Developing & Using Ontologies for the Semantic Web: An Inspirational and a Synthetic Approach to Ontological Engineering & Utilizing WordNet in Information Retrieval

Tseliso Molukanele
tmolukan@cs.uct.ac.za

Supervised by:
Associate Professor Sonia Berman
sonia@cs.uct.ac.za

Rapelang Rabana
rrabana@cs.uct.ac.za

Department of Computer Science
University of Cape Town
Private Bag
Rondebosch 7701

1 Abstract

Ontological engineering is explored via the inspirational approach of creating ontologies through one's own imagination and the synthetic approach of developing ontologies through the synthesis of existing ones. The inspirational approach is implemented in two toolkits, one with minimal constructs and guidance and the other with constructs such a functional contexts and granularity. The synthetic approach is implemented in a toolkit that makes use of WordNet to generate semantically similar concepts across source ontologies, so as to support semi-automated merging. Lastly, an ontology-based information retrieval system is developed as a characteristic example of the use of ontologies in common applications.

2 Introduction

Ontologies hold more than just the future of the semantic web, but represent a significant effort by science to utilize mankind's most valuable resource, knowledge. The creation and use of ontologies encourage ways of acquiring and consolidating information into a form required by end-users.

An ontology can be termed to be 'an explicit formal specification of how to represent the objects, concepts and other entities assumed to exist in some area of interest and the relationships that hold among them' [DLI98]. By being formal specifications, ontologies reduce the need provide a means of sharing knowledge efficiently without ambiguities, among humans as well as computers. This is a critical feature as both humans and computers must be able to process information correctly by being able to gather the context and semantic meaning of information from an ontology.

This first part of this paper is centred on building tools that can be used by persons without knowledge representation expertise to develop ontologies. Two complementary approaches are explored here; the inspirational and the synthetic approach [Gru02]. The Inspirational approach is taken by a user who attempts to capture knowledge of a particular domain of interest by engineering an ontology from their individual imagination, creativity and personal insights. In contrast the Synthetic approach is taken by the user, who has identified existing ontologies from a particular domain and desires to amalgamate various ontologies into one unified ontology with all the relevant information identified. This approach involves ontology mapping and merging.

The latter part of the paper describes a simple ontology-based information retrieval system. The system, named OntSearch utilizes the WordNet ontology. This application demonstrates a possible direct use of ontologies, among many others that are not discussed.

3 Background

3.1 OWL

OWL (Web Ontology Language) is an ontology specification format that arose out of the need to make the web a better medium for information and knowledge.

The roots of OWL lie in HTML (Hyper-Text Mark-up Language, that was further improved by XML (eXtensible Mark-up Langauage) which separated the context of the information for the actual data represented. RDF (Resource Description Framework) was later introduced to support the semantic processing of information by machines in addition to humans. OWL is the direct successor of RDF that enables robust ontologies to be built and distributed across the web and allows data and exact meaning to be shared.

OWL relates concepts through generalization/specialization and aggregation relationships. These taxonomies are acyclic graphs that hold concepts as nodes and edges as the relationships that exist between the concepts. OWL is the standard used for all ontology applications discussed.

3.2 Functional Contexts and Granularity

During the design of Geographical Information Systems (GIS), [Sem04] errors relating to semantic heterogeneity, inappropriate data collection and misleading interpretations were encountered. A methodology that accommodated granularity and context was proposed for reducing such errors.

'Granularity is defined as the level of influence or impact at which objects interact. Context is defined as the operational subsystem in which the object interaction is being studied' [Sem04], it is seen as expressing the builders point of view and primary concern for building the ontology. To enforce these two concepts in the ontology engineering process, it is required that initially the system to be modelled is decomposed into functional context subsystems which are subsystems with specific functions that are independent of each other in terms of definition. Following this decomposition, a further decomposition is required to disintegrate each functional context into niches where each niche has a level of influence, otherwise known as granularity. The primary entity in the functional context has a granular level of one and other levels are given positive or negative values to reflect lesser or more influence.

The extra constructs are helpful in suggesting ways of combining different ontologies that have been built by users with different points of view. Such ontologies would have different functional contexts and may have some concepts the same but defined differently. For the concepts occuring in both ontologies it can be decided which will be carried over by comparing the impact of those concepts. Alternatively it may be decided to keep information from both concepts but the impact will aid in pointing out what is important.

This methodology also places constraints on relationships like generalization or aggregation that are intended to help in detecting and remedying errors and omissions that may have occurred during the modelling process. This methodology is further explored under the Extension Toolkit.

3.3 Ontology Mapping

Establishing semantic mappings between ontologies is a keen area of research. The GLUE system [Doa02] applies machine learning techniques to semi-automatically create semantic mappings between ontologies. In finding mappings between ontologies, the system must establish similarity between concepts, as is required by the synthetic approach to ontological engineering. The GLUE system favours the distribution-based Jaccard coefficient of P(A ∩ B)/ P(A U B) for any two concepts A and B. The coefficient is zero if the A and B are disjoint, or not related and one when they are the same concept. Distribution measures like the Jaccard coefficient model each concept as a 'set of instances taken from a finite universe of instances'.

Solving for the Jaccard coefficient requires the determination of the joint distributions of A and B via a sampling process that assumes that the sample used is a 'representative sample of the instance universe covered by the taxonomy'. The Jaccard coefficient is effectively an estimation and thus GLUE system avoids having to determine a specific similarity value directly.

A simpler and possibly more accurate approach, that does not use distribution methods, is employed in the Synthetic Toolkit. The approach used is adapted from the techniques employed in computational linguistics for ontology-based information retrieval systems, which are discussed in the following section, 3.4.

3.4 Ontology-Based Information Retrieval

Ontology-Based Search Engines move away from traditional keyword based models to concept-based models. The typical ontology used to support such a model is WordNet [Mil92]. WordNet represents an extensive electronic English thesaurus where terms with the same meaning are grouped together to form a synset (set of synonyms). Liu *et al.* [Liu04] make use of WordNet as well as several other techniques to develop an ontology-based information retrieval system.

The approach taken is to identify phrases including proper names of people or entities (using an entity recognizer), dictionary phrases (using WordNet), simple phrases and complex phrases. The fundamental assumption made is that phrase similarity is more important than term similarity.

WordNet is used to generate synonyms and related concepts from query terms with which to match conceptually related documents. Word sense disambiguation is performed using adjacent words in the query and the basic constructs of WordNet, namely, synsets of the query terms, definitions of the synsets, the hypernyms (sub-synsets) and the definitions of the hypernyms.

A simplified version of this approach is discussed further in under the ontology-based information retrieval system, OntSearch.

An adaptation of this approach of determining semantic similarity between queries and documents is also used for identifying semantically similar concepts across different ontologies and further explored under the discussions of the Synthetic Toolkit.

3.5 Ontology Merging

'Merge is the process of building an ontology in one subject reusing two or more different ontologies on that subject' [Pin01]. In a merge, as is done by the Synthetic Toolkit, the source ontologies are synthesized into one ontology.

The merging of ontologies is still largely a manual process, but many tools have become available, such as PROMPT [Noy00].

4 Supporting Tools

All Toolkits are implemented using the Jena 2 API [McB02] that provides a programmatic environment for RDF and OWL [McG04]. This API is generally used for building Semantic Web applications. It effectively parses ontologies written in OWL and builds an ontology model that provides support for the creation and retrieval of typical objects found in ontologies.

Another supporting tool used was HyperGraph [Hyp], a visualization tool that uses hyperbolic geometry to display hierarchical structures. It is a highly useful tool where large data volumes must be displayed as it employs a fisheye-like distortion that offers 'local detail and global context'.

5 Inspirational Approach

5.1 Methodologies

In the investigation of the inspirational approach to ontological engineering, two methodologies that support the user's individual creativity in developing ontologies were considered. The first methodology allows for ontology development where classes, individuals and properties can be added at will anywhere on the ontology so long as each concept has at least one is-A or has-A relationship to some other concept. The implementation of this basic, unrestricted methodology was fulfilled in the Inspirational Toolkit.

The second methodology employs several extensions over and above the functionality of the Inspirational Toolkit and enforces a more structured approach to ontological engineering. The extensions include the enforcement of functional context and granularity in the ontology modelling process. Granularity can be used as a means to govern the macro-structure of large object-oriented applications. It is suited to the organization of small computational units but is too finely grained for the organization of larger applications. This methodology is carried out in the Extension Toolkit. Note that the extensions mentioned are not compliant with OWL.

5.2 Inspirational Toolkit

The first phase of the implementation of the Inspirational approach brought the Inspirational Toolkit to bear.

The user can make use of two relationships in their ontology development. The generalization/specialization relationship allows for 'is-a' relationships between a super class and its sub-class. An 'is-a' relationship can also exist between a class and instances of the class. The second relationship of aggregation, allows for a 'has-a' relationship between a class and its properties. Figure 1 illustrates the graphical user interface of the Toolkit and some of the functions available.

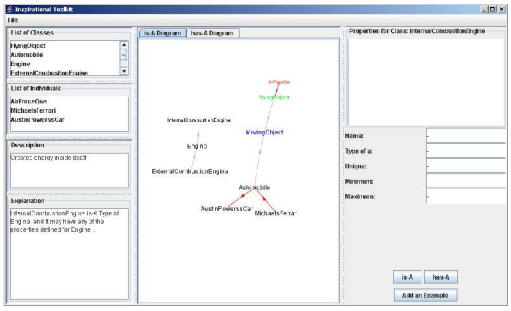

Figure 1: Inspirational Toolkit Interface

The nodes created by the user on the interface, are used to automatically generate the OWL syntax that specifies the ontology model drawn by the user. An interactive graph model enables the user to change the current area of focus in the ontology, simply by selecting a node in the area of interest at which point the area is instantly magnified.

5.3 Extension Toolkit

The second phase of implementation of the inspirational approache focused on the Extension Toolkit and extends the basic functionality of the Inspirational Toolkit.

The extensions of functional contexts and granularity are intended to aid in disambiguation. For each concept included in the ontology, the functional context and granularity is identified before the concept can be included in the ontology. Hence, functional contexts and niches must be inserted before nodes can be inserted. The functional context can also be considered to be the knowledge domain and the niche as the specific subset of the domain, which the concept of has the most influence in.

Each functional context has a niche with a default impact and the nodes of this niche have a default impact. Other niches in the functional context are given higher or lower impact; this being a number associated with each niche. Figure 2 illustrates functional contexts and granularity. There are restrictions that govern how concepts are related in order to keep relations semantically correct. Note also that multiple tabs enable the user to access different views of the ontology.

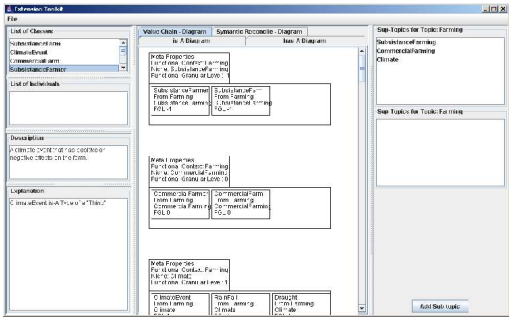

Figure 2: Illustration of Functional Context and Granularity

5.4 Results

A group of six users were required to develop ontologies using the Inspirational Toolkit and the Extension Toolkit. The users felt that more constructs, as found in the Extension Toolkit were helpful gave much required guidance. The additional constructs of the Extension Toolkit resulted in more comprehensive knowledge capturing that may not have otherwise been captured by the user.

This result indicates that novice ontology developers may require considerable support in the engineering process through the use of constructs, so as to ensure that as much information is extracted from their minds and captured.

5.5 Conclusions

The toolkits and testing conducted prove that it is possible to implement a toolkit for the creation of ontologies by users without knowledge representation expertise. The implication of this is that knowledge can be captured by persons with domain knowledge and ultimately, the inspirational approach is within the reach of the average user.

Furthermore, even though users felt the Extension Toolkit required more time to learn, users found it considerably better to build ontologies using the Extension Toolkit over the Inspirational Toolkit. The additional constructs employed by the Extension Toolkit provided users with a greater understanding of the intended structure of the ontology under construction. Users understand how this would be helpful in combining ontologies with same concepts but built with different functional context that represent the builders point of view.

6 Synthetic Approach

6.1 Merging Methodology

The merging methodology to synthesize ontologies is adapted from that of the toolkit, PROMPT [Noy00]. The Synthetic Toolkit takes two ontologies as input and guides the user in the creation of one, merged ontology as its output.

Guidance is provided through a list of suggestions of classes to merge. The user can choose to execute one of the suggestions or perform other operations such a direct merger or a basic function like adding another class. Figure 3 demonstrates this methodology.

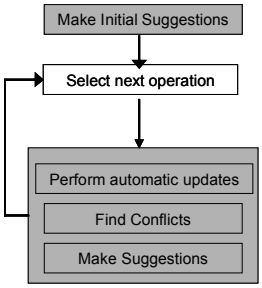

Figure 3: Synthetic Toolkit Methodology

The methodology also makes use of the notion of a 'Preferred Ontology'. Between the two source ontologies, this ontology, as the name implies, is of greater importance. It has often been acknowledged that source ontologies are not of equal relevance to a user [Noy00] and there is often the one source ontology that the user chooses to improve upon by drawing from the information of others.

The Synthetic Toolkit requires that the preferred ontology is designated. The implication of this designation is that merging occurs in one direction from the other ontologies to the preferred ontology, such that input from other source ontologies is amalgamated with the information already found in the preferred ontology. Figure 4 illustrates the interface of the Synthetic Toolkit and how two ontologies can be displayed simultaneously. On the left is the Preferred Ontology and in the centre is the other ontology.

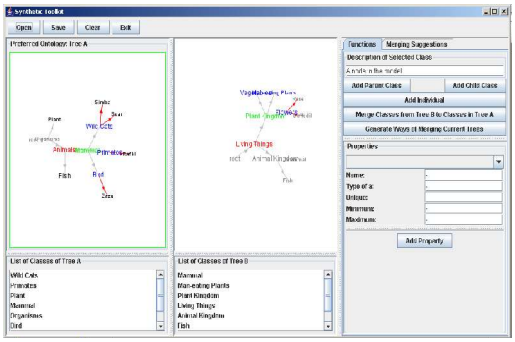

Figure 4: Synthetic Toolkit Interface

6.2 Implementation

The area of focus in the implementation of this toolkit is its ability to identify semantic similarities across ontologies and suggest them to the user.

The process of matching classes across the source ontologies is divided into a few steps, decreasing in the expected accuracy of the matches generated. Suppose any class in the Preferred Ontology, Ontology A, is c_1 and any class from the other ontology, Ontology B is c_2.

- Step One: if any class c_2 from Ontology B, or the synonyms of any c_2 match c_1 or a sub-term of c_1, then suggest a merger between c_1 and c_2.
- Step Two: if any class c_2 or a sub-term of c_2 from Ontology B matches c_1 or any synonyms of c_1 then suggest a merger between c_1 and c_1.
- Step Three: if any class, c_2, from Ontology B or the synonyms of any c_2, is found in the resource description of any class c_1, suggest a merger between c_1 and c_2.
- Step Four: if the resource descriptions of any classes c_1 and c_2 have more than two content words in common, suggest a merger between c_1 and c_2.

Using the synsets and description of classes, related concepts are discovered. All steps are executed regardless regardless of whether a previous step returned a suggestion or not. This ensures that options are available to the user to select from. Figure 4 illustrates the presentation of merging suggestions on the right panel.

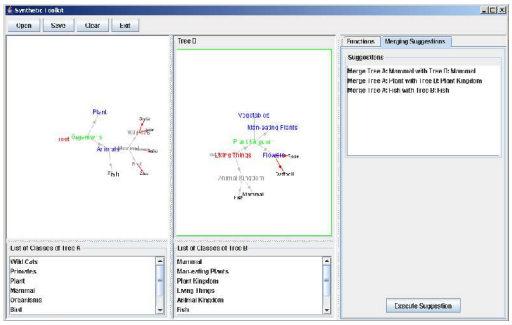

Figure 5: Synthetic Toolkit Merging Suggestions

6.3 Results

User testing was conducted for six users of various disciplines in the final year of undergraduate study without any prior exposure to knowledge representation. Users had to merge example ontologies as well as ontologies they had to create. Unmet needs or desires of users were the direct manipulation of ontologies such as dragging, and short-cut functionality using the right button of the mouse. Though the dynamic structure of HyperGraph was very appealing, it still struggled with clutter when numerous nodes are displayed.

Nonetheless, the merging methodology and process was readily accepted and understood making it plausible for general use by non knowledge representation experts. Only 38% of merging suggestions generated by the Toolkit were actually used by the user suggesting that a more stringent algorithm is required. 78% of the total mergers performed by the user (manually and automatically from suggestions) were automatic mergers from the suggestions generated by the Toolkit. This indicates that the Toolkit did not overlook key areas of semantic similarity across the source ontologies and was highly successful to this regard.

6.4 Conclusions

The question under investigation is the Toolkit's capacity to aid ontology engineering for the average user who does not have knowledge representation expertise. By the fact that the six users of the study were able to merge example and self-built ontologies with another ontology of a similar theme is key indicator that the Synthetic Toolkit is highly accessible to the average user.

Furthermore, the Toolkit demonstrated a strong capacity to identify similarities across ontologies. Though there is still significant room for improvement the Synthetic Toolkit has succeeded in its primary goal and has brought the

synthetic approach of ontological engineering within the understanding of non-experts.

7 ONTSEARCH

7.1 Overview

OntSearch in an ontology-based information retrieval system designed specifically to be able to compare search results with and without an ontology. It a classic example of how ontologies can be used. The system conssequently produces two lists of results as illustrated in Figure 5.

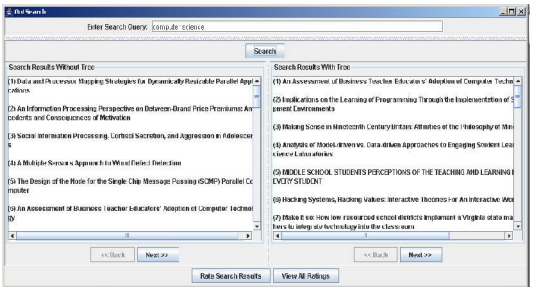

Figure 5: OntSearch Interface

A collection of documents were processed and an inverted file structure created where each term from the documents is associated with a list of weighted document identifiers. A typical term frequency * inverse document frequency are used to ascertain the weighting of each document. Weighting are then sorted to provide a ranked order [Arasu]. These aspects are not discussed further, but rather the focus on how WordNet was utilized in the application.

7.2 Implementation

OntSearch makes use of WordNet and word sense disambiguation techniques to generate terms that are semantically similar to the query terms entered by the user. By expanding the query to cover all strongly semantically related concepts, other relevant documents are discovered

WordNet [Mil92] is an English word database where nouns, verbs, adjectives and adverbs are arranged into sets of synonyms which each represent a single sense. WordNet is used to help clarify the sense of each query term and thus be able to generate the correct synonyms. The process of determining the correct sense of a word is referred to as word sense disambiguation. Performing word sense disambiguation also involves the use of adjacent terms in the query statement.

The process of disambiguation is divided into a few steps [Liu04], decreasing in the expected accuracy of related concepts generated. Suppose the term for which sense must be determined is t_1 and the other term from the query that is being used to help ascertain the sense of t_1 is called t_2.

- Step One: if t_2 or a synonym of t_2 is found in the definition of the synset of t_1, say S, S is determined to be the sense of t_1.
- Step Two: if the definition of each synset of t_1 is compared against the definition of each synset of t_2. The combination of the synset of t_1 and the synset of t_2 that have the highest positive number (at least 2) of content words in common yields the sense for t_1.
- Step Three: if t_2 or one of its synonyms appears in the definition of a synset S, containing a hyponym (sub-class) of t_1, then the sense of t_1 is determined to be the synset of S_1, which contains t_1 and has the descendant S.
- Step Four: Let t_1 be contained by a synset S_1. S_1 has a hyponym (sub-class) synset U that contains a term h. If h appears in the definition of a synset S_2 containing t_2 , then the sense of t_1 is determined to be S_1.

Using the synsets, definitions and hyponyms of terms, concepts related to the query terms are discovered and relevant documents containing the related concepts.

7.3 Results

The same participants involved in the evaluation of the Synthetic Toolkit, rated the searching results produced by OntSearch. The ratings compared the quality of results produced with and without the use of WordNet.

The ratings results did not convincingly substantiate the belief that ontology-based information retrieval systems provide superior results to traditional search engines. This indicates that a review of algorithms is required to arrive at more conclusive answers.

8 GENERAL CONCLUSIONS

Ontological engineering characterizes a significant growth area, particularly with regards to the use of ontologies by non-knowledge representation experts. Both the inspirational and the synthetic approach to building ontologies represent viable methods that are within the grasp of novice users.

9 FUTURE WORK

9.1 Formal Evaluation of Toolkits

A formal evaluation of the Toolkits must be done to compare them against existing ontology development applications. It is only after such formal evaluations that conclusive evidence as to the usability and quality of the Toolkits can be established. Such evaluations are critical if the Toolkits are to be improved upon substantially.

9.2 Further use of OWL constructs

OWL is a very expressive language and further constructs can be implemented in the Toolkits. Such constructs include the use of restrictions on the relationships between concepts. One such restriction on the has-A relationship expresses a concept definition based on a specific has-A relationship between concepts. For example the definition of “red car” is based on the restriction “car” has-a “color” and is expressed as a “red car” is any “car” that has-A “color” value of “red”. Without this restriction it is not possible to define a human being as having 2 walking legs without explicitly creating instances of legs and having a maximum value restriction on those.

9.3 Diagnostics

Performing diagnostics would go very far in improving the quality of ontologies that inexperienced users might produce [McG00]. Possible diagnostics tests include checking for completeness of the ontologies by ensuring

all descriptions and definitions are complete; a syntactic analysis to indicate an incidence of words or sub-strings, or possible acronym expansion; a taxonomic analysis indicating redundant super classes, sub classes or instances; and a semantic evaluation that identifies a property mismatch. More thoroughly completed ontologies also allow for more semantic similarities to be drawn across ontologies.

9.4 Automated Concept Search

At this point in the development of the Synthetic Toolkit, the user must specify the source from which concept matches will be generated. In future, such searches must be automated in the Toolkit is to significantly assist in the reuse of ontological data [Kal]. Such searches must explore the local ontology collection and ultimately crawl the web in search for relevant ontologies that might be useful to the user. The Toolkit would then dynamically make suggestions as the user is developing their ontologies.

10 References

[DLI98] DLI-UIUC Glossary. 1998. Retrieved from http://dli.grainger.uiuc.edu/glossary.htm

[Doa02] Doan, A., Madhavan, J., Domingos, P., and Halevy, A. 2002. Learning to map between ontologies on the semantic web. In *Proceedings of the 11th International Conference on World Wide Web* (Honolulu, Hawaii, USA, May 07 –11, 2002)

[Gru02] Gruninger, M. and Lee, J. 2002. Ontology Applications and Design, *Commun.* ACM 45, 2 (Feb – 2002), 39 – 41

[Hyp] HyperGraph. Retrieved Aug 2005 from: http://hypergraph.sourceforge.net/

[Kal] Kalyanpur, A., Hashmi, N., Golbeck., J., and Parsia., B. Lifecycle of a Causal Web Ontology Development Process. University of Maryland. Retrieved Aug 2005 from: http://www.mindswap.org/~aditkal/WWW04_COD.pdf

[Liu04] Liu, S., Liu, F., Yu, C., and Meng, W. 2004. An effective approach to document retrieval via utilizing WordNet and recognizing phrases. In *Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Sheffield, United Kingdom, July 25 - 29, 2004). SIGIR '04. ACM Press, New York, NY, 266-272. DOI= http://doi.acm.org/10.1145/1008992.1009039

[McB02] McBride, B. 2002. Jena: A Semantic Web Toolkit. *IEEE Internet Computing* 6, 6 (Nov. 2002), 55-59. DOI= http://dx.doi.org/10.1109/MIC.2002.1067737

[McG00] McGuinness, D. L., Fikes, R., Rice, J., and Wilder, S. 2000. The Chimaera Ontology Environment. In *Proceedings of the Seventeenth National Conference on Artificial intelligence and Twelfth Conference on innovative Applications of Artificial intelligence* (July 30 - August 03, 2000). AAAI Press / The MIT Press, 1123-1124.

[McG04] McGuinness, D.L., Harmelen, F. 2004. OWL Web Ontology Language Guide. Retrieved Aug 2005 from: http://www.w3.org/TR/owl-features/

[Mil92] Miller, G. A. 1992. WordNet: a lexical database for English. In *Proceedings of the Workshop on Speech and Natural Language* (Harriman, New York, February 23 - 26, 1992). Human Language Technology Conference. Association for Computational Linguistics, Morristown, NJ, 483-483.

[Noy00] Noy, N. F. and Musen, M. A. 2000. PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment. In *Proceedings of the Seventeenth National Conference on Artificial intelligence and Twelfth Conference on innovative Applications of Artificial intelligence* (July 30 - August 03, 2000). AAAI Press / The MIT Press, 450-455.

[Pin01] Pinto, H. S. and Martins, J. P. 2001. A methodology for ontology integration. In *Proceedings of the 1st international Conference on Knowledge Capture* (Victoria, British Columbia, Canada, October 22 - 23, 2001). K-CAP '01. ACM Press, New York, NY, 131-138. DOI= http://doi.acm.org/10.1145/500737.500759

[Sem04] Semwayo, D. and Berman, S. 2004. A Model and Methodology for GIS Design based on Granularity and Context. Department of Computer Science. University of Cape Town.