A COMPARATIVE STUDY OF THESAURI TOOLS
A Perspective from Integrability in Information Systems

Beatriz Pérez-León and M. Mercedes Martínez-González
Department of Computer Science, University of Valladolid, Campus 'Miguel Delibes' s/n, Valladolid, Spain

Keywords: Thesauri, Thesauri tools, Software integrability, SKOS, RDF, Semantic web.

Abstract: The Semantic Web has brought a renewed interest in thesauri as support for semantic searches and other added-value services. Tools that manage thesauri permit to create, edit, and query thesauri. But there is also the possibility to import thesauri and to integrate thesauri. In fact, integrability at the information level has also received an important push with the stabilisation of the SKOS standard as a W3C Recommendation. In this paper several thesauri tools are evaluated and compared.

1 INTRODUCTION

Thesauri are conceptual tools that permit to organise topics from a domain in a hierarchy of concepts and subconcepts (topics and subtopics). This permits navigation through topics and *concept-based* searches, which are so valued in the Semantic Web context.

The W3C SKOS (Standard Knowledge Organization System) standard (Alistair Miles and Sean Bechhofer, 2009) has stabilised very recently. But even before its stabilisation, numerous initiatives have been proposed to represent various thesauri with RDF/SKOS (Lacasta et al., 2007; Faro et al., 2008; Polo et al., 2008). SKOS and RDF are the two Semantic Web standards that will permit thesauri to get into the integration arena, so important in web information systems.

While traditionally thesauri were created manually or with text editors, new thesauri tools permit to create, edit, or use (manage) thesauri. Previous reviews of thesauri tools (Martinez and Leiva, 2001), (Gilbert and Butler, 2003) focused on software features, like creation and management of thesauri and features related to software output (visualisation of thesauri on the screen, printer, etc). In this paper we review a set of thesauri tools, but we focus on compatibility with Semantic Web standards (RDF/SKOS) and integrability or reusability of software. We define a set of criteria for comparison, which include purpose and functionality, interoperability and integration at the information level, and reusability of software as components of third-party applications in information systems.

In this section we present a brief introduction to the main characteristics of thesauri, and the criteria used for evaluation (and comparison) of thesauri tools. In section 2 we present the tools evaluated and the results of the evaluation. Section 3 summarises the conclusions obtained from the comparison of these tools.

1.1 Thesauri

Thesauri are *controlled vocabularies* that collect and organise terms from a domain. Terms are grouped into *concepts* and relationships (equivalence, hierarchical, associative) are established among them. Notes are used to clarify their meaning.

The *Equivalence relationship* divides terms in preferred, which will be used for the indexing and retrieving process, and non-preferred terms. The hierarchy of concepts is indicated by *Broader Term* and *Narrower Term* relationships. The *Associative relationship* is established between terms that are conceptually related, but their relationships are neither hierarchical nor equivalence relationships in nature. These are referred to as *Related Terms*. Finally, *Notes* serve to clarify the meaning and application of a term in relation to other terms in the thesaurus. Notes can be divided in scope notes, historical notes, editor notes, usage note, etc.

Perez-LeÃşn B. and Mercedes MartÃ■nez-GonzÃąlez M.
A COMPARATIVE STUDY OF THESAURI TOOLS - A Perspective from Integrability in Information Systems.
DOI: 10.5220/0002800502030206
In *Proceedings of the 6th International Conference on Web Information Systems and Technology (WEBIST 2010)*, page
ISBN: 978-989-674-025-2

1.2 Methodology Applied

The methodology followed is inspired on the one proposed by (Martinez and Leiva, 2001). It organises the work in six steps:

1. Define criteria and tests
2. Search on the web for software of management of thesauri that provide evaluation versions[1]
3. Download the software
4. Read documentation
5. Perform the tests
6. Perform the analysis of results and draw conclusions based upon the tests

1.3 Criteria for Evaluation

In this part we present the criteria used for the evaluation of the tools considered in section 2. Criteria 4 and 5 are related with compatibility with Semantic Web (RDF/SKOS), integrability and extensibility of software. The criteria for thesauri tools' evaluation are:

1. *Purpose*. Purpose for which the tool was created.
2. *Functionalities*. We have selected functionalities considered relevant in some previous analysis about what thesauri tools should offer (Severino, 2007). In this paper, because of space limits, we will only consider the most relevant functionalities: creating and editing thesauri, searches and retrieval in thesauri, and navigation.
3. *Thesauri Structures Supported*. The thesauri structured considered in this study are: terms (Preferred/descriptor, Non-preferred/non-descriptor), equivalence, hierarchical, and associative relationships. Notes (scope, historical, editor, ...) complete this list.
4. *Formats Supported*. The ability to import and/or export thesauri represented with different formats (including the Semantic Web SKOS standard) is an important one for the aim of interoperability.
5. *Provision of Software as Packages or Services*. Software integration is achieved by means of software packages that can be used by other tools (.jar packages for Java applications, widgets, services that exchange XML messages, ...).

1.4 Tests Performed

The set of tests we have prepared to compare and draw conclusions are:

- *Test 1*. To create a new Thesaurus from scratch.
- *Test 2*. To insert terms, relationships and notes.
- *Test 3*. To create a term which is at the same time a narrower and broader term of other.
- *Test 4*. To search different terms.
- *Test 5*. To import a Thesaurus.
- *Test 6*. To export a Thesaurus.

Checking automatically thesaurus integrity is considered an important function of thesauri tools by experts in thesauri creation ((Sánchez, 2009), (Martinez and Leiva, 2001)). For this reason, we included some tests to check this capability (test 3).

2 EVALUATION OF THESAURI TOOLS

2.1 ThManager

ThManager [2] is distributed under the terms of the GNU Lesser General Public License as published by the Free Software Foundation. It was first released in 2007.

- *Purpose*. ThManager is a thesaurus editor tool that facilitates the creation and management of thesauri using the SKOS format.
- *Functionalities*. ThManager has two modes of operation, the Thesaurus Viewer and the Thesaurus Editor (which guides the user through the edition process). Terms can be added in the thesaurus one by one or imported from files in RDF/SKOS format.The main limitation of this editor is that integrity checks are not automatically done by the tool, which means that it did not pass Test 3.

  The Thesaurus viewer is used to navigate through concepts. In this tool it is not possible to perform an advanced search as a combination of logical operators (AND, OR, NOT). However, "Exact match", "Starts with" and "Contains" searches are provided. To encourage re-use of thesauri, ThManager offers a metadata profile (title, language, etc) of its thesauri based on Dublin Core.
- *Thesauri Structures Supported*. ThManager supports the three most important relationships in a

[1] Except in the case of PoolParty. In this case we could not have access to the software despite the fact that we registered for evaluation, and the tests were performed by checking in its documentation if test operations were supported.

[2] http://thmanager.sourceforge.net

thesauri, the equivalence, the hierarchical, and the associative relationship. Regarding notes, ThManager only supports scope notes.

- *Formats Supported*. ThManager supports RDF/SKOS as a format not only to import but also to export data.
- *Provision of Software as Packages or Services*. Although ThManager is supposed to use an API to access the thesauri stored in its repository (Lacasta et al., 2007), it is distributed as an 'off-the-self' application, with which this API is not provided. As well, its interface is not included in the documentation provided with the tool. So, we could not access or know the methods that this API offers.

2.2 TemaTres

TemaTres [3] is a web application to manage documentation languages. In August 2009 the 1.032 version was released.

- *Purpose*. TemaTres is oriented to the development of hierarchical thesauri, on which several editors can be working at the same time.
- *Functionalities*. When creating a new term in a thesaurus, TemaTres defines a workflow with three states for a term, *Candidate*, *Active* and *Rejected*. These states define what actions can be performed with the terms. Validation is left to the user, as the tool does not automatically check the coherence of the thesaurus.

  TemaTres presents both a systematic and an alphabetical list of terms. It offers different options to perform searches: simple search, expanded search through related or hierarchical terms, search terms suggestion (*Did you mean:...*).
- *Thesauri Structures Supported*. The relationships allowed are the equivalence, the hierarchical, and the associative relationship. Other relationships can be established between terms of different thesauri. Any term can be annotated with scope notes, historical notes, bibliographic notes or private notes.
- *Formats Supported*. In TemaTres there is no option to import a thesaurus from any format, but any thesaurus in TemaTres can be exported to text format, Dublin Core, SKOS-Core, Zthes, MADS, TopicMap (XTM 1.0), SiteMap 0.8, RSS.
- *Provision of Software as Packages or Services*. TemaTres provides web services since version 1.0.32.

2.3 Term Tree

Term Tree [4] was first released in 1999.

- *Purpose*. TermTree is a specialised thesaurus software for creating and maintaining complex thesauri and taxonomies.
- *Functionalities*. In this application, a thesaurus can be created inserting terms one by one, or importing terms from different file formats, such as ASCII structured tag file or MultiTes database. When creating a relationship between terms the validity of the requested link is checked.

  Term Tree offers a powerful tool for searching. The user may choose to search terms, fields associated with a term, search by relation, or search by date ranges (creation and modified dates). Standard wild cards, "*" and "?" are supported.
- *Thesauri Structures Supported*. Term Tree supports the three most important relationships in a thesauri: the equivalence, the hierarchical, and the associative relationship. Regarding annotation, Term Tree supports the scope note and usage note.
- *Formats Supported*. A thesaurus can be exported to several text formats, Microsoft Excel spreadsheet, HTML, XML, and others. However, the formats used in the import process are restricted to Term Tree tagged, MultiTes database and Aka database.
- *Provision of Software as Packages or Services*. According to Term Tree documentation, there is an API available for custom application development. However this API is not distributed in the demonstration kit or documented elsewhere, which means that we could not have access to it.

2.4 PoolParty

PoolParty [5] is a thesaurus management system for the semantic web. We selected it because it is the first tool in which software integration in third-party applications and Semantic Web goals are basic pillars of its design.

- *Purpose*. PoolParty can be used to build and maintain multilingual thesauri providing a simple user interface.
- *Functionalities*. PoolParty is a web based application. The interface of PoolParty displays the information in two areas: the Concept hierarchy tree

[3] http://tematres.r020.com.ar

[4] http:/ /www.termtree.com.au

[5] http://poolparty.punkt.at/

and the Concept details view. A thesaurus in PoolParty can be created by manual editing. When creating a term or setting relationships among terms there are drag and drop and autocomplete facilities to help the user. PoolParty has the ability of extracting words and phrases that characterise a document automatically. The only type of search referred in the manuals is search based on suggestions.

- *Thesauri Structures Supported*. PoolParty supports the three most important relationships in a thesauri, the equivalence, the hierarchical, and the associative relationship. Regarding annotation, PoolParty only supports the scope note. PoolParty is compliant with most of the elements of the SKOS standard. To avoid building a thesauri which violates SKOS, it provides quality queries to check the thesaurus' integrity.
- *Formats Supported*. PoolParty is able to import existing SKOS thesauri, which meet certain prerequisites, and to export to different formats such as RDF/XML, N-Triples, Turtle, N3, TriX and TriG.
- *Provision of Software as Packages or Services*. According to its documentation, PoolParty offers Web Services to access data via SOAP and REST.

3 CONCLUSIONS

There are not many free thesauri tools available. We evaluated four tools that can either be downloaded for free, either offer an evaluation version for free. Basic functionalities are supported in all the tools evaluated in a similar manner. All the tools revised permit the construction and management of thesauri. However, validation is only provided in some of them (Term Tree and PoolParty). There are also differences in the provision of advanced search. All of them support basic thesauri structures, with some differences on the range of notes supported.

Interoperability has received the main attention at the information exchange level. Almost all the tools studied support thesauri import/export by means of plain text formats or standard formats. SKOS/RDF Semantic Web standards are not yet supported in all tools. We think that this will change in the future with the spread of RDF/SKOS.

Despite some tools are supposed to be designed in layers that include APIs for thesaurus management (Lacasta et al., 2007), (Ferreyra,) they do not provide a public API or package that can be reused independently. Thus, interoperability at this level seems to be the weak aspect of free thesauri software.

ACKNOWLEDGEMENTS

This work was partially supported by the *Consejería de Educación de la JCyL* research projects VA010B06 and VA012B08.

REFERENCES

Alistair Miles and Sean Bechhofer (2009). *SKOS Simple Knowledge Organization System Reference. W3C Recommendation 18 August 2009*. http://www.w3.org/TR/2009/PR-skos-reference-20090818/.

Faro, S., Francesconi, E., Marinai, E., and Sandrucci, V. (2008). Eurovoc studies lot2 d2.3 -report on execution and results of the interoperability tests. Technical Report 10118, Publications Office of the EC, Institute of Legal Information Theory and Techniques ITTIG.

Ferreyra, D. Tematres: software libre para gestión de tesauros. URI: http://www.r020.com.ar/tematres/index.html.

Gilbert, J. and Butler, M. (2003). Review of existing tools for working with schemas, metadata, and thesauri. Technical Report HPL-2003-218, Digital Media Systems Laboratory, HP Laboratories Bristol.

Lacasta, J., Nogueras, J., Lopez-Pellicer, F. J., Muro-Medrano, P., and Zarazaga-Soria, F. (2007). Thmanager: An open source tool for creating and visualizing skos. *Information Technology and Libraries (ITAL)*, 26(3):39–51.

Martinez, G. M. and Leiva, I. G. (2001). Evaluacion de softwares de gestion de tesauros. *Ciencias de la Información*, pages 32(3), 3–23.

Polo, L., Alvarez, J. M., and Rubiera, E. (2008). Promoting government controlled vocabularios for the semantic web: the eurovoc thesaurus and the cpv product classification system. In *ESWC 2008, 1-5 june 2008, Tenerife, Spain*, volume 50212 of *Lecture Notes in Computer Science*, pages 111–122. Springer.

Severino, F. (2007). What thesaurus to define eu/acp relations? *European journal of developmental research, n 2*, pages 327–351.

Sánchez, J. P. (2009). Diseño de un sistema colaborativo para la creación y gestión de tesauros en internet basado en skos. Master's thesis.