Collaborative and Assisted SKOS Generation and Management

Pierfrancesco Bellini, Antonio Cappuccio, Paolo Nesi
DISIT, Distributed Systems and Internet Technology, Department of Systems and Informatics,
University of Florence, Firenze, Italy, tel: +39-0554796523, fax: +39-055-4796363
email: pbellini@dsi.unifi.it, cappuccio@dsi.unifi.it, nesi@dsi.unifi.it , www: http://www.disit.dsi.unifi.it

Abstract— Despite to the presence of many systems for developing and managing structured taxonomies and/or SKOS models for a given domain, the production and maintenance of these domain knowledge bases is still a very expensive and time consuming process. This paper proposes a solution for assisting expert users in the collaborative development and management of knowledge SKOS. The proposed solution accelerates the SKOS production by crawling and exploiting different kinds of sources (in multiple languages and with several insistencies among them). The proposed WEB based tool supports the experts in defining relationships among the most recurrent concepts reducing the time to SKOS production and allowing collaborative production. The solution has been developed for Open Mind Innovative Space project, with the aim of creating a portal to allow industries at posing semantic queries about potential competences in a large institution such as the University of Florence.

Keywords-SKOS, Semantic Web, Skills Management System

I. Introduction

In the Semantic Web era, representing knowledge in the form of ontologies, thesaurus, taxonomies and other type of semantic data has became increasingly mandatory. The semantic markup is widely available, both to enable sophisticated interoperability among agents and to support human web users in locating and making sense of information.

Among the models for content semantic classification and enrichment the simplified knowledge organization system, SKOS is probably the most diffused [1]. It is a data model for sharing and representing knowledge organization systems (KOS) such as thesauri, classification schemes, term lists, controlled vocabulary and taxonomies within the framework of the Semantic Web. The SKOS model is typically defined using the Resource Description Framework (RDF) [2] and it allows information to be machine-understandable and processable by automatic software agents. Thus, for content classification, the adoption of a SKOS is one of the first step to enter in the semantic world.

The adoption of a knowledge representation structures for content involves many advantages:

- the semantic information is available in machine readable format and can benefit from the emergent technologies of the semantic web;
- the increasing spread of the semantic search engines like Swoogle [3], Watson [4] and SIndice [5] helps to find information with a high degree of precision;
- the users are helped in tagging their content using a predefined set of terms (also called vocabulary or terms, which can be multilingual and supported by a thesauri) belonging to the SKOS that should be validated and accepted by the community. The terms may be also provided in different languages and their translations validated as well.

The SKOS can be also a model at which one may tend after the collection of free tags from the users, the so called folksonomy. The free tags could be statistically analyzed to build a taxonomy and finally a SKOS adding related relationships. Tools for SKOS editing are quite diffuse, such as thManager [7] (a simple SKOS editor), SKOSEd [8], a Protégé plug in for editing SKOS.

However, despite the presence of emerging systems for developing and managing structured taxonomies and/or SKOS models for a given domain, the automated production and maintenance of these domain knowledge bases is still a process often very expensive and time consuming. In the literature, there exist few tools that support the process to pass from text to ontologies and/or SKOS, among them: PoolParty [9], TextToOnto [10], other methods [13] for semi-automatic conversion from well-defined thesaurus like MESH [14] or NCI [15] thesaurus in SKOS format.

In point of fact, software products on the market still require an important phase of manual collection of information from the domain and do not provide satisfactory mechanisms for the coordination of a collaborative production process performed by several groups. Thus, the modeling of domain-based SKOS often turns out to be a manual process and, in most case, it is time consuming and hard. It may involve a large number of personnel that are not easy to be coordinated to collaborate to a unique SKOS.

These problems become very critical when a large knowledge modeling is needed. For example, when one has to SKOSify the activities/documents of a local govern, or of a large University, or of a commercial/industrial district, or of a large editor/press such as IEEE, ACM, Springer, etc. The already in place standard classifications for companies (e.g., European and national codes, for example the ATECO or ISTAT); or for institutions or researchers (e.g., in Italy SSD, and CUN areas) are not suitable to match documents and content with the definitions of their "terms" and "key phrases". Most of those classifications have been frequently produced years ago and with the purpose of manual fitting of activities into them, and not with the purpose of using them for automated classification and reasoning with semantic tools. Moreover, in most cases the manual production of SKOS leads to stress some knowledge area/subarea depending on the knowledge of the experts involved, without taking into account the effective distribution of data to be ingested in the knowledge base.

As a matter of fact, the task of creating a SKOS requires a deep knowledge of the specific domain, and implies:

- the precise understanding of the semantic model behind a SKOS to avoid the production of terms which are not related each other by a specialization/generalization and/or relation relationships;
- the adoption of skilled personnel in both modeling knowledge and application domain or sub-domain;
- the domain analysis and thus the collection of terms in an organized form and relationships;
- a mechanism for coordination of activities in the various stages of the production task;
- the adoption of rules to avoid over-classification (over specialization in the SKOS hierarchy) and under-classification in some areas;

The previous description put in evidence why the knowledge base production process is a time and resource consuming task and not free of errors, even if the target is only the production of a SKOS.

A solution could be to start from the data/content to be classified, and directly extract from these sources the SKOS with an automated or semi-automated process and tool.

This paper proposes a novel solution for assisting expert users in the collaborative development and management of a knowledge modeled as a SKOS. The main idea is to realize a solution and tools to strongly accelerate the process of SKOS production exploiting the real documents/content and web pages to be indexed, and involving the experts in creating relationships among the most recurrent concepts. The solution proposed has been developed in a wider project called Open Mind Innovative Space, which has been founded by ECRF. Open Mind Innovative Space project has as main objective the realization of a portal on which the industries can pose questions with the aim of identifying the competences in terms of researchers and groups in the large knowledge of the University of Florence. In the literature, there is a number of systems that have been proposed to solve the above described problem of helping modeling knowledge bases, may be matching the *demand* (semantic query) against the *offer* (knowledge about domain).

Thus, the problem is still there; how to accelerate the production of SKOS when the domain knowledge is very large and the amount of information to be processed is huge.

The rest of paper is organized as follow. In Section 2, an overview of the Open Mind Innovative Space is provided. Section 3 reports the requirements of the identified tool for collaborative and assisted generation and management of SKOS. Section 4 shows the software architecture that implemented the solution in the global project framework, putting particular emphasis on multi language RDF-SKOS editor. Finally, in Section 5, the achieved experimental results are reported. Conclusions are drawn in section 6.

II. Open Mind Innovative Space Overview

As previously stated, the main goal of the Open Mind Innovative Space project is to realize a service to industries on which they can pose questions with the aim of identifying researchers and groups with the needed competences, knowledge among those of the University of Florence. The University of Florence includes more than 50 different departments belonging to all the scientific sectors areas, and hosting about 2000 researchers and more than 400 labs with their web pages. Each researcher may also teach at 2-3 courses; thus about 6000 course programs that may be considered competence descriptors as well. Moreover, the several research departments and researchers participate to research projects, for a total of about other 20.000 descriptors, etc. In such context it is very hard to identify a manageable number of people that could be reasonably entitled in terms of skill to create a shared common SKOS. This is due to the fact that the whole knowledge model has to be extracted from a huge amount of information, ranging from health care to geometry and math, from engineering to agriculture, from mechanics to statistic and pharmaceutics, etc. And, the sources of this knowledge may change quite dynamically, every year the courses are updated, the CV of people change, other publications and projects arrive, etc.

On the basis of the above description, the available information can be ingested from a large amount of different sources. This highly dynamic collection of sources may be automatically gathered through the use of software agents and crawling tasks. The information gained can be used by a semantic search engine to answer user queries with a high degree of precision. For example, by using an assisted semantic query interface with natural language query engine such as Aqualog [11]. This scenario is shown in Figure 1.

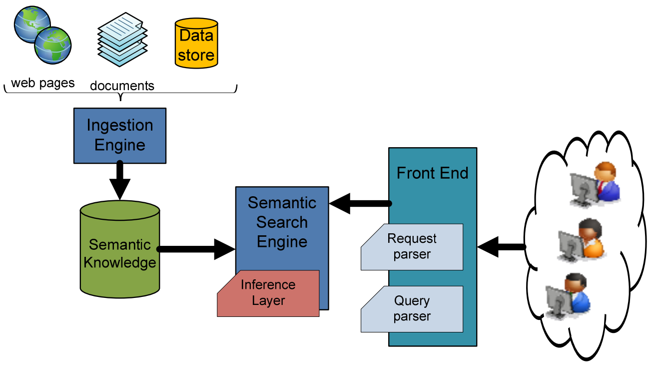

Figure 1. Ingestion and querying semantic knowledge base

This general model, is similar to those used in other mentioned state of the art solutions, and poses its bases on the availability of a good ontology, may be including a SKOS of the domain knowledge. The general ontology can be composed of dynamic blocks or subsections that describe the various features of the domain. They include both concepts that describe the physical structure of the domain and concepts that describe relations about personnel.

As a first step we had taken into account the fact that, the huge amount of data to be processed belong to governmental institutions such as the university and thus also on people. More specifically, the domain knowledge is composed by three self-supporting ontologies which are related by semantic relationships. Therefore, the basic elements of the knowledge base are those regarding:

- **Friend of a Friend (FOAF) ontology** [18] used to model many properties about Person and Organization class (professor, phd students, students, researchers, contractors, their relationships, research classification as SSD, CUN, etc.): the name, the surname, the e-mail properties and the *knows* relationship (applicable to individual belonging to the Person class).
- **Academic life ontology** is an ontology, we developed specifically for the Italian University case structure and terminology, that defines elements for describing universities and the activities that occur at them (labs, departments, faculties, research centers, groups, projects, courses, curricula, matter, projects, integrated labs, etc.). The main OWL entities and classes described by ontology are:
  - **Organization** class describes physical structures of university like research center, departments and laboratories;
  - **People and role** describe instances like full-professors, researchers and PhD students, related and derived from FOAF concepts;
  - **Activity entities** that cover concepts like past projects, ongoing projects and academic publications; To each person the specific publications are added as well, establishing in this way also relationships among the different authors.
- **Competences SKOS:** it is the SKOS ontology that describes the hierarchy of the technical skills of structures and people belonging to the given application context. This part of the knowledge is the most dynamic.

The components related to the **Academic life ontology** and to the **FOAF** are initialized and directly populated by gathering information from the central database of the University and of other institutions. Among them the central CINECA servers [19]. This operation is performed with a set of crawling tasks realized by using SOAP Client implemented in JAVA making use of JAX-WS [20].

On the basis of the described project, the most critical aspect is the modeling and population of the above mentioned **Competence SKOS** for the whole university area. Typically, in these cases the solution proposed is to manually produce a coarse classification. On the other hand, what it is really needed is to arrive at a SKOS strongly related to the real sources of descriptors to allow the automated classification and reasoning.

For these reasons we started with the idea of producing a solution for assisting expert users in the collaborative development and management of a **Competence SKOS**, the ***Collaborative SKOS Accelerator and Manager,* CoSKOSAM**. With the aim of accelerating the process of SKOS production and population. In the next section the identified requirements are presented.

III. Requirements of CoSKOSAM

The CoSKOSAM is not only an editor, its requirements put in evidence that the aim was to create a collaborative environment in which several experts can contribute to the production and management of the same SKOS. And the systems may help them in identifying the keywords and concepts which are located in the real documents and sources. The main functional requirements for the CoSKOSAM have been to provide the capability of:

- ingesting and analyzing content from a large amount of different sources (web pages, cv, documents, etc.) to extract keywords and concepts and keep them linked with the original context/source;
- updating the crawling and ingestion of the content and thus the update of the semantic structures related to the identified keywords and concepts;
- helping the area editors with suitable tools that allow them to identify the most relevant keywords and concepts;
- managing content coming from different languages so that to map the concepts into a multilingual knowledge base, also providing/exploiting translation services and utilities;
- creating and editing a multi language SKOS about the identified concepts/competences/skills of personnel and research centers of the academic structure;
- supporting collaborative crawling, management and editing of the SKOS structures;
- allowing the incremental and distributed production of knowledge;
- supporting the integration of produced SKOS in terms of related terms and synonymous.

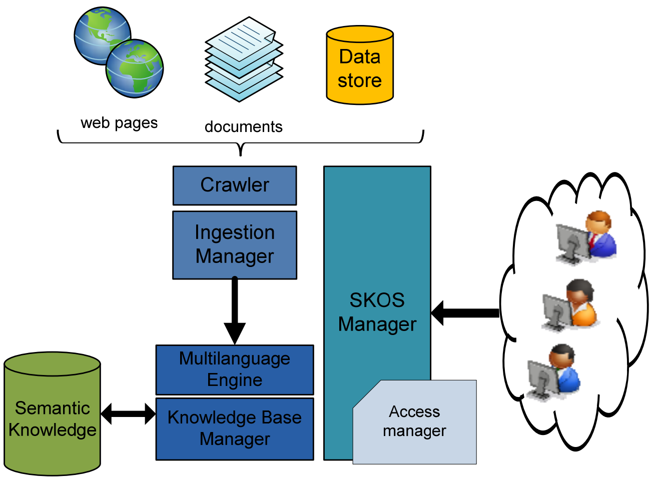

Figure 2. Architecture of Knowledge Base Management System

The benefits of creating a **Competence SKOS** via a WEB-based collaborative environment and management are many. Using CoSKOSAM skilled staff, such as directors of departments, or sector referent people, are enabled to build and manage the knowledge base incrementally and collaboratively confronting with their colleagues toward the common goal of realizing a high quality product.

IV. CoSKOSAM Architecture

The CoSKOSAM architecture is shown in Figure 2. It has been implemented in JAVA as a web application, using a client server architecture.

The architecture consists of some primary building blocks:

- ***SKOS Manager*** provides services for creating and maintaining a SKOS of keywords and concepts formalized according to a SKOS vocabulary.
- ***Access Manager*** offers services for the security login and grant check. It helps to synchronize and coordinate the access to the various sub-sections of the SKOS among different users. The privileges pyramid provisioned by system can be grouped in the following three main categories:
  - *Administrators users*: users who have the highest privileges on semantic store.
  - *Writer users:* users who can edit, read and write the working knowledge base but cannot execute backup and crawling activities;
  - *Reader users:* guest users. They can only read the knowledge base but cannot change it in any way;
- ***Crawler and Ingestion Manager*** offer the ability for crawling and ingesting sources to feed the knowledge base. The crawling task performs a breadth-first search on graph of the web structure of University and ingests some kind of domain-information like personnel, courses, staff curriculum, advertising, research centers, etc. Specifically, it makes use of the GATE NLP Platform [16] to implement the crawler architecture;
- ***Multilanguage Engine*** provides services for the management of multilingual SKOS taxonomy. The information ingested by crawling task is automatically translated by Multilanguage Engine in the languages handled by module. The sources are frequently produced in only one language while the keywords and concepts have to be declined in multiple language (e.g., in Italian and English at least).
- ***Knowledge Base manager*** provides the software API interface for manipulating the domain ontology. The knowledge base has been built by using the API provided by the Sesame framework [17].

A. *SKOS Manager*

The SKOS Manager provides services for creating, managing, and maintaining a multilingual SKOS model as comprised of concepts compounded from keywords extracted from several kinds of sources by the Ingestion Manager. The approach allows the organization of concepts into concept schemes where it is possible to indicate semantic relationships between terms. The SKOS Manager enables the complete development of a SKOS by providing a range of services, for the incremental, collaborative and multilingual development of the common knowledge via WEB.

Each information about Competence SKOS that is ingested by previous crawling processes is stored in the knowledge base as an instance of the skos class: *Concept*. A user of the SKOS Manager, which is expert in a given sub-domain such as a single department, has the chance to change the knowledge base by adding new concepts and inserting semantic relationships among already existing concepts.

The relations between concept over the SKOS vocabulary allow to add semantic information to the knowledge base.

The allowed relations are:

- **skos:conceptSchema** relation provides the ability to express the origin of a concept in a concept scheme;
- **skos:hasTopConcept** relation provides the ability to express the mayor topics that are wrapped up into a concept scheme;
- **skos:broader** relation must be used to express the fact that a concept is in some way less general than another.
- **skos:narrower** is the inverse relation of skos:broader;
- **skos:related** relation provides the ability to create associative links between concepts. The property carries weak semantics and it expresses the fact that two concepts are in some way related, and that the relationship should not be used to create a hierarchy but for create links between branches of a hierarchy of concepts;
- **skos:prefLabel** is the preferred label associate to concept in a given language. A label is any word, phrase or symbol that can be used to refer to the concept by people. When a user adds a label to SKOS the system provides to automatically translate the text-label in the right language by exploiting an external service (in any case the translation may be corrected by the user).

A graphic explanation of generated SKOS vocabulary is shown in Figure 3.

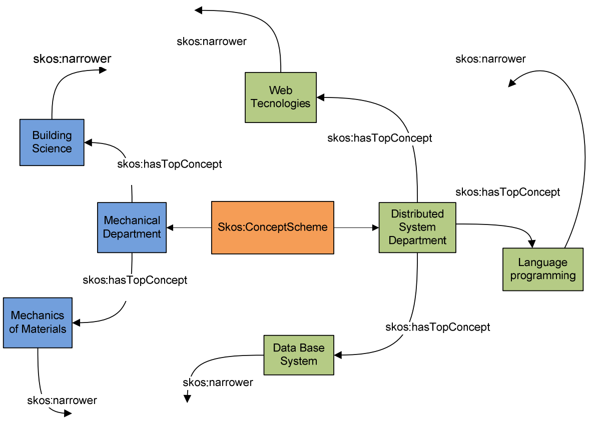

Figure 3. Generated SKOS Knowledge Base

Multilingual and multicultural issues are dealt to assure a wider and more effective exploitation of data beside the background of the operator and their location. Furthermore a multilingual approach helps to improve *precision and recall* of popular search engine, which are very good at retrieving the accurate information.

Other features provided by the system are:

- searching by label in the working taxonomy and discovery branches that contain the target label;
- view the semantic information about personnel or structures related to each *skos:Concept*; that is the link to the original sources;
- view the frequency with which the concepts occur in source web documents;
- filtering the concepts by frequencies. In this way the user has the opportunity to discard some labels that may be considered statistically less remarkable or highly specializing;
- ability to edit concepts in multilingual mode taking advantage of automatic translations. An user can also manually re-translate a label for achieving an higher quality result;
- visual system log for provide additional information about working knowledge base;
- more than a graphical view for displaying the SKOS concepts:
  - only label: displays the label in the current language
  - label with frequencies: displays next the label, the frequencies of the current concept;
  - label with language: displays next to the label, the dual language to the current one.

V. Experimental Results

In this section, the achieved results and some user experiences are presented. The proposed tool has been used to develop an experimental knowledge base related to DSI Department, one of the most active departments of the University of Florence [21]. It is under application by 3 different departments, while the development of the Competence SKOS for each department is autonomously developed as separate SKOS branches. They are lately merged together in an unique SKOS by joining concepts and adding skos:related relationships.

The aim of this experimentation has been: (i) the ingestion of the whole sources related to DSI department; and (ii) the building of a multilingual Competence SKOS about department and its personnel. The department included about 62 researchers (28 full professors, 13 associate professors, 21 researchers), and about 160 courses and programs, related publications and web pages, for more than 2000 publications. All staff accessible via web.

The department includes 5 different sectors/research-areas such as math, operating research, computer science, computer engineering, automated control. The people and activities of those areas have different competences and skills and are difficult to be represented by a unique person.

The CoSKOSAM started crawling and ingesting all data and producing the list of basic concepts as presented in Figure 4 in which the SKOS Manager is shown.

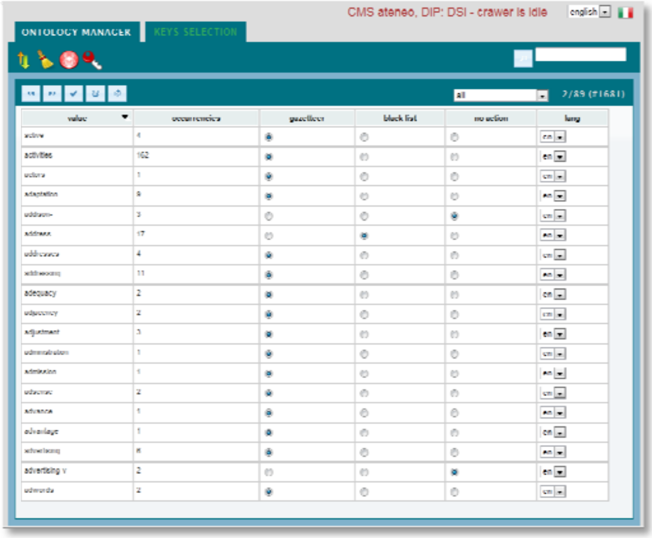

Figure 4. SKOS Manager Web user interface – setting up crawling task stage for basic concepts

The extracted basic concepts/keywords has been browsed by the reference domain expert to identify those that have been erroneously classified to the wrong language. The list of them can be filtered to present only those that have a frequency greater than a given value. This allows the expert to focus its time on the most relevant one. At the end of this process, the selected keywords are put in the gazetteer of the semantic Ingestion Manager based on GATE [16]. In the case of DSI, the process extracted about 1600 terms aligned by the system in both languages, revised by the expert in about 1 hour of work.

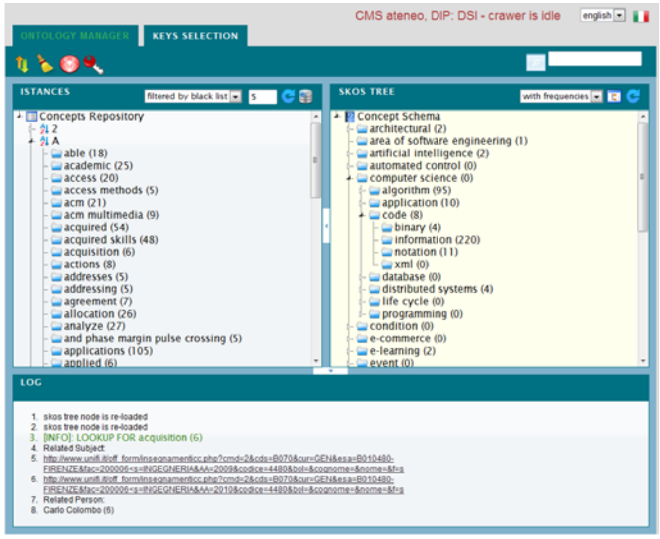

Figure 5. SKOS Manager Web user interface – setting up Conpetences SKOS from the bsic concepts.

After this phase, the basic concepts and keywords can be used to create the Competence SKOS passing to a different interface of the SKOS Manager, as shown in Figure 5. On the left side, the list of basic concepts with their frequency (alphabetically ordered), while on the right side; the produced Competence SKOS are reported. By using the mouse, the user of the SKOS Manager may look up at the original sources referred by both basic concepts or SKOS concepts and reported in the log box below.

The SKOS on the right side can be manually created by adding new concepts or by using the drag and drop paradigm by exploiting and arranging the basic concepts on the left side.

The produced Competence SKOS for the DSI consists of different classes and has required 1 day of work for its completion.

At the end of the Competence SKOS production for DSI, the remaining basic concepts not allocated were 1937 and the produced SKOS was based on 786 (for each language) concepts organized in 27 levels. The domain OWL ontology for the DSI only has 3270 individuals and 45 OWL classes.

The Competence SKOS produced has been validated by experts of the sectors by using the look up facility and in short by observing the sources identified and connected a large sample of broader concepts. Another form of validation has been the production of simple SPARQL queries with and without the usage of the inferential engine. The inferential engine, in this case, exploited both the hierarchy of concepts and the related relationships the SKOS.

VI. Conclusions And Future Work

This paper proposed CoSKOSAM, a web based solution for accelerating the production and management of SKOS derived from sources that can be located on Web. The solution and proposed tool has been used for the automated ingestion and analysis about the university life, including afferent organization and technical skills. The hierarchical structure of the competences and the semantic relation among them have been formalized using the SKOS vocabulary, providing a developing method collaborative, computer-aided and coordinated activity.

The methodology greatly reduces the time spent in the development process aiding the users in all stages of the production process. Furthermore, the ontology is produced according to the OWL/RDF/SKOS rules and can benefit from emerging technologies and innovations offered by the semantic web. The generated ontology can be used as information domain by a demand and supply system about academic skills. It is currently in connection with a semantic database for allowing performing SPARQL queries allowing:

- semantic search engine to retrieve ranked information. For computing ranking it is possible making use of term frequency as a factor weighting within the ranking algorithm;
- semantic indexing for search engine optimization and fuzzy queries.
- exploiting inferential engine to increase the system intelligence,
- improving the engine for providing results to the users and permitting them to navigate in the mesh of relationships among FOAF entities and results;

In addition of this, some guide lines for short-term future work have been already planned, in particular regarding usability and user support.

Acknowledgements: The authors would like to thank all contributor to the Open Mind project and in particular Prof. M. Lombardi and Prof. P. Rissone.

References

[1] SKOS: Simple Knowledge Organization for the Web. [Online]. Available: http://www.w3.org/2004/02/skos/

[2] RDF – Vocabulary Description Lanaguage 1.0: RDF Schema. [Online]. Available: http://www.w3.org/TR/rdf-schema/

[3] D. Li; T. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V. C Doshi, and J. Sachs (2004). "Swoogle: A Search and Metadata Engine for the Semantic Web". Proc. of the Thirteenth ACM Conference on Information and Knowledge Management (CIKM). ACM. pp. 652–659.

[4] Mathieu d'Aquin and Enrico Motta, "Watson, more than a Semantic Web search engine", in press.

[5] G. Tummarello, E. Oren, and R. Delbru. Sindice.com: Weaving the open linked data. In Proceedings of the 6th International Semantic Web Conf and 2nd Asian Semantic Web Conf (ISWC/ASWC2007), Busan, South Korea, vol.4825 of LNCS, Berlin, Nov. 2007. Springer Verlag.

[6] M. d'Aquin. Building SemanticWeb Based Applications with Watson. In WWW2008 - The 17th International World Wide Web Conference - Developers' Track, 2008.

[7] ThManager – an Open Source Tool for creating and visualizing RDF SKOS vocabularies

[8] SKOSEd-Thesaurus editor for the Semantic Web. [Online]. Available: http://code.google.com/p/skoseditor/.

[9] Pool Party - SKOS Thesaurus Management. [Online]. Available: http://poolparty.punkt.at/

[10] Alexander Maedche and Steffen Staab. The TEXT-TO-ONTO Ontology Learning Environment. Institute AIFB, University of Karlsruhe, 76128 Karlsruhe, Germany.

[11] V. Lopez, E. Motta, E., V. Uren, and M. Pasin, (2007) AquaLog: An ontology-driven Question Answering System for Semantic intranets, Journal of Web Semantics, 5(2), p.72-105.

[12] Protégé, Protégé enviroment. [Online]. Available: http://protege.stanford.edu.

[13] M. van Assem, V. Malais´e, A. Miles, and G.Schreiber. A method to convert thesauri to SKOS. In Proceedings of the third European SemanticWeb Conference (ESWC 2006), pages 95–109, Budva, Montenegro, June 11-14 2006.

[14] Soualmia L.F., C. Goldbreich, and S.J. Darmoni. Representing the mesh in owl: Towards a semi-automatic migration. InProc. of the 1st Int'l Workshop on Formal Biomedical Knowledge Representation (KR-MED 2004), pages 81–87, Whistler, Canada, 2004.

[15] J. Goldbeck, G. Fragoso, F. Hartel, J. Hendler, B. Parsia, and J. Oberthaler. The National Cancer Institute's Thesaurus and Ontology.Journal of Web Semantics, 1(1), Dec 2003.

[16] Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.:GATE: A Framework and Graphical Development Enviroment formRobust NLP Tools and Applications. In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL '02), Philadelphia (2002).

[17] Arjohn Kampman, Frank van Harmelem, Jenn Broekstra: Sesame: Ageneric architecure for storing and querying rdf and rdf schema: in proceedings of ISWC 2002, October 7-10, Sardinia, Italy (2002).

[18] The Friend of a Friend (FOAF) Project. [Online]. Available: http://www.foaf-project.org/.

[19] Centro di supercalcolo, Consorzio di Università. [Online]. Available: http://www.cineca.it/page/chi-siamo

[20] JAX-WS. [Online]. Available: http://jax-ws.java.net/.

[21] DISIT Lab – Department of Systems and Informatics, University of Florence. [Online]. Available:http://www.disit.dsi.unifi.it/.