# SWETO: Large-Scale Semantic Web Test-bed

Boanerges Aleman-Meza, Chris Halaschek, Amit Sheth, I. Budak Arpinar, Gowtham Sannapareddy
*Large Scale Distributed Information Systems (LSDIS) Lab*
*Computer Science Department, University of Georgia*
*Athens, GA 30602-7404*
*{boanerg, ch, amit, budak}@cs.uga.edu, gowtham@uga.edu*

**Abstract.** The emergent Semantic Web community needs a common infrastructure for testing the scalability and quality of new techniques and software which use machine processable data. Since ontologies are a centerpiece of most approaches, we believe that for an accurate evaluation of tools for quality, scalability and performance, the research community needs a freely available ontology with a large description base. If the use of tools is to be for advanced semantic applications, such as those in business intelligence and national security, then instances in the knowledge base should be highly interconnected. Thus, we propose and describe a Semantic WEb Technology evaluation Ontology (SWETO) test-bed. In particular, we address the requirements of a test-bed to support research in semantic analytics, as well as the steps in its development, including, ontology creation, semi-automatic data extraction, and entity disambiguation.

## 1. Introduction

Considering that there are somewhere between 20 to 50 ontology tools alone [16, 17], the question arises: how do we test and compare them? Similarly, applications that utilize ontologies for inference, semantic integration, and semantic analytics, require a benchmark for quality, scalability and performance evaluations. Thus, the emergent Semantic Web community needs a common infrastructure for both testing and evaluations. In particular, we feel there is a need to have a large, high quality test ontology from which various ontology tools can assess and test their scalability and other properties.

Of particular interest is not just the schema of the ontology, but also the population (instances, assertions or description base) of the ontology. A highly populated ontology (ontology with instances or assertions) is critical for assessing effectiveness, and scalability of core semantic techniques such as semantic disambiguation, reasoning, and discovery techniques. Ontology population has been identified as a key enabler of practical semantic applications in industry; for example, Semagix[1] reports that its typical commercially developed ontologies have over one million objects [18]. So far, such ontologies have not been available to the research community.

Another important factor related to the population of the ontology is that it should be possible to capture instances that are highly connected (i.e., the knowledge base should be deep with many inter-entity relationships). This will allow for a more detailed analysis of current and future semantic tools and applications, especially those that exploit the way in which entities are related. This is exemplified in our SemDis[2] project, in which new complex semantic relationships can be queried and discovered through traversing sequence of links among the entities of interest. Clearly, an ontology and corresponding knowledge base of real-world scale are needed as a benchmark for evaluating and comparing such tools and techniques.

To this end, we propose a Semantic Web Technology evaluation Ontology (SWETO[3]), that captures real world knowledge with tens of classes populated with a growing set of relevant facts, currently at about one million instances. As part of the creation of SWETO, we have adopted the following iterative process that allows the periodic extension the ontology and its instances:

(i) Designing SWETO schema using an ontology design toolkit (detailed later),

(ii) Identifying knowledge sources that can be used to populate parts of SWETO without focusing on a specific domain,

(iii) Utilizing extractors (written by humans using a toolkit) to periodically and automatically extract parts of knowledge from various open and public sources,

(iv) Semi-automatically applying disambiguation techniques to extracted instances in the ontology (with limited human involvement) to eliminate redundancies and improve quality of the knowledge base,

[1] http://www.semagix.com
[2] http://lsdis.cs.uga.edu/Projects/SemDis/index.php
[3] http://lsdis.cs.uga.edu/proj/Sweto

(v) Providing capabilities for exporting SWETO and its instances from an internal representation to World Wide Web Consortium (W3C) standards, namely either OWL [13] or RDF [14]; thus allowing open use of SWETO.

The remaining sections of this paper are organized as follows: Section 2 details related work in this area; Section 3 describes the overall methodology of our approach for creation of SWETO; Section 4 presents the current results of our work; Section 5 provides conclusions and some future directions for SWETO.

## 2. Related Work

Due to the infancy of the Semantic Web, little research has been focused on the development of an evaluation benchmark or test-bed for it. One current and ongoing effort however is TAP [2], which provides a large knowledge base annotated using RDF and is described as a "... shallow but broad knowledge base ..." [2]. Our work differs in that we provide a smaller schema, but with a much larger number of instances that are highly interconnected. Additionally, we provide the option to serialize the ontology using OWL, allowing for more constraints and expressiveness at the schema level.

## 3. Methodology

SWETO is an ontology that incorporates instances extracted from heterogeneous sources. Automatic population is created by extractors (detailed in Section 3.3).

### 3.1. Ontology Creation

The test-bed has been created in a bottom-up fashion where the data sources dictate the classes and relationships defined in the ontology, similar in spirit to the concept of emergent semantics [1, 15].

To illustrate with an example, consider the listing of "people" in a computer science department. Typically, they would be listed separately as Faculty, Students and Staff. In such cases we create appropriate classes in the ontology and populate them with instances.

In SWETO, the ontology was created using Semagix Freedom, a commercial product which evolved from the LSDIS lab's past research in semantic interoperability and the SCORE technology [6]. The Freedom toolkit allows for the creation of an ontology, in which a user can define classes and the relationships that it is involved in using a graphical environment. Thus, the user is relieved of the burden of serializing the ontology to the OWL syntax.

### 3.2. Selection of Data Sources

Creation of a solid test-bed requires meticulous selection of data sources. We focused our selection of data sources by considering the following factors:

(i) Selecting sources which were highly reliable Web sites that provide entities in a semi-structured format, unstructured data with parse-able structures (e.g., html pages with tables), or dynamic web sites with database back-ends. In addition, the Freedom toolkit has useful capabilities for focused crawling by exploiting the structure of Web pages and directories.

(ii) We carefully considered the types and quantity of relationships available in a data source. Therefore we preferred sources in which instances were interconnected.

(iii) We considered sources whose entities would have rich metadata. For example, for a 'Person' entity, the data source also provides attributes such as gender, address, place of birth, etc.

(iv) Public and open sources were preferred, such as government Web sites, academic sources, etc. because of our desire to make SWETO openly available.

### 3.3. Knowledge Extraction

In SWETO, all knowledge (or facts that populate the ontology) is extracted using Semagix Freedom software. Essentially, extractors are created within the Freedom environment, in which regular expressions are written to extract text from standard html, semi-structured (XML), and database-driven Web pages. As the Web pages are 'scraped' and analyzed (e.g., for name spotting [19]) by the Freedom extractors, the extracted entities are stored in the appropriate classes in the ontology. Additionally, provenance information, including source, time and date of extraction, etc., is maintained for all extracted data. We later utilize Freedom's API for exporting both the ontology and its instances in either RDF [14] or OWL [13] syntax. For keeping the knowledge base up to date, the extractors can be scheduled to rerun at user specified time and date intervals.

Automatic data extraction and insertion into a knowledge base also raise issues related to the highly researched area of entity disambiguation [7, 8, 9, 10]. In SWETO, we have focused greatly on this aspect of ontology population. Using Freedom, entity instances can be disambiguated using syntactic matches and similarities (aliases), customizable ranking rules, and relationship similarities among entities. Freedom is thus able to automatically disambiguate entities as they are extracted [6].

Furthermore, if Freedom detects ambiguity among new entities and those within the knowledge base, yet it is unable to disambiguate them within a preset degree of certainty, the entities are flagged for manual

disambiguation with some system help on possible matches.

Lastly, there a special cases in which neither the software, nor humans can directly determine if two entities are the same. For example, consider two persons named 'John Smith'. Without metadata attributes, neither the system nor humans can determine what to do by only looking at the entity name. This is a future research direction we wish to follow in which semantic similarity will be used to state with some degree of certainty that these two persons (i.e. 'John Smith'), are in fact the same person. For now, we remove these types of entities from the knowledge base in order to maintain both cleanliness and consistency.

## 4. Results

Our aim of achieving a test-bed of over 1 million instances is near completion. The current population includes over 800,000 entities and over 1,500,000 explicit relationships among them. Here we provide initial statistics that illustrate the size in terms of entities and relationships connecting them.

Table 1 summarizes a subset of the classes of the ontology that are representative of the majority of instances currently in SWETO ontology.

**Table 1. SWETO test-bed ontology initial metrics**

| Subset of classes in the ontology | # Instances |
|---|---|
| Cities, countries, and states | 2,902 |
| Airports | 1,515 |
| Companies, and banks | 30,948 |
| Terrorist attacks, and organizations | 1,511 |
| Persons and researchers | 307,417 |
| Scientific publications | 463,270 |
| Journals, conferences, and books | 4,256 |
| TOTAL (as of January 2004) | 811,819 |

What makes this work more valuable is in respect to how inter-connected the instances are (this currently is not available in a taxonomy and in most current ontologies that are freely available). As mentioned earlier interconnectedness becomes critical in semantics analytics applications (such as [3]). Table 2 summarizes a subset of the relationships connecting instances in the ontology. Note that some relationships apply to a variety of instances, such as the "located in" relation.

**Table 2. SWETO statistics on relationships**

| Subset of relationships | # Explicit relations |
|---|---|
| located in | 30,809 |
| responsible for (event) | 1,425 |
| Listed author in | 1,045,719 |
| (paper) published in | 467,367 |

As mentioned in Section 3.3, a variety of techniques for entity disambiguation has been employed in order improve the knowledge base. The frequency and type of disambiguation method is presented below in Table 3.

**Table 3. SWETO statistics on disambiguation**

| Disambiguation type | # Times used |
|---|---|
| Automatic (Freedom) | 248,151 |
| Manual | 210 |
| Unresolved (Removed) | 591 |

In addition, SWETO details can be found at its homepage (http://lsdis.cs.uga.edu/proj/Sweto/). There, we provide a graphical user interface for browsing of SWETO ontology (through the use of Touchgraph[4]) as illustrated in Figure 1, the latest version of the knowledge base (instances), our own native API for easy use (alternately tools such as Jena [12] could be used), and a detailed description of the data sources.

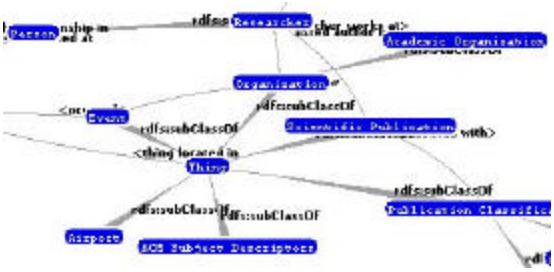


**Figure 1 Subset of SWETO schema visualization**

## 5. Conclusions and Future Work

In this paper, we presented SWETO, a test-bed for testing effectiveness and scalability of current and future semantic Web applications and techniques.

As mentioned earlier, the ontology-driven Semagix Freedom toolkit has been used for graphical creation of the ontology schema, as well as for automated population of the ontology with extractors. Additionally, Freedom was used for entity disambiguation. Lastly, we provided a summary of the statistics that make up for the current population of over 800,000 entities and over 1,500,000 explicit relationships among them.

Our research with SWETO test-bed has primarily been driven by the discovery of semantic associations [4] and their ranking [5]. Therefore, we aim for continuing the population of the ontology by further inter-connecting instances in order to provide a diverse test-bed for testing semantics analytics research ideas.

[4] http://www.touchgraph.com/

As mentioned in Section 3.3, we also wish to further investigate the use of semantic similarity for entity disambiguation.

## 6. Acknowledgements

SWETO test-bed is an effort that incorporated ideas and suggestions from different people in the LSDIS lab to whom we are thankful. Additionally, we would like to acknowledge our UMBC collaborators, especially Tim Finin, Anupam Joshi, and Li Ding who we are jointly working with on the SemDis project.

We also thank Semagix, Inc. for providing its Freedom product. In particular, we would like to especially thank David Avant and Yashodhan Warke for their insightful comments and reviews.

This work is funded in part by National Science Foundation (NSF) Awards 0219649 ("Semantic Association Identification and Knowledge Discovery for National Security Applications") and IIS-0325464 ("SemDis: Discovering Complex Relationships in Semantic Web"). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

## References

[1] S. Staab: Emergent Semantics. IEEE Intelligent Systems 17(1), 2002. pp. 78-86

[2] R. Guha and R. McCool, "Tap: A Semantic Web Test-Bed", Journal of Web Semantics, 1(1), Dec. 2003, pp. 81-87

[3] A. Sheth, B. Aleman-Meza, I. B. Arpinar, C. Halaschek, C. Ramakrishnan, C. Bertram, Y. Warke, D. Avant, F. S. Arpinar, K. Anyanwu, and K. Kochut, Semantic Association Identification and Knowledge Discovery for National Security Applications, Special Issue of Journal of Database Management on Database Technology for Enhancing National Security, Eds: L. Zhou and W. Kim, 2004 (Accepted).

[4] K. Anyanwu, and A. Sheth. r-Queries: Enabling Querying for Semantic Associations on the Semantic Web. Twelfth International World Wide Web Conference, Budapest, Hungary. May 20-24, 2003; pp. 690-699

[5] B. Aleman-Meza, C Halaschek, I. B. Arpinar, and A. Sheth, Context-Aware Semantic Association Ranking, First International Workshop on Semantic Web and Databases, Berlin, Germany, September 7-8, 2003; pp. 33-50

[6] A. Sheth, C. Bertram, D. Avant, B. Hammond, K. Kochut, and Y. Warke. (2002). Managing semantic content for the Web. IEEE Internet Computing, 6(4), 2002. pp 80-87

[7] R. Mihalcea, and S. I. Mihalcea: Word Semantics for Information Retrieval: Moving One Step Closer to the Semantic Web. ICTAI 2001: 280-287.

[8] P. Resnik, "Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language", Journal of Artificial Intelligence Research, 1999.

[9] V. Kashyap, and A. P. Sheth, Semantic and schematic similarities between database objects: A context-based approach. VLDB Journal, 5(4):276—304, 1996.

[10] M. Rodriguez, and M. Egenhofer, Determining Semantic Similarity among Entity Classes from Different Ontologies, IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No. 2, March/April 2003.

[11] S. Handschuh, S. Staab. CREAM - CREAting Metadata for the Semantic Web. Computer Networks. 42, pp. 579-598, Elsevier 2003.

[12] B. McBride. Jena: A semantic Web toolkit. IEEE Internet Computing, 6(6), 55-59, 2002.

[13] S. Bechhofer, F. Harmelen, J. Hendler, I. Horrocks, D. McGuinness, P. Patel-Schneider, et al. (2003). OWL Web Ontology Language Reference. W3C Proposed Recommendation, from http://www.w3.org/TR/owl-ref/

[14] O. Lassila, & R. Swick. (1999). Resource Description Framework (RDF) Model and Syntax Specification. W3C Recommendation, from http://www.w3.org/TR/REC-rdf-syntax/

[15] V. Kashup and C. Behrens. "The Emergent Semantic Web: A Consensus approach for Deriving Semantic Knowledge on the Web", Proceedings of the International Semantic Web Working Symposium, July 2001, Stanford, USA.

[16] M. Denny. "Ontology Building: A Survey of Editing Tools", available at http://www.xml.com/pub/a/2002/11/06/ontologies.html

[17] "A survey on ontology tools." OntoWeb Consortium, 2002. http://www.aifb.uni-karlsruhe.de/WBS/ysu/publications/OntoWeb_Del_1-3.pdf

[18] A. Sheth, C. Ramakrishnan. Semantic (Web) Technology In Action: Ontology Driven Information Systems for Search, Integration and Analysis. IEEE Data Engineering Bulletin, Special issue on Making the Semantic Web Real, 26(4), pp. 40-48, 2003.

[19] B. Hammond, A. Sheth, and K. Kochut. Semantic Enhancement Engine: A Modular Document Enhancement Platform for Semantic Applications over Heterogeneous Content. In V. Kashyap & L. Shklar (Eds.), Real World Semantic Web Applications (pp. 29-49): Ios Pr Inc. 2002.