A Collaborative Semantic Web Layer to Enhance Legacy Systems

Alfio Gliozzo[1], Aldo Gangemi[1], Valentina Presutti[1], Elena Cardillo[2], Enrico Daga[2], Alberto Salvati[2], and Gianluca Troiani[2]

[1]Laboratory for Applied Ontology, ISTC-CNR, Rome, Italy
[2]URT-CNR, Rome, Italy

Abstract. This paper introduces a framework to add a semantic web layer to legacy organizational information, and describes its application to the use case provided by the Italian National Research Council (CNR) intraweb. Building on a traditional web-based view of information from different legacy databases, we have performed a semantic porting of data into a knowledge base, dependent on an OWL domain ontology. We have enriched the knowledge base by means of text mining techniques, in order to discover on-topic relations. Several reasoning techniques have been applied, in order to infer relevant implicit relationships. Finally, the ontology and the knowledge base have been deployed on a semantic wiki by means of the WikiFactory tool, which allows users to browse the ontology and the knowledge base, to introduce new relations, to revise wrong assertions in a collaborative way, and to perform semantic queries. In our experiments, we have been able to easily implement several functionalities, such as expert finding, by simply formulating ad-hoc queries from either an ontology editor or the semantic wiki interface. The result is an intelligent and collaborative front end, which allow users to add information, fill gaps, or revise existing information on a semantic basis, while keeping the knowledge base automatically updated.

1 Introduction

A legacy information system can be defined as any information system that significantly resists modification and evolution. Legacy systems are affected by problems such as lack of documentation, obsolete hardware and cost of maintenance software. On the other hand, most of the systems currently in place in large institutions and companies belong to the aforementioned category. Therefore, the new trend is to develop methodologies to allow the information to migrate from legacy systems to more flexible data structures that enable interoperability, reusability, and integration with the current Semantic Web (SW) technologies.

In this paper we propose a SW-based solution for the problem above, and we describe its application to the use case provided by the Italian National Research Council (CNR) intraweb. The main goal of this project is to develop a SW layer on top of the databases and web publishing systems that are currently in place at the CNR.

CNR is the largest research institution in Italy, employing around 8000 permanent researchers, organized into departments and institutes. Its 11 departments are focused

on the main scientific research areas. Its 112 institutes spread all over Italy, and are subdivided into research units, which are characterized by different competences, research programmes, and laboratories.

The overall structure of the CNR is then rather complex: departments express a "research demand", while institutes perform a "research supply". The activity of planning and organization of such a huge institution is then strictly related to that of matching the research demand and the research supply. It can be performed only by having in mind a global picture of the interrelations between the entities in such a huge network, an operation almost impossible without the semantic facilities provided by the recent ICT technology. As a matter of fact, only recently research units from different institutes and departments have slowly started some synergies, and at the cost of lengthy meetings and substantial push from a new set of 83 management units, called "progetti" (frameworks). 749 further units, "commesse" ("research programmes") have been created in order to direct local projects and synergies, and 704 researchers lead them as chief scientists. Each research programme is structured into local workpackages that channel research funds to institutes.

The complexity of this structure has proved to increase the potentiality for synergies within the CNR, but also the effort for its maintainance, monitoring, channelling of external funds and requests, etc. An adequate support for extracting and matching the competence-related knowledge that is scattered within local research units appears more and more relevant. In this context, information sharing and interoperability is crucial from a project management perspective. For example, in order to achieve a particular subgoal of a research project, it is often necessary to look for external qualified and highly specialized human resources, while having a clearer picture of the competences spread in the various departments of CNR would allow to avoid the use of external resources, since the probability of finding the desired profile among the internal members of the organization is very high.

SW technologies contain viable solutions to overcome the problems above, so that we have developed a prototype system that allows data migration from legacy databases to a wiki portal. Our system employs SW technologies such as OWL ontologies, reasoning systems, text mining tools, and ontology-driven wiki site management.

The general architecture of our system is described in Section 1, while the remaining sections analyze each component. Section 2.1 describes the migration process we implemented to unify the information spread into the different legacy systems in place at CNR, while Section 2.2 describes the domain ontology we developed and its further population. Section 2.3 focalizes on the text mining component we developed to induce `on_topic` relations among instances of the ontology, while Section 2.4 illustrates the deployment of the so obtained ontology into the wiki portal. Pros and cons of the proposed approach are described in the evaluation section, while Section 4 concludes the paper illustrating the new application scenarios opened by this work.

2 General Architecture and Information Workflow

As introduced in the previous section, the main goal of the information workflow presented in this paper is to enhance accessibility and interoperability of the information

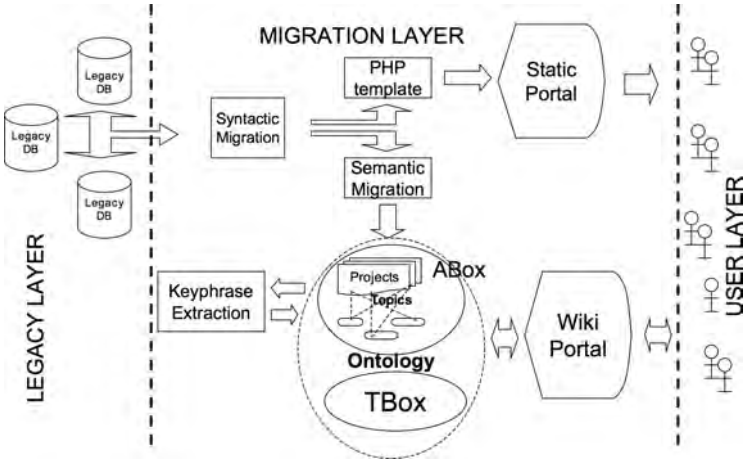

Fig. 1. The CNR semantic intraweb workflow and architecture

spread in different legacy systems, making it reusable by final users, and allowing them to interact with the resulting knowledge in a collaborative way. This section broadly describes the methodology and the algorithms we implemented to this aim. Details of each component will be illustrated in the appropriate subsections.

The overall system architecture is represented in Figure 1. The first component of our system performs a syntactic migration of the information spread into different CNR databases. The databases contain administrative and financial data, research organization data, and personal data of CNR employees. For privacy reasons, in this paper we focus on the semantic intraweb created for research organization data, which also supports a preliminary *expert finding* system. The result of this operation is a unified view of the overall CNR activities, including people involved, departments, research projects, and so on. This information is directly put at disposal of final users by means of an intraweb portal, presenting it by means of dynamically generated web pages. Details of this operation are described in Section 2.1.

Syntactic integration only provides a static and partial view of the information contained into the database. Basic semantic operations, such as expert finding and similarity reasoning, cannot be performed intellectually or with simple queries, given the huge amount of data actually contained into the databases. To this aim, we have perfomed a semantic migration, with the main objective of porting and then distilling knowledge from the information released by the previous integration. We have obtained a structurally richer representation, and on top of it reasoning and other cognitive operations can be performed. In particular, we have developed an ad-hoc domain ontology describing the information contained in the databases at a semantic level (see the TBox from Figure 1), and we have automatically populated it by implementig a semantic migration process that transfers the information from views on the database entries, to appropriate OWL-RDF code. Details of this step are illustrated in Section 2.2.

As a matter of fact, since a large part of the CNR data is composed of textual material (e.g. abstracts of research projects, descriptions of work done, internal reports, etc.), the benefits of semantics in this phase alone are limited to a more explicit, rigorous maintenance of information contained in databases. However, this is not yet something that "makes a difference" to final users, and can convincingly support the need for a real migration to the semantic (intra)web.

Therefore, we have followed the direction of adopting automatic text mining techniques to further enrich the structure of the knowledge base (and eventually of the ontology). In particular, we have acquired `on_topic` relations, by adopting automatic keyphrase extraction techniques. This operation is described in detail in Section 2.3. Based on this richer structure, the inferencing capabilities provided by OWL and SPARQL reasoners have been key to a substantial enrichment of the knowledge base (see Table 1).

Finally, we have adopted a wiki-based approach to deploy the knowledge base into a wiki site available to the final users. To this aim, we have exploited WikiFactory [4], an environment that can automatically generate a wiki portal mirroring an existing ontology. In addition, WikiFactory allows the final user to modify the ontology and the knowledge base (e.g. introducing new classes or properties, modifying the existing ones, introducing new individuals and property values, etc.) by simply modifying the generated wiki pages.

2.1 Syntactic Integration of Legacy Systems

All the components adopted for the syntactic integration are based on web services able to provide information from different knowledge sources into multiple standard formats, such as XML, RSS, SOAP etc. Each different legacy system can be accessed by means of an ad-hoc web service, providing information into a standardized XML format. These formats are transformed by applying appropriate templates. The system matches the XML retrieved from the web service to the template by means of XSLT datasheets, and returns the information organized in another (XML, RDF or OWL) format. To this aim, the system adopts different technologies (mainly Java and Ibm Web Datablade).

This simple strategy is used to provide information to external systems (for example, a web site of a CNR institute through RSS), to allow system integration, to retrieve knowledge expressed in RDF-OWL, and finally to build HTML pages for the current intraweb portal in place at CNR, already accessed by thousands of users.

The syntactic migration and integration of CNR data does not allow to reason over the knowledge contained in those data, because it only addresses data manipulation without any explicit assumption on the semantics (either in the linguistic or logical sense) of those data. For example, the reason why a certain table from a database is ultimately transformed into a certain part of a HTML page is not explicit (no logical semantics), and the associations between the terms used across the records of the databases are implicit (no linguistic semantics).

In order to add some semantics to the CNR data, we have adopted a twofold strategy: on one hand, an OWL conversion (see Section 2.2) of legacy data has been made by creating a template for each class from the CNR ontology, which is filled for each

instance extracted from a database. On the other hand, a text mining-based enrichment of semantic relations among terms from textual records (see Section 2.3) has been also performed, and then formalized in OWL. This twofold strategy provides logical and linguistic semantics to a substantial amount of CNR data.

2.2 Representing and reasoning on legacy information in OWL-RDF

In order to make the semantic migration effective, we have firstly developed an OWL ontology describing the CNR scientific organization. The ontology engineering process was rather simple as we took advantage from the existing XSD schemata defined to produce a set of HTML templates as presented in the previous section. Based on those templates, we produced a formal description of the domain ontology in OWL(DL).

Figure 3 shows the TBox of the ontology that we developed to describe the CNR scientific organization.[1] It encodes the relations between individuals of classes such as Researcher, City, Department, Research Programme, Institute and Framework. The actual expressivity exploits a fragment of OWL(DL): cardinality restrictions, property range and domain, disjointness, transitive and symmetric properties, etc.

Then we have populated the ontology by exporting RDF code from the syntactic module described in section 2.1. The second column of Table 1 reports the size of the ontology collected after the simple migration from legacy data. After basic reasoning, 3148 individuals and 14695 property values have been created in the CNR knowledge base. In the next section, we will explain why these data increase so dramatically after applying learning techniques to unstructured data.

As a tool for the ontology lifecycle and reasoning over the large ABoxes created after reasoning with topics extracted via NLP and LSA (see next section), after some testing, we have decided to use TopBraid Composer,[2] a commercial software based on the open source development platform Eclipse,[3] providing advanced visualization and querying tools, as well as efficient and substantially bug-free interaction between the Pellet[4] reasoner, the storage mechanism, and the interface.

2.3 Acquiring `on_topic` relations by applying text mining technologies

One of the main limitations characterizing typical OWL ontologies (at least for OWL1.0) is their weakness in modelling semantic proximity among concepts. For example, names of persons and organizations typically belong to different taxonomies, and just in a few cases they are actually related. On the other hand, semantic proximity is very well modeled by adopting geometrical models, such as the Vector Space Model [8] or contextual similarity techniques [2], elaborated in the Information Retrieval and Computational Linguistics areas.

Geometrical models for semantic proximity are at the basis of successful applications like search engines, while at the same time they constitute one of their bigger

[1] http://www.loa-cnr.it/ontologies/CNR/CNR.owl
[2] http://www.topbraidcomposer.com/
[3] http://www.eclipse.org/
[4] http://www.mindswap.org/2003/pellet/

ABox	**before**	**after**
Department	11	11
Institute	112	112
ChiefScientist	704	704
Framework	83	83
Programme	749	749
Workpackage	1166	1166
Topic	499	3148
City	66	66
Total individuals	**3393**	**6042**
owl:PropertyValue	**14695**	**158441**
TBox	**before**	**after**
owl:Class	11	11
owl:DatatypeProperty	33	33
owl:FunctionalProperty	10	10
owl:InverseFunctionalProperty	7	7
owl:ObjectProperty	28	48
owl:Ontology	1	1
owl:TransitiveProperty	2	2
owl:SymmetricProperty	0	10

Table 1. Ontology size before and after the reasoning and text mining procedures (original Italian class names are translated)

limitation. In fact, even if search engines are very powerful in providing information somehow related to the query, they are clearly not able to go deeper than that. For example, logic constraints cannot be imposed on the type of objects we are interested in during search, making the retrieval technology clearly inadequate for the new needs of the Web 2.0 and ultimately the semantic web.

On the other hand, logical models, such us those represented by domain ontologies, are characterized by approximately opposite properties. They allow us to easily perform semantic queries, for example by reasoning over OWL-RDF models, or by imposing a semantics over a query language like SPARQL, but they typically rely on manually designed descriptions written in some formal language, which are typically very costly and not available on a large scale.

In the context of several research projects, we are following the direction of fuzing empirical and logical approaches for knowledge representation, trying to integrate assessed text processing and information retrieval techniques with traditional knowledge engineering methods. The outcome of this integration would provide a much more powerful framework for knowledge representation, integration and acquisition to be used as a basic infrastructure for the SW.

To accomplish this goal, we followed the direction of acquiring `on_topic` relations to create new links among instances in the ontology. On topic relations are automatically induced via text mining, by analyzing the textual material connected to the instances in the ontology. We added a new class in the ontology, called topic, whose instances are

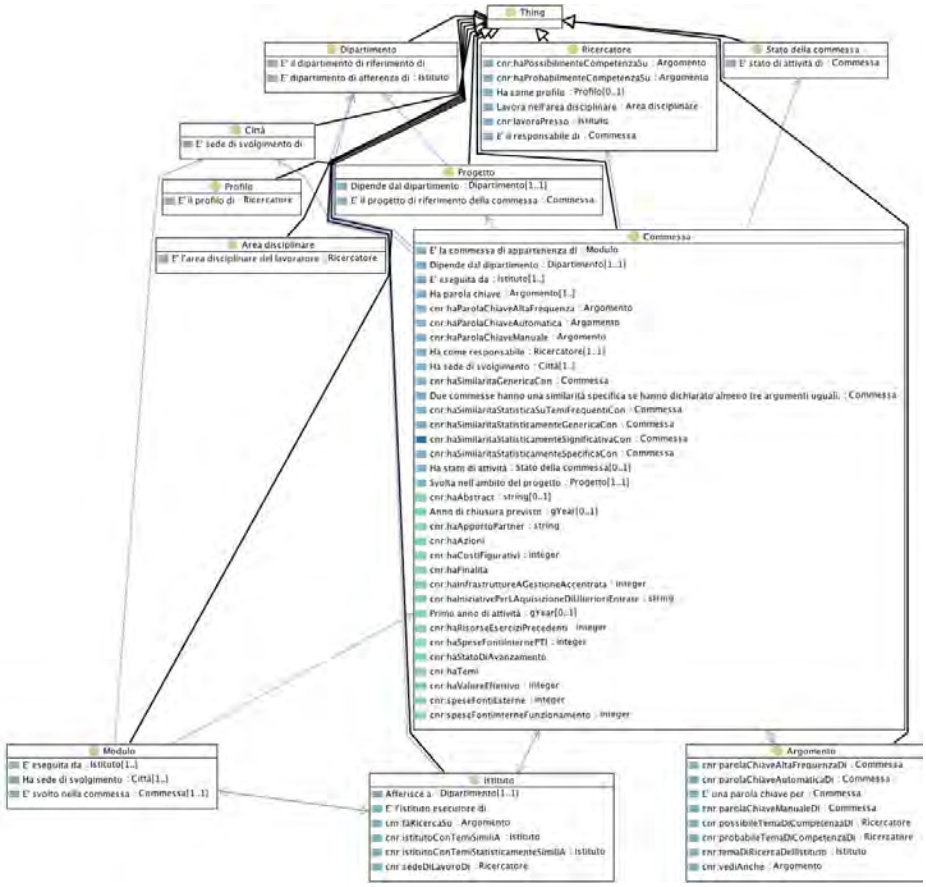

Fig. 2. TBox of the CNR ontology

different key-phrases extracted from documents, and we related them to entities in the ontology, such as research programmes, institutes, researchers, etc.

Extracting keyphrases from documents is a very well established technique in Natural Language Processing (NLP). In the literature, several methodologies have been proposed, ranging from applying supervised learning techniques [10], to pattern based approaches. Key-phrases are in general noun compounds, usually composed of 2 or 3 words, and can be identified by specifying syntactic patterns. Statistical measures are in general adopted to measure the internal coherence among words of the same terms and the distributional properties of the term as a whole inside documents in a corpus.

For the purposes of this work we implemented a novel approach for term extraction, based on Latent Semantic Analysis (LSA). Our approach identifies first a set of candidate terms from the whole document collections, by applying pattern-based approaches on the output of a Part of Speech tagger (e.g. all the sequences composed by Noun + determiner + nouns are candidate Italian terms). Then it filters out incoherent

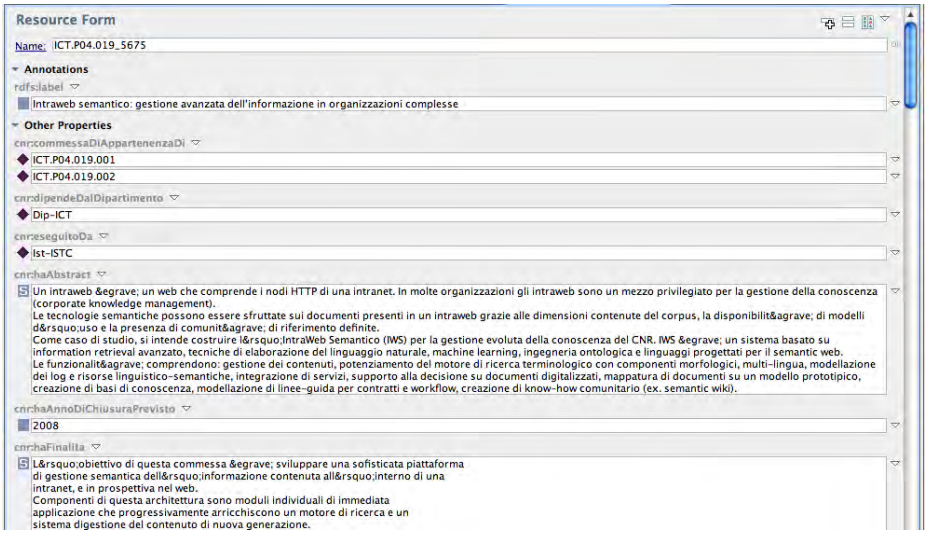

Fig. 3. An Instance of the Research Program class

terms by estimating the mutual information between the compound words for each term (e.g. “new economy” is a term, while “new English” is not a term). Finally, it represents documents and terms in the LSA space, a geometrical space in which the similarity between them can be estimated by taking into account second order relations. Keyphrases for each document are then selected by looking for all the neighboring terms of the document vector in the LSA space. This process is illustrated by figure 4. Further details on the LSA technique adopted are reported in [3].

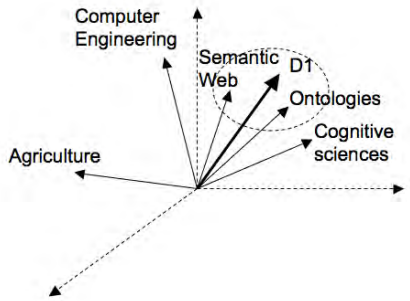

Fig. 4. Terminology Extraction in the LSA space

2.4 WikiFactory: a collaborative environment for knowledge representation, mantainment and upgrading

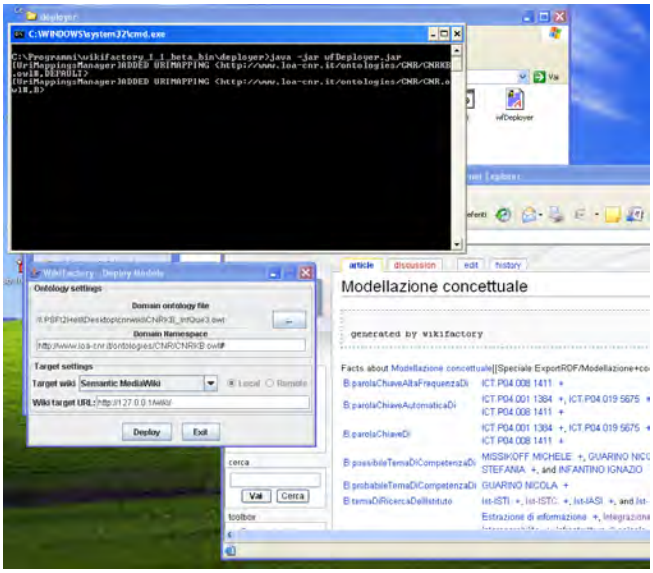

Fig. 5. Deployment on WikiFactory

As a final step of our workflow, we used the obtained ontology and its associated knowledge base in order to build a web portal based on a semantic wiki platform. To this aim, we exploited WikiFactory [4]. WikiFactory is a server application that takes as input an OWL ontology (including individuals and facts) and automatically deploys a semantic wiki-based portal. Wiki applications share the same basic philosophy of open editing and provide asimple text-based syntax for content editing. Currently, the most popular semantic wiki is Semantic MediaWiki [12] (a MediaWiki [6] extension), which enables semantic features such as defining categories, relations and articles, which corresponds to OWL classes, properties, and individuals, respectively. Semantic MediaWiki allows users to define queries by supporting a subset of SPARQL [7]. Although Semantic MediaWiki could be used in order to import our ontology and associated knowledge base into a semantic wiki site, we decided to use WikiFactory, because it provides additional features that are key to our case study requirements:

- It maintains the synchronization between the underlying ontology and the wiki content: this means that users can navigate and evaluate the ontology, and directly modify it from the wiki pages.
- For each wiki page, WikiFactory provides users with suggestions on the usage of semantic relations: users are not supposed to know all defined relations and the way to apply them, they can rely on the “light” reasoning capability of WikiFactory that include in each page all applicable semantic relations.

– It enables users to handle simple restrictions: each suggested relations is associated with a link that enables users to express some restriction on that relation.

For the sake of our case study, WikiFactory synchronization capability and user support for applicable relations has been particularly useful. Figure 6 shows an example of the visualization made possible by the wiki deployment.

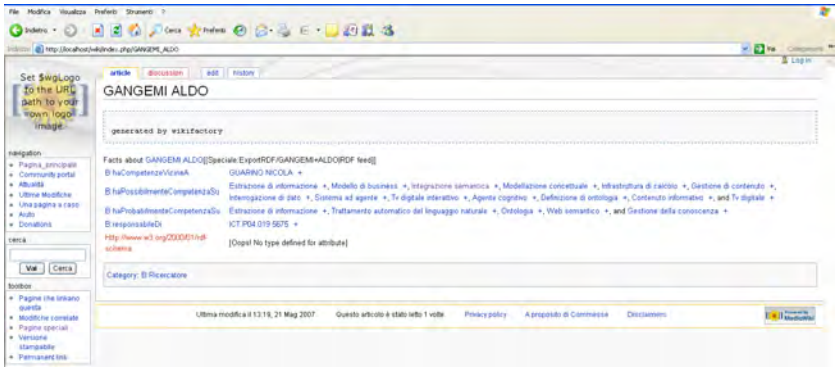

Fig. 6. Visualizatiion of an instance of the class researcher on the wiki

2.5 A remark on related work

It's not our intention to suggest here best practices on how to extract legacy data from relational databases: there exists a large literature on this topic, even in the semantic web proper (see e.g. [9], which also contains a large review of related work). Our focus is here on what to do with the extracted legacy data, and how to use NLP, semantic reasoning, and collaborative semantic tools to improve and enrich those data. Within our scope, there is related research, which differentiates from ours in scope and techniques used. For example, [5] focuses on using semantics for expert discovery, but does not apply it to a large organizational intraweb. [1] focuses on using a semantic wiki for organizational knowledge, but does not deal with knowledge enrichment as we do.

3 Evaluation

The evaluation of such a complex system is not trivial. Generally speaking, a conclusive judgment on the usefulness of the proposed technology could only be asserted on the basis of user satisfaction on the real use of the front end (the semantic wiki in our case). At the present stage, our application is entering the community usage, and we have formal plans on how to perform a user study in the next months.

For the sake of this work, we have evaluated the different modules independently, by adopting both quantitative and qualitative criteria. In particular, we have evaluated the accuracy of the keyword extraction system, and some functionalities, such as expert finding, and we can express a qualitative judgment on them.

3.1 Benefits from semantic migration

Semantic migration of legacy data augments sensibly the range of possibilities of the original information system, since turning it to a knowledge base enables consistency checking, inferences and semantic queries. On the other hand, the pure migration strategy adopted to populate the ontology provides little more than the original material, since only weak inferences can be performed, i.e. inverse relations that are materialized based on the ontology properties, and some SPARQL CONSTRUCT materializations that can e.g. be carried out to associate research programmes to manually inserted topics, and these to the chief scientists leading the programmes. This is an example of CONSTRUCT queries executed on the legacy knowledge alone (the original Italian vocabulary is translated here). The first query constructs owl:PropertyValues between researchers and the topics they probably have competence on (since they are responsible for the programme for which that topic has been manually inserted). The second constructs owl:PropertyValues that assert the similarity of competence between any two researchers that have probably competence on at least two common topics.

```
CONSTRUCT   { ?r hasProbablyCompetenceOn ?k }
WHERE       {
?r responsibleFor ?c .
?c hasManuallyInsertedTopic ?k .}

CONSTRUCT   { ?r1 hasSimilarCompetenceAs ?r2 }
WHERE       {
?r1 hasProbablyCompetenceOn ?k1 .
?r2 hasProbablyCompetenceOn ?k1 .
?r1 hasProbablyCompetenceOn ?k2 .
?r2 hasProbablyCompetenceOn ?k2 .
FILTER (?k1 != ?k2) .
FILTER (?r1 != ?r2) .}
```

Unfortunately, manually inserted topics are just a few (see Table 1), and they hardly co-occur in different programmes, so that very few property values have been inferred between researchers and topics, or between similarly competent researchers. This finding suggests that whenever a legacy database mostly contain string-based, non-structured data, migrating them to semantic technologies does not necessarily starts a virtuous circle of knowledge enrichment.

3.2 Key-phrase extraction

Extracting terminology from domain specific texts is a very well assessed technique in Natural Language Processing [11], and the present state-of-the-art algorithms for this task are highly accurate, achieving precision in general higher than 0.8. The key-phrase extraction problem is more complex, as it requires to associate relevant terms (i.e. keywords) to documents. For this purpose, we have randomly sampled a set of 30 instances from the class ResearchProgramme, and we asked a domain expert to

manually create a "Gold Standard" set of ON_TOPIC associations connecting them to appropriate keywords. [5]. The result is a list of keywords associated to each programme, as illustrated in Figure 7. Then we compared the Gold Standard annotations with those provided by the automatic system, measuring (micro)precision and recall, obtained as the total number of (manually assigned) relations that have been actually matched by the automatic system respectively divided by the total number of assignments made by the system, and by the human, obtaining precision 0.66 and recall 0.51. These results are very encouraging, since the agreement measured on the same task is around 70%. In addition, our system is totally unsupervised and can be easily ported to different tasks and domains, making our technology for extracting on_topic relations widely applicable at very low costs.

ID COMM.	KEYWORD 1	KEYWORD 2	KEYWORD 3	KEYWORD 4	KEYWORD 5
938	DIVERSITA' GENETICA	RESISTENZA A STRESS	ALBERO FORESTALE	MARCATORE MOLECOLARE NEUTRALE	CARATTERE ADATTIVO
973	FONDALE MARINO	RILIEVO ACUSTICO	MACCHINA OPERATRICE	PERCEZIONE SONORA	CRESCITA DI FILM SOTTILE
999	DIRETTIVA EUROPEA	QUADRO LEGISLATIVO	TARGET DI QUALITA'	QUALITA' DELL'ARIA	LEGISLAZIONE ITALIANA
1004	ECOSISTEMI	POPOLAZIONE ITTICA	DISPOSITIVO ELETTROACUSTICO	ATTIVITA' DI DIVULGAZIONE	COLONNA D'ACQUA
1054	ESTRAZIONE DI PARAMETRI	REALIZZAZIONE INFRASTRUTTURA	GRIGLIA COMPUTAZIONALE	TELERILEVAMENTO	ELEBORAZIONE DI IMMAGINI
1066	SONDE FLUORESCENTI	PRODUZIONE DI ANTICORPI	CARATTERIZZAZIONE BIOCHIMICA	KIT DIAGNOSTICI	SALUTE UMANA
1085	RILASCIO DI GLUTAMMATO	MUTAZIONE PUNTIFORME	CELLULE GLIALI	MORBO DI PARKINSON	NEURONI DOPAMINERGICI

Fig. 7. Gold Standard Keyphrases associated to different programmes

3.3 Reasoning on the acquired knowledge

Going back to the sample CONSTRUCT queries shown in Section 3.1, once many more topics are available and are associated to the research programmes described by the texts from which the topics have been extracted, the result of those queries becomes meaningful. In fact, since all research programmes are now associated with many topics, most researchers can be now said to be associated with several topics, as inferrable from a new version of the first CONSTRUCT query, which populates the hasProbablyCompetenceOn property values through the statistically inferred topics:

```
CONSTRUCT   { ?r hasProbablyCompetenceOn ?k }
WHERE       {
?r responsibleFor ?c .
?c hasHighFrequencyStatisticallyInferredTopic ?k .}
```

[5] We concentrated on the class ResearchProgramme since a lot of text is usually contained in the ABSTRACT and GOALS fields of each individual, as illustrated by Figure 2.

Since researchers have now more topics, which they can be competent upon, researchers can be associated on a richer competence similarity basis, thus originating a potential social network, needed to achieve the expert finding task. For example, by firing a revision of the second CONSTRUCT query:

```
CONSTRUCT   { ?r1 hasSimilarCompetenceAs ?r2 }
WHERE       {
?r1 hasProbablyCompetenceOn ?k1 .
?r2 hasProbablyCompetenceOn ?k1 .
?r1 hasProbablyCompetenceOn ?k2 .
?r2 hasProbablyCompetenceOn ?k2 .
FILTER (?k1 != ?k2) .
FILTER (?r1 != ?r2) .}
```

As an example, one of the authors has discovered his own social competence network, made of Aldo Gangemi, Nicola Guarino, Michele Missikoff, Mario Mango Furnari, who are actually the chief scientists that work on semantic web and semantic integration at CNR. The measure of the population performed is found in the third column of Table 1: from 499 to 3148 topics, and from 14695 to 158441 property values. The precision of the results, based on the first evaluation sessions on a sampling of topics that are closely related to authors' competence, is very good.

4 Conclusion and Future Work

In this paper we have described the application of semantic web technologies to the problem of enhancing knowledge extracted from organizational legacy systems. We integrated the information from different databases that describe the scientific organization of CNR into an OWL ontology, and enriched the resulting knowledge base by automatically learning on_topic relations based on the analysis of the textual fields from the databases. Finally, we performed reasoning on the top of the learnt knowledge, largely expanding the ontology by inferring new relations through the materialization of SPARQL queries. In particular, we concentrated on the expert finding problem, by inducing competence proximity relations between researchers.

The resulting knowledge base has been deployed in a collaborative environment, generated by the WikiFactory tool, allowing communities of users to access the knowledge base and to modify the information there contained, for example by cleaning data, correcting wrong property values inferred on top of automatically learnt topics from the key-phrase extraction system, by adding new instances and relations among them, etc.

This paper has focused on the first stage of the *competence finder* project at Italian National Research Council, a research programme involving several departments and research units. As future work, we will further study the integration of topicality information into structured knowledge bases, for example by formalizing continuous distances within the ontology as a way to reason over semantic proximity. We are also going deeper in integrating text processing and ontologies, for example by applying named entity recognition and text categorization tools on the textual fields. Another

parallel development direction is to make user studies on the collaborative environments for annotating and validating the results of automatic text processing techniques. To this aim, we are further developing the WikiFactory platform, concentrating on its compatibility with more sophisticated OWL constructs. Finally, we are going to propose the overall architecture presented in this paper as a general industrial solution for knowledge management, being the verticalization effort limited to the engineering of new domain ontologies, and leaving the remaining components as domain independent.

Acknowledgments

We are grateful to the members of the NeOn consortium who contributed to the NeOn vision being funded by the European Commission 6th IST Framework Programme. Further information on NeOn is available on http://www.neon-project.org.

References

1. Soeren Auer, Thomas Riechert, and Sebastian Dietzold. Ontowiki - a tool for social, semantic collaboration. In *5th International Semantic Web Conference, LNCS 4273*, 2006.
2. I. Dagan. Contextual word similarity. In Rob Dale, Hermann Moisl, and Harold Somers, editors, *Handbook of Natural Language Processing*, chapter 19, pages 459–476. Marcel Dekker Inc, 2000.
3. A. Gliozzo. *Semantic Domains in Computational Linguistics*. PhD thesis, University of Trento, 2005.
4. Angelo Di Iorio, Valentina Presutti, and Fabio Vitali. Wikifactory: An ontology-based application for creating domain-oriented wikis. In *ESWC 2006*, pages 664–678, 2006.
5. Steven Kraines, Weisen Guo, Brian Kemper, and Yutaka Nakamura. Semantic web technology for expert knowledge sharing and discovery. In *5th International Semantic Web Conference, LNCS 4273*, 2006.
6. *The MediaWiki Website*. http://www.mediawiki.org/wiki/MediaWiki. Viewed on November 30th, 2006.
7. Eric Prud'hommeaux and Andy Seaborne. SPARQL query language for RDF, 2005.
8. G. Salton and M.H. McGill. *Introduction to modern information retrieval*. McGraw-Hill, New York, 1983.
9. Jinmin Tang, Chunyin Zhou, Huajun Chen, and Yimin Wang. From legacy relational databases to the semantic web: an in-use application for traditional chinese medicine. In *5th International Semantic Web Conference, LNCS 4273*, 2006.
10. P. Turney. Learning algorithms for keyphrase extraction. *Information Retrieval*, 2(4):303–336, 2000.
11. P. Velardi, R. Navigli, A. Cucchiarelli, and F. Neri. *Ontology Learning from Text: Methods, Evaluation and Applications*, chapter Evaluation of OntoLearn, a Methodology for Automatic Learning of Domain Ontologies. IOS Press, 2005.
12. WikiMedia. "Semantic MediaWiki". http://meta.wikimedia.org/wiki/SemanticMediaWiki. Viewed on November 30th, 2006.