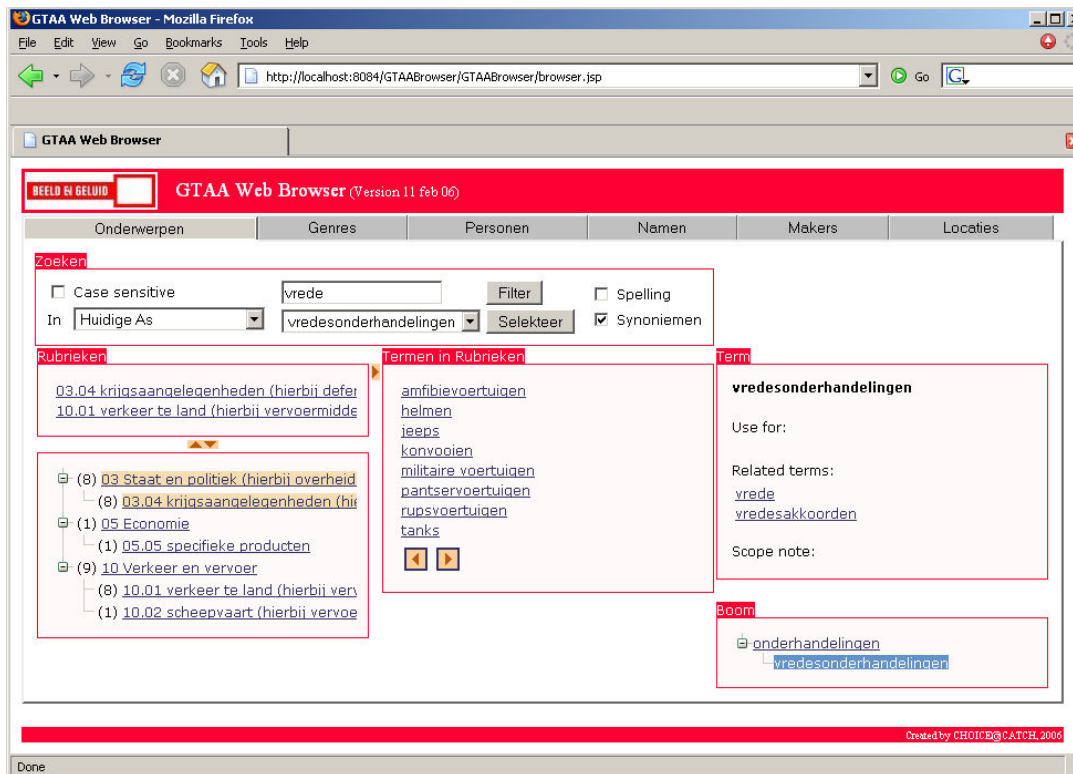


Figure 1: GTAA Browser user interface

Since the GTAA's facets are disjoint each facet's terms can be accessed on a different tabbed pane. Most thesaurus facets currently have little structure. In these cases their terms are presented as scrollable alphabetic lists. When the number of terms in a facet gives rise to this, terms on a tab are presented per start character.

The Keyword facet is the one with most structure. All keyword terms are classified in Categories. This Category hierarchy is displayed on the left side of the Keyword tab (labeled "Rubrieken"). Clicking on a Category displays its terms in the scrollable list in the middle (labeled "Termen in Rubrieken").

Clicking on a term link, in any of the panels on a tab, results in selecting this term. The panel on the right (labeled "Term") shows all available information about the selected term, including Non-Preferred terms, Related Terms and Scope Notes. The bottom right panel (labeled "Boom", for "tree") shows the selected term situated in its Broader/Narrower Term hierarchy. All categories that a selected term belongs to are highlighted in the Category hierarchy on the left.

First user tests showed that the average number of terms per (sub)category is too large to allow efficient selection of an appropriate term in the middle term list. Therefore we decided to use the fact that most terms are in more than one category in the following way: instead of one, more categories can be selected by clicking on them. Selected categories are shown in the top part of the "Rubrieken" panel. For all categories, the category hierarchy shows (in parenthesis) the number of terms in that category that are also in all of the selected categories. When this number is zero for a category, the category is

not displayed anymore. The middle panel shows the cross section of the terms from the selected categories. When more than one category is selected, the remaining number of terms is usually small enough to allow quick human selection of terms. For example, selecting the categories 'military affairs' and 'traffic on land' leaves just 8 terms related to military land vehicles. This 'faceted' navigation allows users to gradually narrow down the number of terms to display based on a multi-thematic classification.

The top of each tab in the GTAA Browser shows a search panel that is shared between all facet tabs. It allows users to search for terms starting with characters entered in a text box. The user can select a term from the result list and set it to be the active term in its facet. To expand the set of text values that lead to proper thesaurus terms two additions to the search functionality were made. First, we used an existing spelling suggestion tool (GSpell⁸) to find thesaurus Keyword terms that can be considered spelling variations of the text entered by the user. The best suggestions are added to the search result list. Second, we generated a list of synonyms for thesaurus Keyword terms on basis of external resources. We used information provided by the online Van Dale reference dictionary⁹ and another dictionary called Muiswerk Woordenboek¹⁰ to connect thesaurus terms with their synonyms (by automatic processing and manual checking of the result). So far, we could associate about 1700 Keywords with one or more synonyms.

The structures that we added to the thesaurus are used in the GTAA Browser's user interface. For example, the

⁸ <http://lexsrv3.nlm.nih.gov/SPECIALIST/Projects/Summary/gSpell.html>

⁹ <http://www.vandale.nl/opzoeken/woordenboek/>

¹⁰ <http://www.muiswerk.nl/WRDNBOEK/INHOUD.HTM>

cross-facet links generated from scope notes in the Person Name facet can be used as bidirectional links. When looking at the term information about 'Alexander II' a link to the Keyword 'czar' is available, from 'czar' a list of links to specific czars in the Person Names can be shown.

3.3. Architecture and Implementation

Figure 2 shows the architecture of the GTAA Browser.

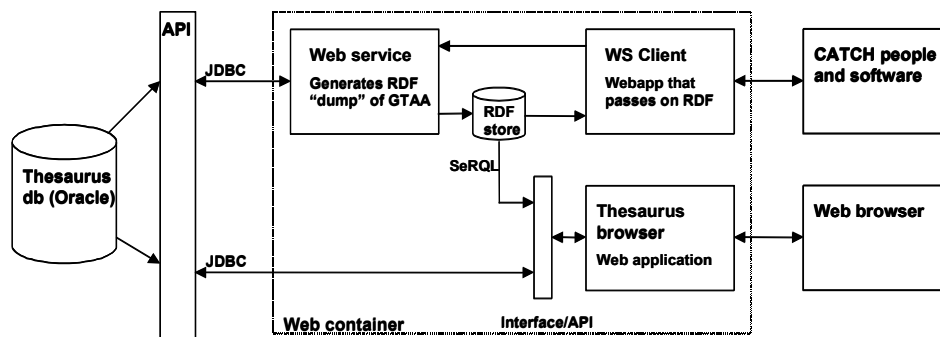


Figure 2: GTAA Browser Architecture

The Browser is implemented using a 3-tier web application architecture. The web application can retrieve the thesaurus data from an extensible set of data sources. Because radio and television professionals need the most recent content of the thesaurus at all times, the data can be directly retrieved from the relational database that is the primary GTAA source. And because CHOICE and CATCH research requires an RDF/OWL representation of the thesaurus, the browser can also retrieve the data from an RDF store using the SeRQL¹¹ query language. The RDF store can be updated on request using a separate web application. This will allow researchers to either download RDF or to query the store over the internet (using SeRQL).

4. Evaluation

We performed a set of user studies to test the reliability and the ease of use of the browser. The test was divided in three parts: first, the users got acquainted with the functionality of the tool and its general display by performing a guided tour, then they used the browser to select keywords for annotating a TV news item, and finally they filled in a satisfaction form. The first observations that we present here result from the analysis of this satisfaction form and from some general comments that were made during the test.

The users who tested the browser were on the one hand cataloguers from Sound and Vision, who use the thesaurus as an alphabetical list of terms in their daily work, and on the other hand people from broadcasting corporations, who create and index their TV programs using different indexing schemes and vocabularies. So the first set of users were specialists on the thesaurus content and on the task of indexing TV or radio programs,

whereas the second set of users were only experts on the latter.

The users from Sound and Vision, despite the fact that they knew the thesaurus content, had never seen it displayed with all its hierarchies. They found its organization complex, but they liked the possibility to browse these multiple hierarchies to find out terms. Nevertheless, they typically started their search for

relevant terms by using the search functionality. The hierarchies were mostly used as a starting point for term retrieval by users from broadcasting corporations, as they did not know how the notions that they were looking for were expressed in the GTAA. Many users expressed interest in additional thesaurus structure for browsing Person Name, Names, Makers and Location facets. Some changes to the browser's user interface and layout are already implemented in response to user's comments.

5. Future work

Since the Browser uses a fully SKOS compliant representation of the thesaurus data, and already deals with most of the core SKOS constructs in some way, we intend to turn it into a generic browser for SKOS.

Additionally, the browser should be reusable as a thesaurus browsing component in other software systems. It could for example be part of a metadata editor (to provide controlled vocabulary to the user), part of a search user interface (to provide structured access to query terms or documents) or part of a terminology mapping tool (to map legacy terminology to concepts in an ontology).

6. Acknowledgements

This work was made possible by NWO. The authors wish to thank Lora Aroyo and Cristian Negru for their active participation in our user studies. And of course our work would not have been possible without the active and interested participation of the people at Sound and Vision.

7. References

- Mark van Assem, Véronique Malaisé, Alistair Miles and Guus Schreiber(2006). *A Method to Convert Thesauri to SKOS*. To appear in ESWC 2006.
- Wouter van Atteveldt and Stefan Schlobach (2005). *A Modal view on Polder Politics*, presented at Methods for Modalities (M4M) 2005, (Berlin, 1-2 December)
- ISO 2788 standard (1986): International Organization for Standardization. *Documentation-guidelines for the establishment and development of monolingual thesauri*. Iso 2788-1986, 1986.

¹¹ <http://www.openrdf.org/doc/sesame/users/ch06.html>