A Semantically Enriched Competency Management System to Support the Analysis of a Web-based Research Network

Paola Velardi
University of Roma La Sapienza
Italy
velardi@di.uniroma1.it

Alessandro Cucchiarelli
Università Politecnica delle Marche
Italy
cucchiarelli@diiga.univpm.it

Michaël Pétit
University of Namur
Belgium
mpe@info.fundp.ac.be

Abstract

While it is generally acknowledged that domain ontologies can significantly improve knowledge management systems (KMS) within organizations and among distributed web communities, we have little evidence of operational ontology-based KMS and their practical utility in real settings in the literature. We describe here the INTEROP KMap, a fully implemented, semantically indexed, competency management system, used to facilitate research collaboration and coordination of a Network of Excellence (NoE) on Enterprise Interoperability.

Since the main highlighted advantages of ontologies are improved information access and interoperability, our aim in this paper is to give experimental support to these claims. We provide a summary description and usage data on the KMap, as well as experiments to quantify the added value of semantic search wrt traditional document ranking measures.

1. Introduction

Throughout the past decades the *management of competencies* has become an important topic for many medium-large enterprises, public institutions and more in general, "communities of knowledge" [1]. Many companies have implemented a Competency Management System (CMS) since the use of these systems may bring many advantages[1]: behavioral guidelines are given, performance standards are evident, and consequently, communication between employer and employee might improve. CMS provides a quantitative measurement of an institution's knowledge and functions as:

- A communication tool across functional business processes
- Data needed to analyze the actual utilization of workforce competencies
- A mechanism to forecast and monitor future needs
- Continuity of workforce capability across programs and projects
- Alignment of workforce competencies to strategic drivers

Unfortunately, in practice, CMSs often seem to receive a limited consensus from the actual users and therefore their usefulness has become a point of discussion. In a recently published study[2] the major problems of CMSs, highlighted by their end users, are: limited *flexibility* (especially to fit user's competency model: "*not flexible enough to accommodate my design*" or "*don't allow enough customization*"), complex or laborious *editing* and *managing* tools (e.g. adding and indexing files), and more in general, poor attention to *knowledge evolution* and *easiness of use*.

The consideration of these problems has been central to the design of a CMS built during three years of duration of the INTEROP EC Network of Excellence, a project focused on enterprise interoperability that involved about 60 research and industrial institutions in Europe.

One of the main objectives of INTEROP has been to build a so-called "*Knowledge Map*" (KMap) of partner competences, in order to perform a periodic diagnostics of the extent of research collaboration and coordination among the NoE members. In the KMap, information is indexed and classified according to a *domain ontology*, which was created using a semi-automated methodology, aimed at minimizing manual effort and maximizing user consensus. In fact, limited involvement of users in the conceptual modeling process and a static view of the domain ontology are reported to be among the main causes of failure of KMS so far [2] [3].

The ontology population strategy has already been described in [4]. The objective of this paper is instead to analyse, in a qualitative and quantitative way, the advantages of a semantically indexed knowledge management system, both in terms of *clarity of usage* and of *improved information accessibility*.

[1] See e.g. NASA CMS http://ohcm.gsfc.nasa.gov/cms/home.htm.

[2] http://iainstitute.org/pg/the_problems_with_cms.php.

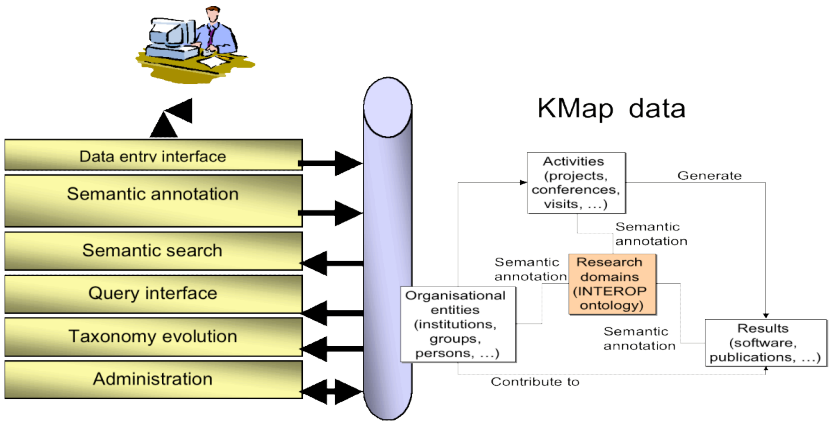

Figure 1. Schema of the KMap data and functionalities

2. Overview of the INTEROP KMap

This section provides a brief overview of the KMap architecture. Its aim is to monitor the status of research in the field of interoperability through a web-based platform that allows the user to retrieve information according to his/her actual needs in a specific situation.

The main benefits of the KMap for its users are:

- To be able to diagnose current interoperability research within INTEROP and in Europe;
- To receive an overview of all European research activities on interoperability and subordinated topics;
- To receive an overview of organisations and experts as well as research results;
- To find relevant information for specific needs quickly;
- To find potential partners for collaborating in research activities.

The KMap architecture and functionalities are shown in Figure 1. The key features are:

1. *Reduction of the knowledge acquisition bottleneck problem*, through the use of automatic ontology population methods, and through the continuous involvement of NoE members in the validation and evolution of the adopted conceptual model.
2. *Semantic indexing* of all KMap entities, either automated, or manual, depending upon the best compromise between the needs of fitting a user's conceptual model, and reducing tedious annotation time.
3. *Semantically-guided search* (complemented by standard database querying), that provides the means for achieving the desired analysis of the enterprise interoperability domain, identifying knowledge gaps, the emergence of new concepts, and progress towards research de-fragmentation, which was the main target of the NoE.

The objective of this paper is to describe features 2) and 3), and to provide an experimental analysis of the benefits of semantic indexing to achieve the main CMS objectives: *improved collaboration and monitoring of a research community*.

As far as feature 1) is concerned, the interested reader is diverted to [4] and [5]. In short, the ontology has been learned semi-automatically, according to the knowledge acquisition value chain briefly sketched in Figure 2. Domain knowledge is automatically acquired from web or user-provided documents, and is then progressively formalized: first, a domain terminology is acquired, then a glossary. Glossary terms are taxonomically structured, and finally, the taxonomy is enriched with additional semantic relations. Each learning step is followed by manual validation and refinement by domain experts, through appropriate collaborative interfaces. For simpler validation tasks, such as terminology and glossary, all the INTEROP members have been involved. For more complex tasks, like taxonomy and ontology validation, a restricted team was appointed.

Finally, the very same knowledge acquisition and validation process is periodically repeated to ensure knowledge evolution.

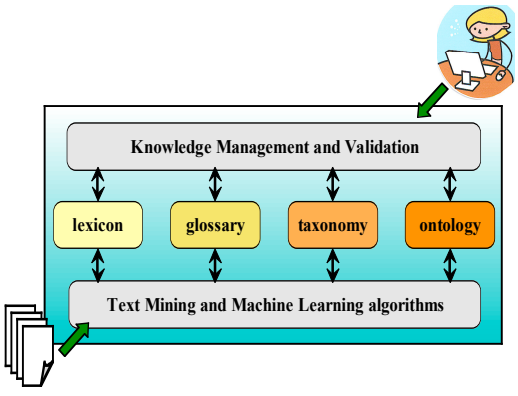

Figure 2. Knowledge acquisition value chain

3. Semantic enrichment

Whichever strategy might be conceived to continuously improve the quality and quantity of a knowledge repository, and to foster user involvement, a primary obstacle to an effective use of the stored knowledge is poor accessibility. Every web user has experimented the frustration of searching for a specific piece of information, and not finding it. The INTEROP collaborative platform[3] did not escape the accessibility issue: a primary complaint of users, as the quantity of stored information had started to grow, was the difficulty to locate what they were interested in.

Semantic enrichment [6] is acknowledged as the most promising technique to improve accessibility of the information on the web and in document warehouses. Ontology-based semantic enrichment has been used in the INTEROP platform to index the KMap and to perform semantic search.

All KMap entities have been semantically enriched with the ontology concepts. There are two ways to add semantic annotations: manual and automatic. Figure 3 shows an example of manual annotation: starting from his KMap page, a researcher can select or search in the ontology a certain number of concepts describing his/her expertise, and assign a weight ranging from 0 to 1 to each concept. A weighted concept vector is thus associated to the "researcher" entity. Instead, to automatically enrich an activity or publication, textual descriptions (or the full publication, if available) are converted from almost any format into a text file, and then analyzed to extract occurrences of ontology concepts[4]. The detected concepts are weighted according to the standard term-frequency-inverse-document-frequency[5] measure used in Information Retrieval. The extracted concept vector is presented to the user for (optional) validation and refinement.

Notice that, in principle, even a researcher or organization can be automatically annotated: for example, the publications and activities associated to a researcher can be used to create his-her concept vector, whether using an automated strategy be the choice of who is entering the information or not. Especially when describing one's own competence domains, the manual strategy seems preferable, while automatic annotation of documents greatly reduces the human effort, so limited, if required, to post-validation.

Semantic enrichment of all KMap entities allows two types of search:

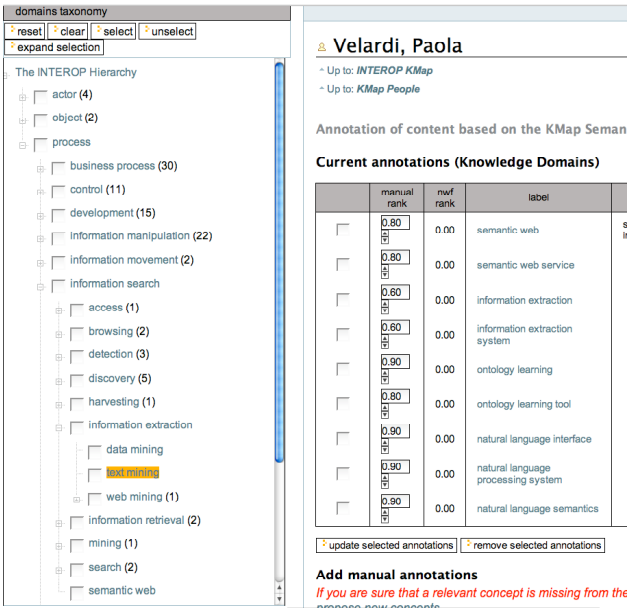

Figure 3. Manual selection of semantic annotations

1) Semantic search;
2) Similarity-based search.

In what follows we briefly describe these two features.

3.1. Semantic search

Traditional information retrieval allows it to search and retrieve indexed information on the basis of keyword match: the user types one or more keywords, and the system retrieves documents including all or some of these keywords. To automatically expand a query with "related" terms, algebraic and statistical methods have been used, like Latent Semantic Indexing[6] and Query Expansion (e.g. [7]). Ontology-based expansion [8] has been successfully experimented only in domains where large ontologies are available, such as medicine. The adoption of knowledge-based approaches has been so far constrained by the difficulty to build large-scale domain ontologies, a limitation that we have overcome thanks to the semi-automated ontology learning methodology introduced in [9] and extensively applied in the INTEROP project [4].

One of the critical problems of ontology-based query expansion is to determine the scope of the expansion: extending a keyword with its synonyms is reasonably "safe", but even one step of generalization or specialization through the ontology may introduce noise[7]. This negative effect is more visible in open domains than in restricted domains.

---

[3] http://www.interop-noe.org.

[4] The ontology stores, for each concept, a list of its lexical variants, including acronyms, e.g. SOA or Service Oriented Architecture.

[5] See the IDF page: www.soi.city.ac.uk/~ser/idf.html.

[6] www.cs.utk.edu/~lsi/.

[7] Because of ambiguity and over-generality.

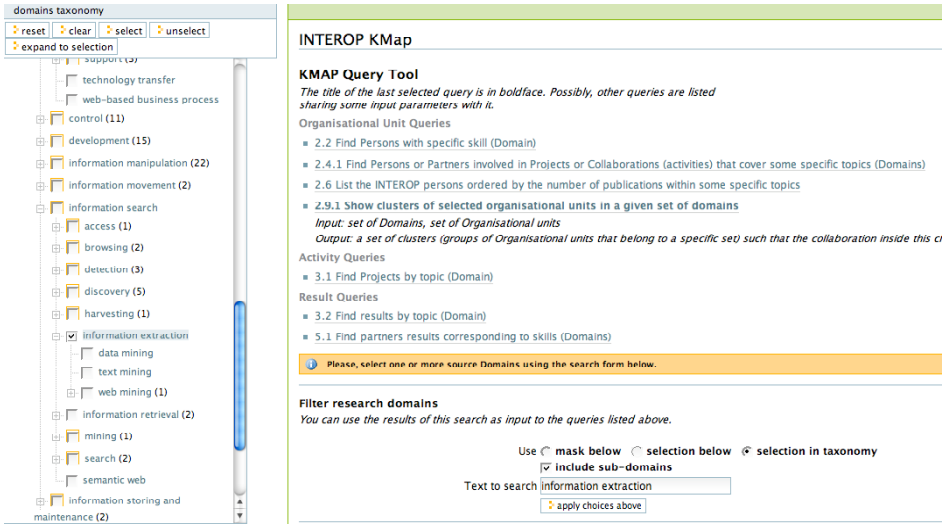

Figure 4. Selection of competence domains for query 2.9.1

In the KMap there are two ways to improve search results through the ontology: the first requires user feedback, as shown in Figure 3. Figure 4 illustrates one among several available query types: to identify a researcher with given competence skills. The user specifies the query keywords (boolean queries are allowed) and, by selecting the "include sub-domains" box, he is presented with a list of ontology concepts matching the query concepts or being a specialization (up to the leaves) of the selected concepts, as shown on the left side of the screen dump in Figure 4. Therefore, if the selected skill is, for example, information extraction, also the concepts {text mining, data mining, web mining} are included. Before actually carrying out the search, the user can select/deselect all the retrieved concepts, thus eliminating overly general or overly specific (according to his intuition) expansions. This may seem tedious, but since users tend to specify keywords that are rather close to the ontology leaves, it has turned out to be a reasonable strategy in order to improve search results. A second way of accessing information is through similarity-based clustering, as discussed in the next section.

3.2. Similarity-based clustering

Semantic enrichment, as described in section 3, associates to each KMap entity E a vector V_E, whose components are the weights associated to ontology concepts automatically detected in the textual description(s) of E, or manually added.

It is often the case that, when a user finds some relevant information in a repository, he then wishes to search for similar information. For example, a researcher may want to find other researchers with a similar competence, or, starting from a project description, to find the best-fitting partners to cooperate, etc. Most web browsers provide a search-by-similarity feature, but they work quite poorly. In a restricted domain, however, similarity search can be effectively computed as follows:

- The similarity measure can be computed only on the basis of domain-relevant concepts, e.g. the semantic vectors V, rather than using all the words appearing in an entity description.
- Given the limited ambiguity of terms in specific domains, a contribution to the similarity computation can also be given by concept pairs that do not match in the compared vectors, but are semantically related in the ontology.

To give a practical example, consider a paper that applies interoperability techniques in a medical domain. Under the interoperability perspective, medical terms are not relevant for a similarity search: the user would rather prefer to find papers that use the same, or similar techniques, or that address the same or similar problems, regardless of the application domain. In a medical domain, exactly the opposite would apply: the concept of "similarity" is indeed domain-dependent. Therefore, to compute a similarity measure, the domain-ontology concepts appearing in a document must have a major role.

Details regarding the semantic similarity measure we used are provided in [4]. An intuition of the similarity formula is provided by the following simplified expression[8], where V_A and V_B are the concept vectors of the entities A and B:

$$(1) \quad sim(A,B) = \alpha \cdot D\cos\_sim(V_A,V_B) + \beta \cdot I\cos\_sim(V_A,V_B)$$

[8] We performed several experiments to calibrate the parameters in (1) and to determine the type of indirect matches. These aspects are not discussed here because they are outside the scope of the paper.

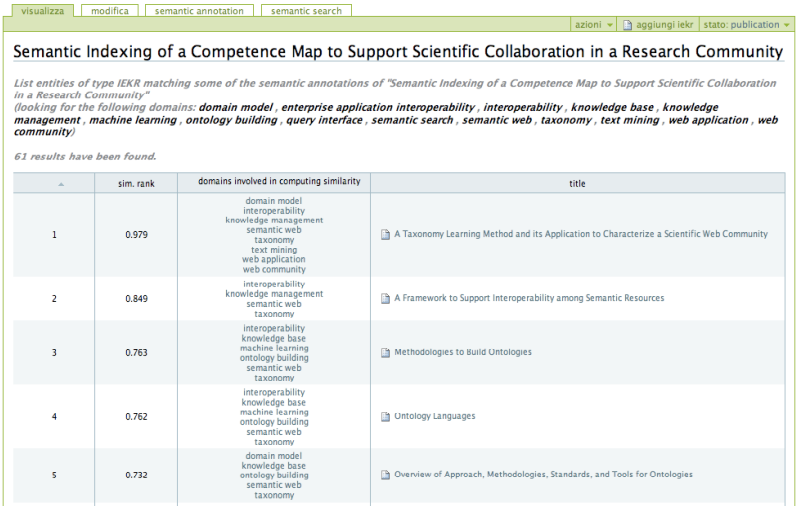

Figure 5. Tabular view of a similarity search result, starting from an IEKR paper

In (1), *cos_sim* is the well known *cosine similarity* formula, *Dcos_sim* accounts only for *direct* matches between concepts in V_A and V_B, while *Icos_sim* accounts for *indirect* concept matches (sibling concepts and concepts related by a direct kind-of relations[9]).

In the INTEROP KMap, similarity-based search is possible starting from any entity or from any search result, and furthermore it is possible to compare entities of a different type (e.g. publications and researchers). The similarity formula is computed automatically, and it is possible, for advanced users, to set the values of the α and β parameters and a threshold for the computed similarity value. We may then analyze the effectiveness of indirect concept matches (or *semantic matches*) in improving search results more precisely.

3.3. Experiments on ontology-based search

The experiments have been performed on the KMap Publications repository, named the Interoperability Document Repository (IEKR). The IEKR stores full papers, therefore their automatically created semantic vectors are in general richer than other manually annotated KMap entities.

The first experiment concerned the tuning of α and β parameters. We selected 10 papers at random from the Interoperability Document Repository. For each paper, we extracted the 10 most similar documents, according to the similarity formula (1) applied with different values of the α and β parameters. Figure 5 represents the tabular view of a similarity search result: concepts contributing to indirect matches are shown in red. By inspecting these results, we may evaluate the impact of semantic matches, using a subjective judgment of similarity. Rather than judging the actual similarity of each document pair, we compared alternative orderings of similarity-ranked papers, for sake of simplicity and whenever it seemed sufficient, only on the basis of the paper titles. The best results were obtained with a "cautious" contribution of indirect matching, obtained with $\alpha = 1.0$ and $\beta = 0.2$.

Table 1 shows the results of an experiment in which two cases are compared: the one in which only direct matches are considered ($\beta=0$, similarity values in column "*Dir*") and the one in which all matches are included, with $\beta=0.2$ (similarity values in column "*All*"). The two "Delta" columns show an average increase of similarity of 24.24% and an increase of 30.40% of retrieved documents when the minimum similarity threshold is set to 0.3.

Table 1. Average document similarity and average # of retrieved documents with and without semantic matches.

Doc.	Average similarity rank of first 10 retrieved document			Number of documents with similarity > 0.3		
	Dir	*All*	*Delta %*	*Dir*	*All*	*Delta %*
1	0.6021	0.6225	3.39	27	30	11.11
2	0.7460	0.7517	0.76	27	29	7.41
3	0.3784	0.4320	14.16	9	15	66.67
4	0.7500	0.7852	4.69	30	35	16.67
5	0.3433	0.5579	62.51	10	26	160.00
6	0.5212	0.5605	7.54	12	16	33.33
7	0.2990	0.2990	0.00	3	3	0.00
8	0.1194	0.2771	132.08	2	2	0.00
9	0.8374	0.8583	2.50	34	37	8.82
10	0.3040	0.3489	14.77	2	2	0.00
		Average Delta	24.24		Average Delta	30.40

[9] Notice that the specific formula that we use in [4] separates the contribution of sibling matches from that of hypernym-hyponym matches.

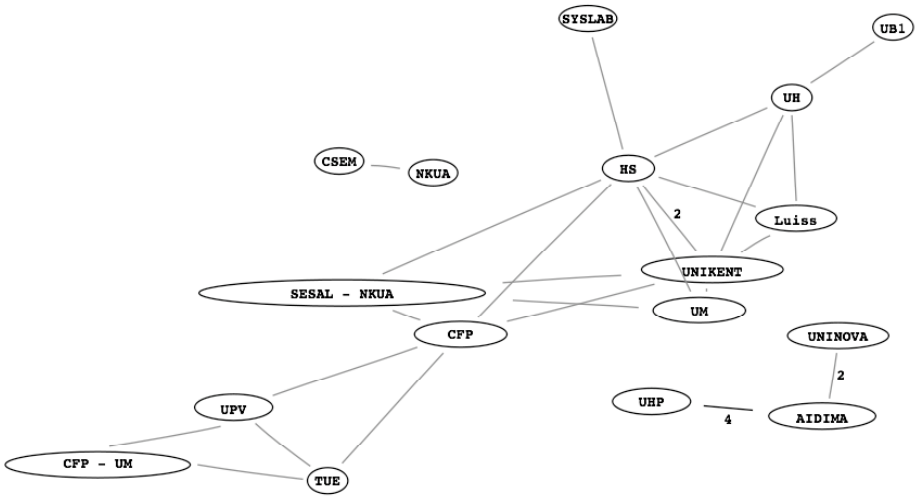

Figure 6. Institution collaborating on "*architecture*" related topics in 2006

One additional advantage of the adopted information search strategy is *clarity*: not only is the user presented, as in traditional information retrieval systems, with a ranked list of retrieved documents[10], but also with an explanation (provided by the column "*domains involved in computing similarity*" in Figure 5) of the concepts directly or indirectly contributing to the similarity computation. A graphical view of a similarity computation result can also be computed, as in the example of Figure 6 (which refers to another kind of search result, as detailed later).

4. Experimental data on KMap usage to support network analysis

This section is dedicated to a summary analysis of the KMap effectiveness, as far as its main intended usages are concerned, that is: i) to support periodic diagnosis of the network status and progress ii) to improve members cooperation and reduce research de-fragmentation. For the sake of space, we only report the most meaningful results and usage data.

One of the major targets of EU NoEs in general, is to foster research de-fragmentation. The INTEROP NoE has had to produce periodic evidence of progress in this direction. This task is usually accomplished in a rather simple way, showing some increase in the number of co-authored publications, co-organized workshops, etc.

The KMap allows a considerably more sophisticated analysis of de-fragmentation. For example, for each of the three main INTEROP research areas (*ontology, interoperability architectures, enterprise modeling*), we may compute the increase of cooperation among organizations and/or researchers. An example of query supporting this analysis is: "*Show clusters of organizations cooperating on a set of specified topics*" (Query type 2.9.1). This query returns a list of NoE partners cooperating on a set of topics, specified by selecting one or more ontology concepts.

In the absence of semantic expansion of the query keywords (see section 3.1), only collaboration types (papers, projects, workshops, …) including precisely the specified keyword (say, "*architecture*") are retrieved. In the KMap, we may instead retrieve all the collaboration types annotated either with the specified keyword(s), or with its related concepts in the ontology. The results are presented either in tabular form, or in a graphical form, as in Figure 6. The figure shows all the collaborating partners with collaboration relevant to the search criterion, and furthermore it shows the number of shared activities between pairs (the thicker the line, the higher the cooperation between connected entities). Nodes and edges can be clicked to visualize details.

In this way it is possible to produce a fine-grained analysis of the improvement of collaborations in specific research areas throughout the duration of the project. Similarly, it is possible to identify knowledge gaps, "hot" research topics, and partners who do not cooperate, even though they have shared interests and competencies.

The KMap features have clearly facilitated the task of partners in charge of producing diagnosis reports, but how far has it been accepted by the INTEROP end-users? This is a more complex question to answer, since the KMap was fully completed only in the final year of the project.

[10] We use the term "document" for analogy with traditional information retrieval systems, but in the Kmap any type of entity (researchers, publications, products, …) can be searched and retrieved in the very same way.

Table 2. Summary data on partner's effort to populate and annotate the KMap (on December 2006)

# of sem. annotations / # of defined organisational entities (person, group or organisation)	2241/718 = 3,12
# of sem. annotations / # activity	330/337 = 0,98
# of sem. annotations / # result	829/712 = 1,16

Table 2 provides a partial picture: it measures the data entry and manual annotation effort of partners, relative to the three main KMap entity types: *organizations*, *activities* and *results* (see Figure 1). The table shows that, as expected, partners understand the utility of semantic annotations, but they prefer to annotate persons or organizations (the average is more than 3 concepts added for each entered organization or person), rather than activities or results, for which they have already been asked to provide a written description. Automated annotation[11] is therefore preferable for these types of entities, to ease and reduce the task of humans.

As far as the acceptance of the conceptual model is concerned, this is apparently satisfactory: as detailed in [5], partners have the chance of proposing new concepts (through an appropriate KMap functionality) whenever the concept they have in mind for an annotation or search, is not found in the current ontology. During the last data entry task, in November 2006, only 3 new concepts have been proposed, and 9 more since April 2007, the official end date of the project. Overall, the ontology has about 2000 concepts.

The future of the KMap tool is meant to go beyond the conclusion of the INTEROP project, thus allowing further improvements and investigations. Three years is a short period to reach the NoE objectives in the domain of interoperability and it will be necessary to pursue these activities beyond the duration of INTEROP. This situation has motivated the creation of the I-ESA Virtual Laboratory, the V-Lab, which was officially launched in March 2007. The ultimate goal is the establishment of a permanent, self sustained, research organization that, capitalizing on the achievement of INTEROP, will continue in the production of high quality research in enterprise interoperability. The V-Lab initial resources will be mainly represented by human capital and infrastructures, in particular (but not only) the KMap, and the methodology adopted to build it. As far as the future of the KMap is concerned, the main ambition is to store the knowledge not only concerning NoE members, but also at a European level, so as to constitute a constantly updated picture of I-ESA research.

[11] Notice in the upper lines of Figure 5 that the annotations associated to the paper (in bold) have been obtained automatically, parsing the document.

5. References

[1] E. Hustad "Supporting end users: knowledge networking in global organizations: the transfer of knowledge" Proc. of 2004 SIGMIS conf. on Computer personnel research

[2] C. Bossen and P. Dalsgaad "Conceptualization and Appropriation: The Evolving Use of a Collaborative Knowledge management System" Proc. of 4th decennial Critical Computing, Aarhns, Denmark, August 20-24, 2005 pp. 99-108

[3] Y. Malhorta "Why Knowledge management Systems fail? Enablers and Constraints of Knowledge management in Human Enterprises" in *Handbook of Knowledge Mangement*, C.W. Holsapple ed., Springer-Verlag, 2002

[4] P. Velardi, A. Cucchiarelli and M. Petit "A Taxonomy learning Method and its Application to Characterize a Scientific Web Community " *IEEE Transaction on Data and Knowledge Engineering (TDKE)*, vol. 19, n. 2, February 2007, pp. 180-191

[5] P. Velardi, R. Navigli and M. Pétit "Semantic Indexing of a Competence Map to support Scientific Collaboration in a Research Community" 20th IJCAI (Joint Conference on Artificial Intelligence), Hyderabad India January 2007

[6] X. Su, S. Hakkarainen and T. Brasethvik "Semantic enrichment for improving systems interoperability" in Proc. of ACM Symp. on Applied Computing, Cyprus, 2004

[7] H. Cui, J. Wen, J. Nie, W. Ma "Probabilistic Query Expansion Using Query Logs" in 11th International World Wide Web Conference (WWW 2002)

[8] J. Bhogal A. Macfarlane and P. Smith "A review of ontology-based query expansion" *Information Processing and Management*, vol 43, issue 4, July 2007

[9] R. Navigli and P. Velardi. "Learning Domain Ontologies from Document Warehouses and Dedicated Websites", *Computational Linguistics* (30-2), MIT Press, June, 2004.