Languages for Special Purposes in a Multilingual, Transcultural World

Proceedings of the 19th European Symposium on Languages for Special Purposes, 8-10 July 2013, Vienna, Austria

http://lsp2013.univie.ac.at/proceedings

Creating ontology from Persian thesauri

Molouk Sadat Hosseini Beheshti; Fatemeh Ejei

Cite as: Hosseini Beheshti, M. S. & Ejei F. (2014). Creating ontology from Persian thesauri. In G. Budin & V. Lušicky (eds.), *Languages for Special Purposes in a Multilingual, Transcultural World, Proceedings of the 19th European Symposium on Languages for Special Purposes, 8-10 July 2013, Vienna, Austria*. Vienna: University of Vienna, 434-441.

Publication date: July 2014

ISBN: **978-3-200-03674-1**

Creating ontology from Persian thesauri

Molouk Sadat Hosseini Beheshti; Fatemeh Ejei
Terminology Research Group,
Iranian Research Institute for Information Science and Technology (IRANDOC)
Iran

Correspondence to: beheshti@irandoc.ac.ir

Abstract. Ontologies can play a significant role in information systems, natural language processing and knowledge engineering. As common lexicon is the prerequisite for knowledge sharing through language, shared ontologies is the prerequisite for knowledge sharing through information technology. To speed up the ontology development process, as ontology developers are reusing all available ontological and non-ontological resources such as different domain ontologies and lexicons, we use the basic sciences thesauri previously developed at IRANDOC as resources for ontology construction. For this purpose, we firstly merge thesauri and transform the data format into ISO 25964 standard. Then, we built conceptual model based on the terms and their relationships in thesauri and the concept maps that were designed by domain experts for each of basic sciences (Chemistry, Physics, Biology, Geology and Mathematics). Ultimately, the ontology was generated by implementing the model in OWL, an ontology implementation language. The aim of this project is to create a standard ontology to be used in information retrieval system.

Keywords. Information retrieval, IRANDOC, ontology, Persian thesauri.

1. Introduction

In recent years, development of World Wide Web and its related technologies influence representing and retrieving knowledge in the field of information science. These new technologies enable machines to understand, process, and retrieve relevant information. In particular, ontologies are used to describe and represent knowledge and can enhance the performance of information processing systems.

However, developing ontologies is a time consuming and labor work, so many ontology developers try to facilitate and speed up this process by reusing other resource such as thesaurus. In particular, (Soergel 2004) and (Kawtrakul 2005) try to reengineer AGROVOC into ontology by building the ontology on the information contained in thesaurus and refine the information as needed. Moreover, in (Huang 2007), the Inspec thesaurus is used to enrich core ontology in the IT domain. In Persian language, Khosravi and vazifedoost (khosravi 2008) work on re-engineering an ASFA thesaurus into ontology in the field of library and information science.

In fact, thesaurus contains semantic information and hierarchical structure that make it an appropriate resource for ontology construction. Therefore, we determined to transform the basic sciences thesauri, previously developed at IRANDOC, into ontology that can be used in our information retrieval system. In the rest of the paper, thesaurus and ontology are compared firstly. Then the ontology development process is described. The ontology refinement issue is mentioned at last.

2. Thesaurus versus ontology

Thesaurus consists of terms and their relationships and its prime application is in information retrieval. The traditional aim of a thesaurus is to guide indexer and searcher to choose the same term for the same concept (ISO25964-1 2011). Terms stand for concepts in thesaurus. Each concept can be represented by one or more terms but just one term is selected as the preferred term per language for a concept. An equivalence relationship should be established between a preferred term and its corresponding non-preferred term.

In addition, two kinds of relationships are distinguished between concepts: hierarchical (BT/NT) and associative (RT). These relationships are established only between preferred terms. Whenever the scope of one concept falls completely within the scope of other concept, hierarchical relationship should be established between them. Similarly, associative relationship is used between terms that are conceptually or semantically related and their relationship is not hierarchical.

On the other hand, Ontologies consist of concepts (also Known as classes), relations (properties), instances and axioms. It is used by people and application to share the meaning of a particular area of knowledge and can be used in formal and informal reasoning (Sowa 2010). Fig. 1 shows a comparison between thesaurus and ontology based on the triangle of meaning. A thesaurus generally works with the left-hand side of the triangle (the terms and concepts), while an ontology, in general, works more with the right-hand side of the triangle (the concepts and referents)(Daconta, et al 2003).

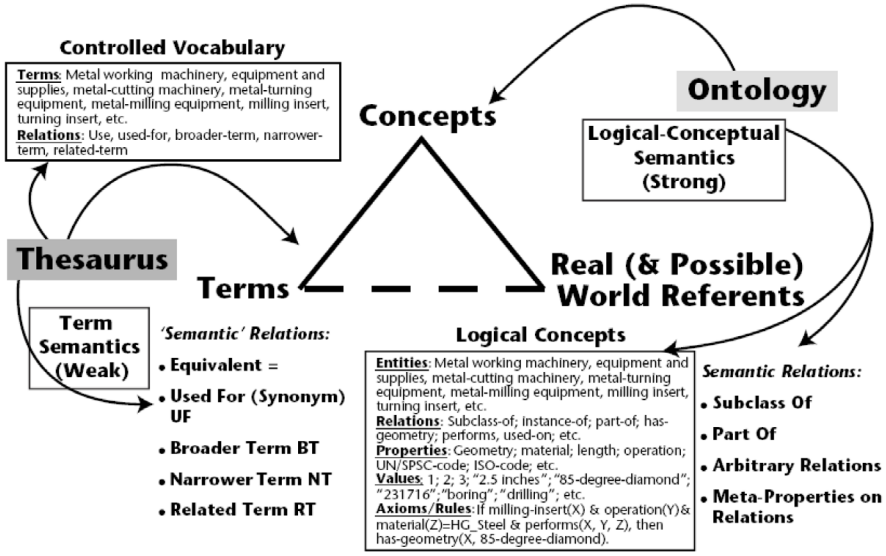

Figure 1: Thesaurus versus Ontology (Daconta, et al 2003)

3. Creating ontology from thesaurus

3.1. Synchronization and integration

We used bilingual (Persian/English) thesauri of basic sciences (chemistry, physics, biology, geology, and mathematics), which were previously developed at IRANDOC, as resources for ontology construction. Within a collection of tens of thousands of terms that were produced in different times and by different experts, we needed to synchronize common concepts in thesauri before integrating them as a macro thesaurus.

To reduce the amount of time and human resources which were needed for synchronizing process, Thesaurus Synchronizer was developed using Thesaurus Builder to illustrate differences between matched cases of two thesauri. The differences between thesauri are examined based on ISO 25964 standard. It also provides powerful tools for demonstrating differences and suggestions for each of the existing matters. Therefore, domain experts synchronized each two thesaurus semi-automatically.

The Thesaurus Synchronizer examines the following issues within two thesauri:

- Differences in transcription of the same concept,
- Differences in narrower terms of the same concept,

- Differences in non-preferred terms of the same concept,
- Differences between the translations of the same term in a specific language,
- Differences in related terms of the same concept,
- Lack of a related term for a concept in one thesaurus,
- Using the same translation for two different terms in a particular language,
- Different selection of a preferred term for one concept,
- Different relationship type between two concepts,
- Infinite loop between concepts (conceptual network),
- Different concepts related to the same term (polysemy).

After domain experts synchronized all thesauri completely, the integration process must be done to produce a macro thesaurus which can be transformed into ontology. The integration of basic science thesauri also was done semi-automatically by domain experts. The thesaurus format was transformed from ISO 5964 into ISO 25964 thereafter.

3.2. Methodology

Our methodology for ontology construction formed based on METHONTOLOGY (Gómez-Pérez 2004). This methodology enables the construction of ontologies at the knowledge level. We also consider the approach for re-engineering non-ontological resources into ontology presented in (Villazón-Terrazas 2010). So we first extracted the conceptual model of our thesaurus based on the concept maps previously designed by domain experts and the structure of thesaurus and then developed ontology using METHONTOLOGY methodology

In METHONTOLOGY, the main activity is ontology conceptualization because it determines the rest of the ontology development process. The aim of this activity is to design the knowledge representation paradigms and regulate knowledge based on the implementation language which will be used to formalize and implement the ontology. After building the conceptual model, it can be transformed into formalized model. The next step in methodology is to implement formalized model in an ontology language, so that the knowledge model is moving gradually to the implementation level during the process and can be understood by a machine. Fig. 2 shows the process model in ontology development. The discontinuous line in the figure shows that the transformation from conceptual model into formalized model may be done incompletely because some domain knowledge may be lost along the conversion process.

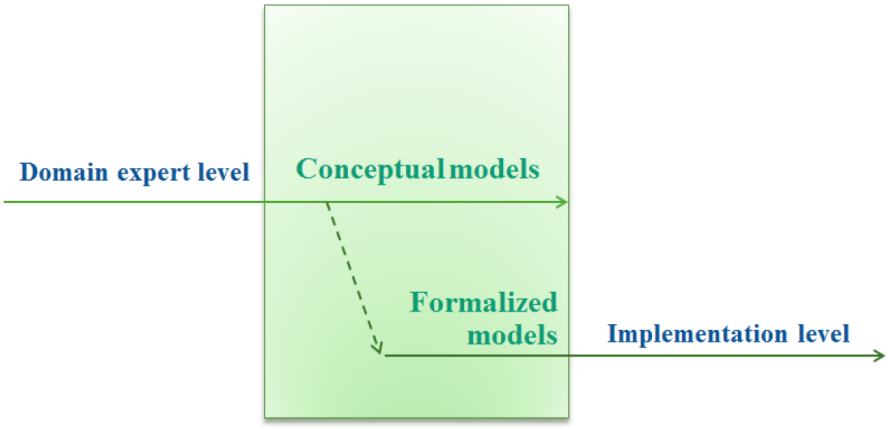

Figure 2: Essential process model in ontology development (Gómez-Pérez 2004)

As shown in the Fig. 3, conceptualization activity in METHONTOLOGY consists of the set of tasks for organizing knowledge. Each task creates a special ontology component (concepts,

relations, instances …) and the arrangement of tasks offers the order which components must be created in along the activity. Following the order of tasks in the model, ensures that the represented knowledge is complete and consistent. We perform first four tasks during conceptualization activity and leave describing details for the next step of our project.

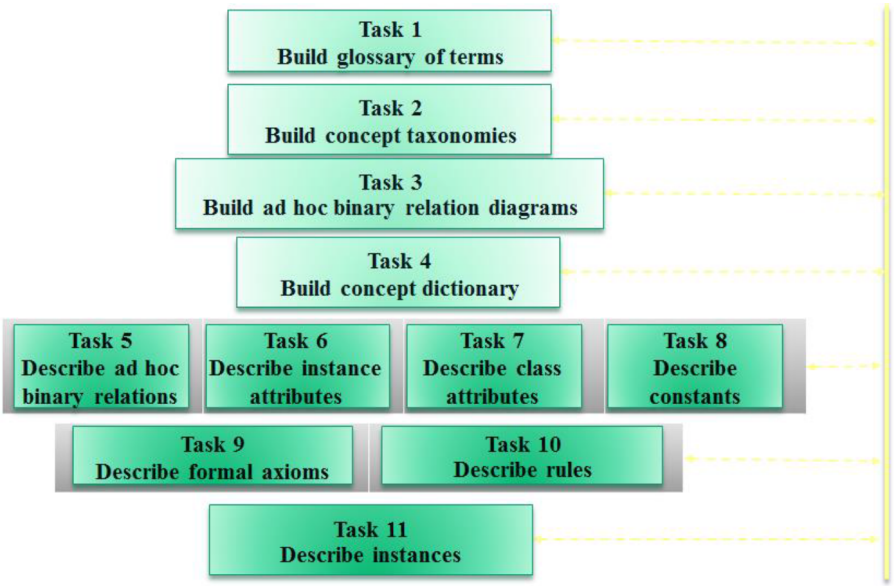

Figure 3: tasks of the conceptualization activity according to METHONTOLOGY (Gómez-Pérez 2004)

3.3. Conceptualization

As first task we built the glossary of terms based on the terms in our thesaurus. It should identify the set of terms which will be included on the ontology, their description in natural language, and their synonyms and acronyms. These terms were selected formerly by our domain experts from multiple resources such as classification schemes, existing thesauri, encyclopedias, dictionaries, periodical indexes, lexical indexes of textbooks, and collection of relevant documents using either inductive or deductive method.

In deductive method, the general framework of the subject is designed firstly. Then each of the topics is divided into subtopics and the process will continue to determine the most specific concepts. In the inductive method, a collection of relevant documents is selected and after indexing them, a set of concepts and terminologies is obtained. In fact, the inductive method is to form a hierarchy of concepts of a domain, while the deductive method tries to design a basic conceptual structure in one or more specialized fields and expand it by appending relevant terms to the structure.

Second task is to build concept taxonomies to classify concepts. Each preferred term designates a concept and concept taxonomies were formed based on the taxonomic relations in thesaurus. Afterward, we identified ad hoc relationships between concepts of the ontology and build ad hoc binary diagrams in task 3. Ad hoc relationships could be established between concepts of the same (or different) concept taxonomy. We mapped the relationships between terms in thesaurus into semantic relationships between corresponding concepts in ontology. BT/NT is converted to super/subclass-type relationship to form hierarchical structure of the ontology and other relationships labeled with their corresponding relationship type in thesaurus. Parts of concept taxonomies and hierarchical relationship between concepts is represented in Fig. 4 and 5.

The last task is to build concept dictionary. Concept dictionary mainly includes the concept instances for each concept, and their ad hoc relationships. We identify non-preferred terms as individuals and associated each of them with the concept which is designated by their corresponding preferred term. Translations and abbreviations are set out as concept attributes.

3.4. Implementation

After building conceptual model based on semantic information in thesauri, we implemented it in OWL, a web ontology language which is recommended by the World Wide Web Consortium (W3C). For this purpose, at first, domain experts contemplated concepts and made common concepts uniform, and then they examined hierarchical and associative relationships, and equalized them semi-automatically. Finally, the Basic Sciences Ontology was developed by converting conceptual model into OWL. Fig. 6 represents part of this ontology in Protégé.

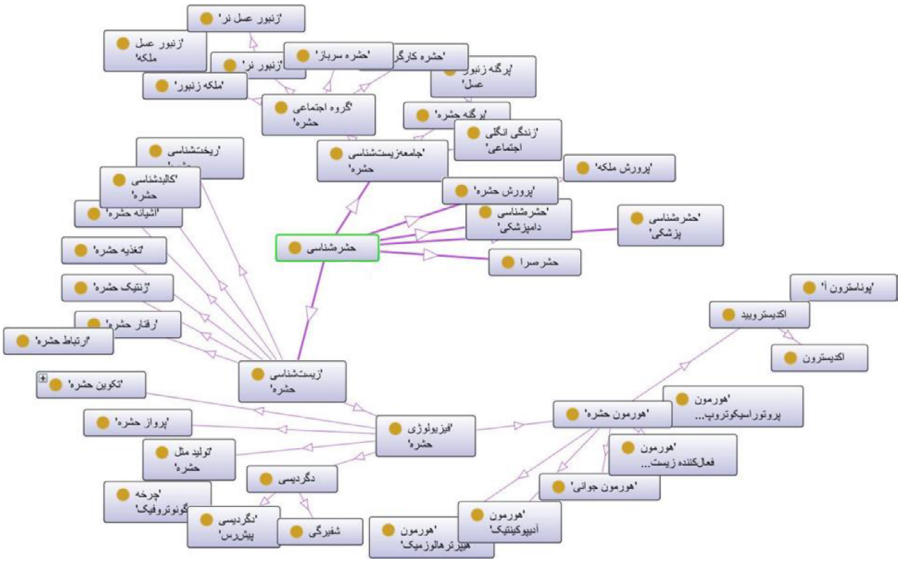

Figure 4: Part of concept taxonomy

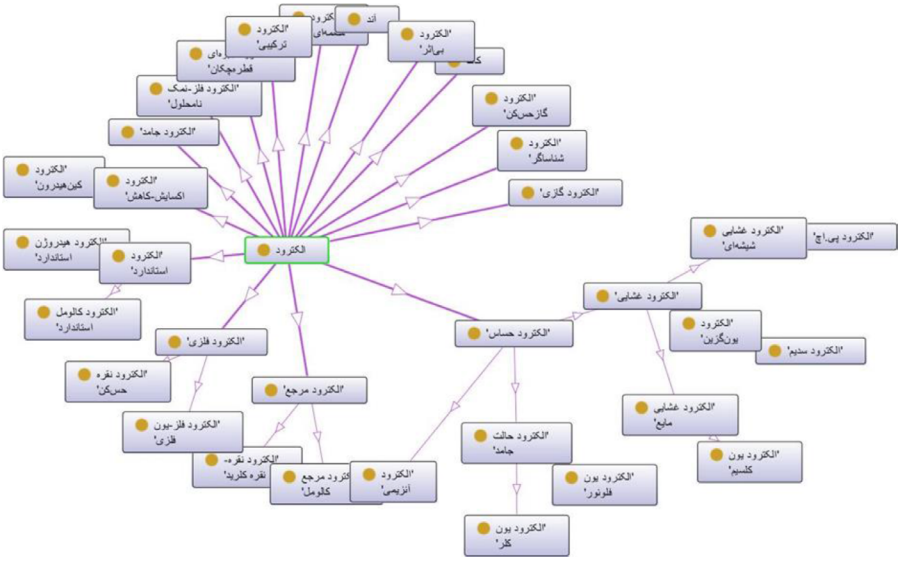

Figure 5: Part of concept taxonomy

4. Ontology refinement

The difference between the applications of thesaurus and ontology and the ambiguity in existing relationships in thesaurus, make the refinement process necessary. The hierarchical relationship in thesaurus may be one of the three types: generic, hierarchical whole-part, or instance relationship. However, in practice few thesauri make the distinction between them (ISO25964-2 2011) and therefore, this kind of hierarchical relationship has insufficient precision

for ontologies. Likewise, the associative relationship is very ambiguous. It is used in many different situations and link any two related terms with non-hierarchical relationship. Thus, its semantic is unspecified and cannot be used for reasoning.

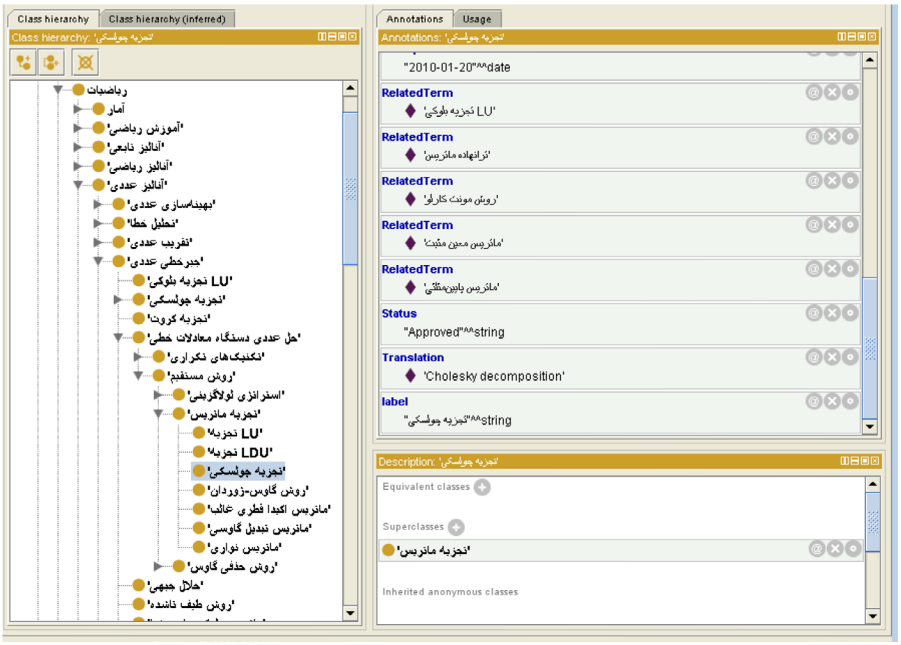

Figure 6: Part of Developed ontology

As a result, the relationships of developed ontology needed to be refined and converted to more precise ones. Our approach of refinement was similar to what proposed in (Soergel 2004). Our experts tried to extract semantic relationships between concepts and make the relationships more meaningful and specific. We also benefit from the concept relationship types, shown in Tab.1, in the first stage of refining the ontology. Hierarchical relationships in thesaurus are usually transformed into one of the concept relationships in first two rows and associative relationships are often converted to one of the relationships in the last row.

5. Conclusion

In this paper, we use thesauri previously developed at IRANDOC as resources to construct basic sciences ontology. At first we synchronized and integrated the thesauri semi-automatically and then transformed the produced macro thesaurus from ISO 5964 into ISO 25964. We use the methodology called METHONTOLOGY for designing the ontology. In this methodology the main activity is conceptualization. We used the conceptual model of our thesauri for this activity and build the ontology conceptual model based on it. At last, ontology of basic sciences generated by formalizing and implementing the model in OWL.

The next step is to refine the relationships to more specific semantic relations. Our domain experts tried to refine some relationships manually based on the Soergel approach (Soergel 2004). But we decide to design an appropriate method for refining the ontology semi-automatically. Also we need to add more details to our ontology and turn it into heavyweight ontology to get more advantage from it in formal reasoning.

X, Y are concepts
Isa
X <*includesSpecific*> Y / Y <*isa*> X
X <*inheritsTo*> Y / Y <*inheritsFrom*> X
Holonymy/meronymy (the generic whole-part relationship)
X <*containsSubstance*> Y / Y <*substanceContainedIn*> X
X <*hasIngredient*> Y / Y <*ingredientOf*> X
X <*madeFrom*> Y / Y <*usedToMake*> X
X <*yieldsPortion*> Y / Y <*portionOf*> X
X <*spatiallyIncludes*> Y / Y <*spatiallyIncludedIn*> X
X <*hasComponent*> Y / Y <*componentOf*> X
X <*includesSubprocess*> Y / Y <*subprocessOf*> X
X <*hasMember*> Y / Y <*memberOf*> X
Further relationship examples
X <*causes*> Y / Y <*causedBy*> X
X <*instrumentFor*> Y / Y <*performedByInstrument*> X
X <*processFor*> Y / Y <*usesProcess*> X
X <*beneficialFor*> Y / Y <*benefitsFrom*> X
X <*treatmentFor*> Y / Y <*treatedWith*> X
X <*harmfulFor*> Y / Y <*harmedBy*> X
X <*hasPest*> Y / Y <*afflicts*> X
X <*growsIn*> Y / Y <*growthEnvironmentFor*> X
X <*hasProperty*> Y / Y <*propertyOf*> X
X <*hasSymptom*> Y / Y <*indicates*> X
X <*similarTo*> Y / Y <*similarTo* > X
X <*oppositeTo* > Y / Y < *oppositeTo* > X
X <*hasPhase*> Y / Y <*phaseOf*> X
X <*growsIn*> Y / Y <*EnviromentForGrowing*> X
X <*ingests*> Y / Y < *ingestedBy*> X

Table 1: Concept relationships: Examples (Soergel 2004)

6. Acknowledgements

The work described in this paper has been done as a research project in Iranian Research Institute for Information Science and Technology (IRANDOC).

7. References

Daconta Michael C., Obrst, Leo J., & Smith, Kevin T. (2003). *The Semantic Web: A Guide to the Future of XML, Web Services, and Knowledge Management*. Wiley. com.

Gómez-Pérez Asunción, Fernández-López, Mariano, & Corcho Oscar (2004). *Ontological Engineering* (2nd ed.). Springer-Verlag, London.

Huang, Jin-Xia; Shin, Ji-Ae; & Choi, Key-Sun (2007). Enriching Core Ontology with Domain Thesaurus through Concept and Relation Classification*. Proc. OntoLex, ISWC.*

ISO 25964-1:2011 (2011). Information and documentation -- Thesauri and interoperability with other vocabularies -- Part 1: Thesauri for information retrieval. *International Organization for Standardization*, International Standard ISO 25964-1, Aug. 2011.

ISO 25964-2 (2011). Information and documentation - thesauri and interoperability with other vocabularies - Part 2: Interoperability with other vocabularies. *International Organization for Standardization*, Draft to be circulated for public comment late 2011.

Kawtrakul, Asanee, Imsombut, Aurawan, Thunkijjanukit, Aree, Soergel, Dagobert, Liang, Anita, Sini, Margherita, Johannsen, Gudrun & Keizer, Johannes (2005). Automatic term relationship cleaning and refinement for AGROVOC. In *Workshop on The Sixth Agricultural Ontology Service,* 247-260.

Khosravi, F. and Vazifedoost, A.(2008) Creating a Persian Ontology through Thesaurus Reengineering for Organizing the Digital Library of the National Library of Iran, *Faslname Ketab*,70, 19-36.

Soergel, Dagobert; Lauser Boris; Liang, Anita; Fisseha, Frehiwot; Keizer, Johannes; & Katz, Stephen (2004) Reengineering Thesauri for New Applications: the AGROVOC Example. *Journal of Digital Information*, vol. 4, no.4.

Sowa, John F. (2010). The role of logic and ontology in language and reasoning. *In Theory and Applications of Ontology: Philosophical Perspectives*, Springer Netherlands, 231-263.

Villazón-Terrazas, B., Suárez-Figueroa, M. C., & Gómez-Pérez, A. (2010). A pattern-based method for re-engineering non-ontological resources into ontologies. *International Journal on Semantic Web and Information Systems*, 6(4), 27-63.

Weinbrenner, S. and Engler, J. (2011) SCY Ontology and metadata scheme, DIV.2.2011, SCY consortium.