Semantic Scout: Making Sense of Organizational Knowledge

Claudio Baldassarre, Enrico Daga, Aldo Gangemi, Alfio Gliozzo, Alberto Salvati, Gianluca Troiani

Semantic Technology Laboratory of ISTC (CNR-Italy)

Abstract. Knowledge takes many forms in large organizations, and a unique opportunity exists to perform substantial integration of heterogeneous knowledge through semantic technologies. We present a sustainable method to create and maintain a data cloud that provides added value to an organization, while not interfering with existing practices. Our method shows one of the first application of knowledge-centric data access, following a web 3.0 paradigm. A use case has been implemented in a large research organization, based on explicit requirements. RDF-OWL datasets generated on the basis of a highly modular, pattern-based ontology are created, enriched by means of inferences and NLP techniques, and are integrated with linked open data. They are presented in different interaction modes that embrace important tasks such as navigation and search of organizational knowledge from any point, expert finding, competence matching etc. The tools implemented have been submitted to end-users for a task-based evaluation.

1 Introduction

An information system for organizations is traditionally thought as a mere technical tool for automation and management of administrative activities. In a scenario where semantic technologies are consistently proving that this idea is too restrictive, we want to reinforce the semantic web vision of *aggregative* information systems. We present the *Semantic Scout*, a software framework that offers semantic support to functionalities such as competence finding, social network discovery, etc.

The need for the *Semantic Scout* is motivated by the quest to provide a flexible decision making support within large organization, and in particular to support expert finding and project management. This is a common requirement within any organization with many stakeholders who are required to work in synergy, and to exploit internal resources, before looking for external competences. The hypothesis at the basis of this work is that the use of semantic technology, and in particular semantic search, automatic text categorization, linked data and ontologies, can make that requirement more easily achievable. In principle, the hypothesis is sensible for two reasons: firstly because semantic technology decouples knowledge from implemented systems, so that data can be consumed in ways closer to specific requirements or new scenarios; secondly, because semantic technology explicitly represents the entities of an organization, which gather an own identity: such identity enables simple and effective data aggregation procedures, and nicely matches the way humans *refer* to relevant things in their environment. A conceptual level that is close to human knowledge management is additionally provided by explicit *conceptual schemata* for the data (ontologies) [3].

In general, semantics improves the flexibility and adaptability of the systems, reducing the problems related to legacy and inconsistent data access, while augmenting the overall productivity. For example, the system described in our use case can be adapted to new requirements by simply changing the way the data are accessed, in a fully transparent and system-independent way.

Part of this work builds upon the results presented in [4], where the authors introduce an approach to migrate legacy data, in the domain of a large research institution, to a format that fosters interoperability and re-usability (RDF/OWL). Consistently with [4] we analyze the case of the *Italian National Research Council* (CNR)[1], and capitalize the capability acquired to integrate information from different databases into an OWL knowledge base (KB). At the same time, we redefine the target goal from [4], expanding the request for tools that supports *organizational research management* both for internal needs, and for opening organizational assets and data to the external world. By *asset* we mean humans, departments, research programs, scientific production (publications, patents), dissemination activities, etc. The objectives pursued by this work include:

- to describe a methodology that spans from an easy and rationalized integration of existing information sources in a variety of formats and media, to appropriate ways to consume the new integrated datasets;
- to improve information exchange and retrieval within and outside of an existing organization;
- to develop a powerful cognitive support for strategic decision makers;
- to reinforce collaboration within the organization.

In section 2 we depict the software architecture of our system when applied to the CNR use case, together with an explanation of the main aspects of the methodology to implement the *Scout* framework. The following sections reflect a more detailed presentation of our general methodology as illustrated by Figure 1. Firstly, we identify the data sources and analyze them in order to figure out the proper ontology able to semantically describe their content; this is described in section 3. Then we perform a reengineering process on the data, as described in section 4. The next step is publishing data, texts and ontologies developed so far on the semantic web, by following the linking open data paradigm: this is described in section 5. Once the data have been represented semantically, it is easy to design applications exploiting data according to different requirements. This is described in a section about data consumption 6. Finally we present an evaluation of the *Semantic Scout* (sec. 7), the related works (sec. 8) and the conclusions (sec.9).

2 Methodology and Software Architecture

The CNR organization presents a fairly complex network of information sub-systems (e.g. accounting, personnel-related, scientific projects and publications, administration documentation, etc.) maintained by different parties. Moreover, there are a number of internal services/procedures (e.g. plan management, contracts repository, activity economic balance etc.) that hardly integrate and interoperate. In [4] the authors explain

[1] http://www.cnr.it

a possible way to overcome these limitations, by designing an OWL knowledge base dense with relations among the main concepts of the CNR domain. We introduce here another aspect: the heterogeneity of user groups like administration, researchers, technicians, executives etc. People belonging to any of these job roles require mechanisms for fetching the information that fits their working style and daily tasks.

The analysis of the CNR user contexts led to the formulation of five core functional requirements to be addressed in order to successfully tackle the problem of managing organizational knowledge supported by semantic technologies:

1 - Browsing the network of organizational resources: requires the capability to traverse the entire collection of resources seamlessly crossing different domains (e.g. human resources, research programs, scientific production, dissemination activities etc.)

2 - Expert Finding: requires the capability to materialize, on demand and in one place, the relevant information about who in CNR is involved in some research or technological context. This activity can be assimilated to performing a sub-network extraction from the network of organizational resources (1).

3 - Semantic search of organizational resources: requires the capability to perform a keyword based search, closer to a classical Google-style search, against the resources in the organization KB (1). In other words, the search results for the user consist in *entities whatsoever* rather than documents only.

4 - Enriching the network of relations among the resources: requires the capabilities to discover degrees of similarity among the resources in the organization (e.g. researchers, institutes, competences, research fields), and to instantiate new relations among them. This requirement extends and supports (1) (2) and (3).

5 - Linking the organizational resources to Web resources: requires the capabilities to instantiate relations between entities belonging to the organization, and entities belonging to knowledge bases available on the Web (e.g. DBpedia[2]).

For what we presented so far (i.e. scenario description and user requirements), we can wrap our concerns into two main requests: (i) on the one hand we are required to make data interoperable, and (ii) on the other hand we are required to keep a sufficient level of specialization when designing information access for a wide range of data consumers, human or machine agents. Such an articulated context includes a spectrum of aspects ranging from systems for persistent data storage, to the tools provided to each user group in order to consume the data relevant to their activity. Figure 1 depicts the five types of methods applied to design and implement the Semantic Scout. The methods are described herewith:

Sources: sources to be reengineered include: (i) *legacy databases*, which are reengineered by following mainstream components for schema transformation and ontology population from databases, as well as specialized patterns for schema exception handling (realistic databases are far less clean than in the idealized situation); (ii) *large textual records* within databases, which deserve to be treated differently, e.g. creating specialized ontology entities to represent them: large textual records are

[2] http://www.dbpedia.org

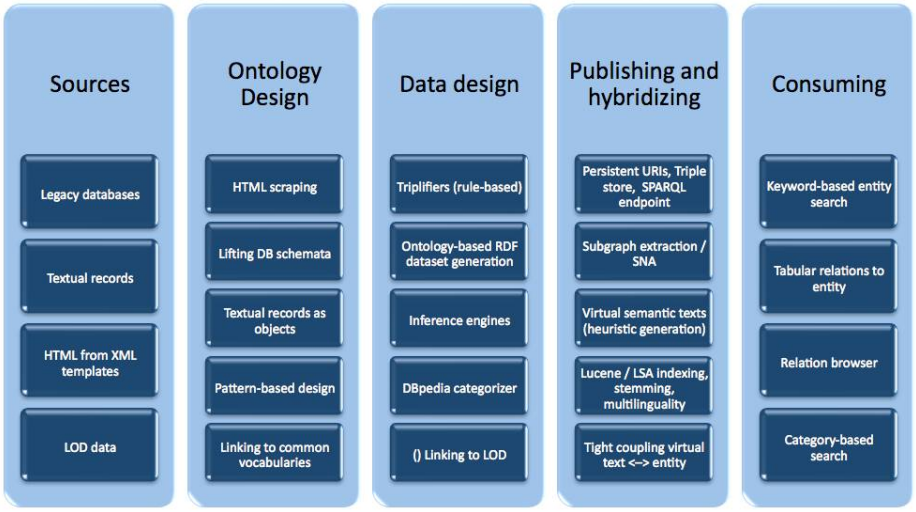

Fig. 1. Semantic Scout Methodology

specially important to build *textual representatives* of the entities, and to facilitate the hybridization of ontology engineering and information retrieval techniques; (iii) *HTML structures* from XML templates, which are a primary source for up-to-date user-oriented views over database data: these are specially useful for ontology design; (iv) *Linked Open Data* [5] from the Web, to be later linked to organizational ontologies and data.

Ontology Design: the methods used for ontology design include: (i) *HTML scraping* in order to derive user-oriented views over data and schemata; (ii) *DB schema lifting* in order to generate the backbone ontology for DB data; (iii) *textual records boosting* in order to create textual objects that will be linked to organizational entities, and used to perform semantic search; (iv) *pattern-based design* in order to create a modular ontology that fits the modelling requirements requirements, e.g. coming from the scraped HTML templates; (v) *linking to common vocabularies* in order to make the organizational ontology interoperable with external ontologies.

Data Design: methods used for data design include: (i) rule-based *rdf-izers* to convert legacy data to RDF, according to the OWL ontology patterns and modules created during ontology design; (ii) *inference engines* such as DL classifiers, rule and SPARQL engines, etc. in order to generate novel RDF triples; (iii) a *text categorizer* to create associations between (the textual representatives of) organizational entities and topics, e.g. DBPedia categories; (iv) *linked data matchers* to link organizational data to linked open data at the data level.

Data Publishing: techniques for publishing and hybridizing data include: (i) *URI schemes, triple stores, SPARQL endpoints* to maintain semantic datasets; (ii) *subgraph extraction and social network analysis* in order to provide synthetic views over the semantic graph induced by the linked RDF-OWL datasets; (iii) *heuristical generation of textual representatives* in order to maintain a textual counterpart to key orga-

nizational entities, e.g. papers for researchers, official descriptions for departments, etc.; (iv) *(multi-)linguistic and indexing techniques*, including LSA indexing, to perform basic and advanced search over textual representatives.

Consuming: technology for consuming organizational knowledge includes: (i) *keyword-based entity search*; (ii) *table- or matrix-based presentation* of relations and attributes of entities; (iii) *graphical browsing* of relations among entities; (iv) *category-based search* of entities.

While the research aspects in figure1 give directions along the methodological dimensions, the functional requirements also drive the design of the *Semantic Scout* software architecture. In figure 2 we have depicted the distribution of the functional components among the architectural layers: an infrastructure of components entirely based on semantic technologies, where we move from the idea of a single data source designed for one client application, as presented in [4], to a service oriented architecture (SOA). Web services allow to integrate functionalities of existing systems, and to build new lightweight clients that enable easy fetching and data consumption from a same underlying KB. From a general perspective, the architecture is deployed considering the three logical layers typically used by any application to organize the functional components of the system: data layer, engineering layer and UI layer. This distinction reflects a good practice in software engineering following the actual trend in semantic web applications [6].

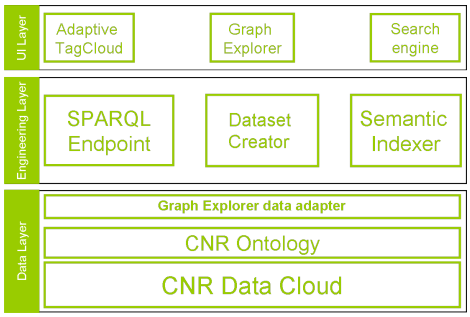

Fig. 2. Architecture

Next sections from 3 to 6 reflect the organization of methods given in figure1, and contain the description of how we achieved the realization of the functional components in Fig. 2.

3 Sources

Although during the analysis of the organization information systems we have run through different ways to expose data, from Internet web portals, to text based documents, the original sources of these data mostly reside in database structures. The

databases are hosted in a distributed fashion, inside the departments to which their maintenance is assigned. The domains that they cover are categorized as:

Organizational: Departments and Institutes inner structure information; Basic activities, Laboratories, Research Units, and International activities. In this set there are also International projects and Partnerships description, for example Spin-offs. This set of repositories belong to the intranet application for structures data management (called "Gestione Istituti", "Gestione Dipartimenti");

Research activities: Institutional research projects and products of research (e.g.journal articles, papers, books, patents etc.). This set of repositories belong to the application for managing the research plans ("Piano di gestione preliminare"), research results ("Consuntivi") and of the research activities ("GeCo - Gestione Commesse").

Administration Contracts, and Statistics; this set of repository belong to the administration management systems.

People: Employeers, Researchers, Technologists, Administratives; Collaborators, and Consultants. This set of repositories belong to the Employeer's management system.

Textual descriptions: Missions of departments, project descriptions, CVs, competence resumes, patent abstracts, publication abstracts.

Not all the data contained in these repositories are relevant to the objectives of producing an integrated organizational management system, hence we need data preparation to produce table views to be further queried. The consistent adoption of the same technology for the databases allowed to extract the data adopting template-based scripts using SQL language. On the other hand, the semantic interpretation of extracted data relies on the analysis of the existing interaction patterns by which the users access and consume the data (e.g. forms in the web portal); this is detailed in section about ontology desing.

4 Ontology Design

The first component of our system performs the reengineering of CNR databases containing administrative and financial data, research organization data, project, publication, and personal data. This component implements the ontology layer of the architecture (Fig. 2).

The reengineering process consists of four major steps: schema reengineering, script-based extraction, dataset generation, and KB evolution. A parallel enrichment process consists of: (1) inference-based dataset generation; (2) datasets created out of NLP-based extraction of implicit associations, and (3) datasets created from semi-automatic linking to Linked Open Data datasets.

A crucial phase in porting databases to semantic datasets is the extraction of the schema. Although several automated procedures exist to transform database schemas to ontologies, the results are usually quite poor when applied to databases that have been evolving for years in large organizations. The reasons for that low quality include the independent evolution of the physical schema of the database with respect to the conceptual schema used at design time, and the "pragmatic" tuning operated on the

physical schema in order to solve local issues emerging during the use of the database. In order to overcome this problem, some methods (e.g. [2] propose to "embed" ad-hoc queries to databases into annotations to the elements of an ontology.
While in a distributed context such ontology can be provided for particular tasks, or even on-the-fly, in case of a single organization like CNR, it is advisable to attempt the construction of a shared ontology. However, since the physical schemas of CNR databases are degraded, we have applied a method for *requirement-based ontology design* that focuses on the actual user consumption of the databases.
In the case of CNR, user consumption is currently ensured by means of HTML pages that are generated on-the-fly by running dedicated scripts on the databases, and by filling 61 dedicated HTML templates with the extracted data. Those scripts play the same role as the embedded queries to databases, and can therefore be reused for porting databases to semantic datasets.
Pattern-based ontology design [7] tries to define the boundaries of an ontology on the basis of explicit requirements provided by users or extracted from reference resources. Requirements are normalized and used as *competency questions*, and an ontology "pattern" is built for each competency question, and has been used as a module of the CNR ontology. In the case of CNR, each HTML template has been considered as a requirement.
HTML templates are structurally and conceptually similar to microformats, consequently, for each HTML template, we have tried to encode a module of the CNR ontology. As usual in realistic projects, the requirements have been massaged in order to obtain a modularization that complies to dependency issues:

- When a strong mutual dependency between two templates has been found (e.g. *departments* and *subdivision in programmes*), we have considered the union of them as a unique requirement
- When a template depends on another (e.g. *research lines* on *programmes*), we have considered the first as a specialization of the second
- When concepts are very general and occur sparsely in several templates (e.g. *localizations*, *subdivisions*, *categories*, etc.), they have been put into "upper" modules that are imported by most of the other modules

The final result is a network of OWL(DL) ontologies, currently consisting of 28 modules, partially ordered in an `owl:import` graph. The whole network includes 120 classes, 162 object properties, 134 datatype properties, 309 restrictions, 543 taxonomic axioms.[3]

5 Data Publishing and Hybridizing

Two rationales have guided the dataset creation according to the approach explained in 4:

1. Each dataset must be focused on collecting the instantiation of a single OWL property (i.e. obtaining an property-centric dataset);

[3] http://www.ontologydesignpatterns.org/ont/cnr/cnr.owl

2. A network of datasets is preferred to a monolithic collection of data materialized in a single file.

After the first rationale, we have generated more then 200 RDF datasets, each of them instantiating the value for a single property. All the collections have been made persistent in files with conventional names, so that the path to an RDF file is composed of a static base address *http://www.cnr.it/rdfgen/*; the prefix of the namespace of the property; the name of the property:

```
Path: base/ns-prefix/property-name.rdf
Ex: owl:ObjectProperty -> commesse:modulo,
Path to the RDF file: http://www.cnr.it/rdfgen/commesse/modulo.rdf
```

With the second rationale, we have generated we have a network of RDF datasets, using the `owl:imports` mechanism. A file containing a "bottom" ontology includes the import closure over the 200 datasets.

This approach has several benefits. First, the granularity of the extraction makes the work easier for debugging and testing w.r.t. to the original data schemas. Additionally, it is easy to manage user access policies, which can have several levels of privacy and sensitivity.

The property-centric organization of datasets support also the lifecycle of reengineered data because it is easier to identify smaller clusters of data to synch, than running the script mechanism on the entire data set, even when we know that a value for the properties is not going to change.

6 Consuming data

The set of functional components, together with the CNR data cloud, and the CNR ontology, are used as the *toolkit* for the Semantic Scout. This section is dedicated to *unfolding* each tool and to explain how and why they fit into the kit. In addition, we present the use cases where they are daily used by CNR people.

The Semantic Scout infrastructure includes an information retrieval engine. As described in section 2, starting from a known interaction pattern is beneficial to the users. Figure 3 shows the result page for the query:

```
{ ethics, sociology, collaboration, social network, reputation }.
```

Traditional information retrieval is performed on, and retrieves, only information objects, typically documents. The variety of semantic search performed by the Scout is still performed on documents, but retrieves *entities*.

Internally, the search engine (traditionally) indexes selected texts, which are however *textual representatives* of entities, generated at data design time by means of regular SPARQL CONSTRUCT queries over the heuristically relevant text data from datatype values in the RDF datasets (for example, publication titles and abstracts for persons). Heuristics is based on context, task, and available data.

Fig. 3. CNR Semantic Scout - Search Engine

This search design pattern is based on a semiotic assumption: each entity can have a typical, although context-dependent, textual representation.

The search engine is able to index both Italian and English text, and implements two types of search, Basic (i.e. keyword based) or Latent (i.e. based on statistical methods to represent texts into a cluster based representation similar to Latent Semantic Indexing). The user has the possibility to choose the desired modality of search before performing the query. In order to implement the multilingual search, we have used two different stemming algorithms for different languages (implemented by the Snowball Analyzer embedded in the standard distribution of Lucene). Latent search is based on Semantic Vectors [4].

In other words, the semantic search design pattern adopted by the Scout tightly couples information retrieval technology for basic search, and ontology design plus linked data to data management, reasoning, and actual consumption of data.

An additional functionality enables the Scout to enrich the emergent semantic social network of CNR with topics, and to link CNR data to DBpedia [1].

We categorize entities with topics by using a (novel, yet unpublished) text categorization system whose goal is to link documents to categories selected from the more than 500,000 categories present in DBPedia, which then provide a rich set of distinctions for the scientific subjects of interest in the CNR case study. The output of the categorizer has been represented in RDF by using the `subject` relation from the SKOS vocabulary, e.g.:

```
<> <http://www.w3.org/2004/02/skos/core#subject>

   <http://dbpedia.org/resource/Category:Knowledge_representation> .

<> <http://www.w3.org/2004/02/skos/core#subject>

   <http://dbpedia.org/resource/Category:Artificial_intelligence> .
```

The categorization data generated so far have been loaded in a dedicated dataset, and used to enrich the knowledge base. This is an example of the application of statistical techniques to enrich the knowledge base. In section 6.1 we show the usefulness of

[4] `http://code.google.com/p/semanticvectors/`

the categorization data for expert finding and semantic browsing of data. The *Semantic Scout* networks the entities of an organization, besides content objects. It is crucial that this approach is backed by tools that can support a proper presentation, and can be coherent with the idea of linked data. The *Graph Explorer*[5] is our choice for the prototype to effectively tackle the presentation of, and the interaction with, the datasets. We investigated other examples of RDF data browser and we found that the relation browser from Moritz Stefaner overcame the limitations of exposing the data in tabular format only, providing an appealing interface for our targeted user groups. We have re-worked some graphical elements in order to adapt them to the classes of data we want to display; these classes are CNR-ontology-concepts considered *hub* subjects of relations in the CNR ontology. Suppose a user looks for CNR researchers that have competence in

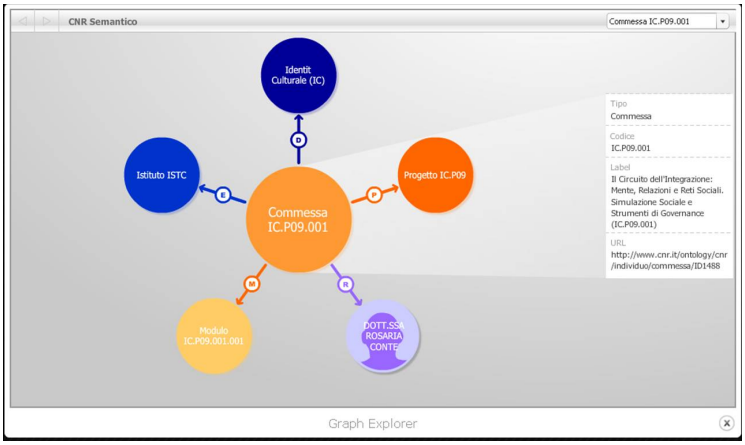

Fig. 4. CNR Semantic Scout - Graph Explorer

the topic *Semantic Web*. A search can be performed with different sets of keywords, but once entities are shown, a user can browse the rich knowledge e.g. in terms of relations between researchers and departments, other researchers, topics, publications, etc. This allows a deeper understanding of who is doing what, explaining how a researcher is involved in the Semantic Web. In addition, it allows us to find additional associated information. Most likely, this information will be very relevant to the user. Figure 4 shows an example of the *Graph Explorer*, with the focused node describing a project and connected to leading researchers, their workpackages, the participating institutes, and related departments. A panel on the right gives a description of the focused node.

We stress the idea of exploring, as opposed to searching, since the former leads to targeted resources along multiple paths, which could be previously unknown to the user; this is emphasized by the graph based representation.

[5] http://moritz.stefaner.eu/projects/relation-browser/

6.1 Expert Finding

Research on expert finding is typically performed on curated data or massive social web data. Our lightweight approach is based on the ability to materialize on demand, and in one place, the relevant information about who in CNR is involved in some academic or technological context. In [4] we explained how to enrich an initial set of relations between researchers, publications, workplaces, conferences, etc. to a much denser set of linking properties. The application of the hybridization processes explained in section 4 is intended to delegate almost most of the complexity of data matching, typical of a process of expert finding, to a reasoning procedure made at design time over the CNR datasets.

Expert finding with the *Semantic Scout* consists in using a combinations of the presented components. A sample scenario includes Carmelo Russo, a project manager in one of the CNR institutes, who is involved in a project on Social Knowledge for e-Governance. Carmelo has been introduced to the problems affecting the area and that the project is expected to solve: "*Reputation is a social knowledge, on which a number of social decisions are accomplished. Regulating society from the morning of mankind becomes more crucial with the pace of development of ICT technologies, dramatically enlarging the range of interaction and generating new types of aggregation. Despite its critical role, reputation generation, transmission and use are unclear. The project aims to an interdisciplinary theory of reputation and to modeling the interplay between direct evaluations and meta-evaluations in three types of decisions, epistemic (whether to form a given evaluation), strategic (whether and how interact with target), and memetic (whether and which evaluation to transmit).*"

Carmelo has been asked to acquire more information about possible researchers, consultants, showcase technologies, and publications related to the project topics.

Identifying the main topics of research Carmelo needs to focus his attention on a few research topics, and to use them as entry points in the network of resources of CNR. Analyzing the text from the the scenario description, an initial step consists in finding out what categories, e.g. from DBpedia, the project scenario can be associated with. Carmelo submits the text from the description to the *Text Categorizer*, and automatically finds the following main topics: *Ethics, Sociology, Collaboration, Social network, Reputation*.

Searching the CNR network of resources The extracted topics are representative key-terms of the problem description; they can be used to trigger a search in the CNR datasets by using the *Semantic Scout* search engine (see sec.6). Carmelo needs to find people involved in any of the areas named by topics, and preferably people who have experience of past collaborations directly (e.g. working on the same activity), or indirectly (e.g. publishing on the same research subject), and might belong to the same group (e.g. same department, institute). He can then trigger a search with input keys: { `Ethics, Sociology, Collaboration, Social network, Reputation` }, which returns a number of results faceted by **Persone**[6](e.g. researchers, consultants, directors), **At-**

[6] People

tività[7] (e.g. projects, workpackages, tasks), and **Strutture CNR**[8]. The faceting of results increases Carmelo's capability to select the scope of the results, and to set the entry points to browse the CNR network of resources, as explained in the next section.

In order to respond to a request for building a team of experts on the topics identified in 6.1, Carmelo needs to explore who, among CNR staff members, and how, is networked through them. The *Graph Explorer* exposes a network not just among people, but among the whole set of organizational assets. Since Carmelo is not familiar with any of the names listed in the **Persona** facet, or in the **Strutture CNR** one, he decides to enter the graph of CNR resources from the **Attività** facet, and chooses the item *Il Circuito dell'Integrazione: Mente, Relazioni e Reti Sociali. Simulazione Sociale e Strumenti di Governance* that looks relevant to his search.

Figure 4 depicts the graph of connections among the activity (central node) selected by Mario, and other CNR resources. Dr. Rosaria Conte holds a relation with the activity to be its key person (i.e. responsible). When the node of Dr. Rosaria Conte is selected to the center, her social network of co-authors is revealed. Starting from the network of names, Carmelo goes back to the **Persona** facet, to investigate those who have a higher score against the search performed in 6.1; the exploration makes it emerge the following list:

```
KEY PEOPLE (ranked): Dr. Rosaria Conte, Ing. Jordi Sabater,
Dott. Mario Paolucci, Samuele Marmo, Daniele Denaro,
Gennaro Di Tosto, Walter Quattrociocchi,
Francesca Giardini, Dott. Paolo Landri.
```

Checking the quality of results The people selected during the previous step have been found relevant to the keyword based search, and connected to a key person (Dr. Rosaria Conte), because she's their co-author in some publication; we expect that all of them share some research topic, and that they match the keywords used in the search. When opening the tabular views about each person (see 5), it is possible to read the subject of research associated with them; in the following we report some:

```
Ing. Jordi Sabater: Cognitive Science;
Dott. Mario Paolucci: Sociology, Psichology;
Gennaro di Tosto: Artificial Intelligence;
Walter Quattrocchi: Interdisciplinary Fields;
```

Most of those people cover research areas that are relevant to the search. On the other hand, some of the topics/keywords from the search remain uncovered. The same process from searching to quality checking is performed using only those topics/keyword found uncovered; the results found are:

```
Giuseppe Castaldi: Ethics;
Aldo Gangemi: Semantic Web, Knowledge representation.
```

By combining a few tools, Carmelo has been capable to collect information about researches and research leaders. A set of relevant publications is also available through the results in the **Attività** facet.

[7] Activities

[8] CNR administrative units

Fig. 5. Tabular view of data about CNR staff member Dr. Mario Paolucci

7 Functional Evaluation

The scenario exemplified in section 6.1 is in fact an existing project named "eRep" ended in March 2009, and whose web portal is available on line[9]. In the eRep project researchers of CNR have been active. We decided to adopt the description of a real case, in order to have a golden standard to compare the results when testing the *Semantic Scout*. This test was performed without an a-priori knowledge of the team from CNR involved in the eRep project, in order to avoid any kind of bias.

The CNR staff members involved in eRep are listed online[10], and out of 10 CNR researchers, we could match 6 people, among which the project coordinator (Dott. Mario Paolucci), plus a project member affiliated with another institution (Jordi Sabater Mir).

In the following, we analyze the aspects that we consider positively qualifying the *Semantic Scout*, and the reasons for missed matching of the remaining people in the list.

Accuracy: All the retrieved people scored among the first 10 in the result from the search engine; we remind that the search was performed with keywords considered topics relevant to the scenario/project description. Their relevance was proved when the same names where found to be part of the social network of Dr. Rosaria Conte, the leader of an activity considered of interest for Carmelo Russo, and returned among the results of the search.

Benefit of integrated data cloud: The entry point to the network of CNR resources has been an activity considered of interest for Carmelo Russo, and returned among the results of the search. In fact, reading the title of the activity triggered a cognitive process of spotting out similarity with the scope and aims of eRep project. This kind of workflow is made available thanks to the information hybridized in the CNR network of resources, as explained in section 4.

Accessibility and Interaction: Carmelo Russo was able to quickly identifying all the key people, by investigating the social network of Dr. Rosaria Conte through the Graph Explorer. The expert finding with the *Semantic Scout* starts with a keyword based search 6.1, and continues by navigating the faceted results, and the data cloud of CNR generated by integrating the distributed data sources 3. The access to this

[9] http://bit.ly/8mOr7Z

[10] http://megatron.iiia.csic.es/eRep/?q=node/36

network of resources is critical to the success of our approach in making sense of organizational knowledge. The Graph Explorer provides here the support for a good level of accessibility and interaction with the data network.

Completeness: The first search retrieved CNR member staff whose field of application was only partially covering the topics/keywords used as input. Carmelo Russo task was to build a team of experts in all of the identified disciplines, without knowing he already acquired the names of mostly all the people involved in eRep project. Since CNR staff members are explicitly related to their subject topics, it has been easy to determine the complementary fields to be searched in order to complete the expertise set needed to cover the project scope.

Reasons for unmatched researchers When comparing the result of the expert finding with the list of CNR members who participated in eRep, we have discovered that four people have not been included in our result set. One of them (Antonietta Di Salvatore) scored below the first 10 people in the list; the other three are Giulia Andrighetto, Marco Capenni, and Stefano Picascia. Giulia Andrighetto is not listed among the people relevant to the query, but belongs to the social network of Dr. Rosaria Conte. Both Marco Capenni and Stefano Picascia are known to our system, but they are neither reported among the people relevant to the search query, nor belong to the network of any of the other researchers. The reason is that they do not share any publications with them, most probably because they have a technician profile. We can safely conclude that the key players in a "dream team" for Carmelo Russo have been found by the Semantic Scout.

8 Related Work

Although this work is not meant to compete with existing corporate knowledge management systems, either distributed or integrated, we notice that all of the technologies we mentioned exist as isolated technological examples, and none of them is intended to cover the same class of problems we are interested in; similarly, the motivations and the objectives underlying those works are different from ours. We intended to prove that it is possible to employ off-the-shelf semantic components, and with a little integration effort, to obtain a nice prototypical toolkit for exploring the information assets of a big organization. It is anyhow worth mentioning that if possible competitors are not present in the open software world, there is a commercial framework called Vivisimo [11] that features a job assistant called *Mr. Stan*. *Mr. Stan* has functionalities similar to what we propose here. We have discovered this only recently. A closer look to *Mr. Stan* assistant shows that in fact it is not empowered with any semantic technologies, which does not make it a direct competitor to our approach. Moreover it has no sign of being a distributed system accessible via services and different kinds of clients.

9 Conclusions and Future Work

We have presented the practices, methods, and implemented components of a framework for integrating existing data and user requirements with semantic technologies in

[11] http://vivisimo.com/

a large organization. The use case is provided by the largest Italian research organization, CNR. Pattern-based ontology engineering and linked open data methods seem to be adequate to generate added value knowledge, simple decoupling of data gathering and consumption layers, and openness to data external to an organization. Among the critical issues, we mention privacy and provenance aspects, which are typically interlaced with internal practices and hierarchical responsibilities in an organization. Those complex interrelations are being studied for the linked open data initiative, where they prove to be non-trivial. On the other hand, within the intraweb of an organization, the same policies that apply to legacy data can be taken as received practices. Nonetheless, the increased dynamics and openness of organizational data pose specific problems, which will be extremely interesting to monitor in the next future, in the context of the NetwOrK project, sponsored by the CNR technology transfer office in order to increase the links between the CNR scientific network, and the industrial or business networks outside of it. Future work has two main objectives; evaluating in detail the user and functional tests, and enriching the number of components that can satisfy requirements such as: (i) social refinement of the CNR datasets through semantic wikis or content management systems, and social bookmarking, (ii) enriching the CNR datasets with new relations inferred from linking to other external data, and finally (ii) extending the capability of matching offer and public request for CNR competences.

Acknowledgements

This work has been supported by the CNR program *Semantic IntraWeb*, the *Semantic Scouting* project funded by the CNR Technology Transfer Office, as well as by the EU projects *NeOn*, funded within the 6th Framework Programme, and *IKS*, funded within the 7th FP.

References

1. Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBpedia – A Crystallization Point for the Web of Data. *Journal of Web Semantics*, 7:154–165, 2009.
2. Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, Antonella Poggi, Riccardo Rosati, and Marco Ruzzi. Data integration through dl-lite ontologies. In Klaus-Dieter Schewe and Bernhard Thalheim, editors, *Revised Selected Papers of the 3rd Int. Workshop on Semantics in Data and Knowledge Bases (SDKB 2008)*, volume 4925 of *Lecture Notes in Computer Science*, pages 26–47. Springer, 2008.
3. Aldo Gangemi and Valentina Presutti. Ontology design for interaction in a reasonable enterprise. In Peter Rittgen, editor, *Handbook of Ontologies for Business Interaction*. IGI Global, Hershey, PA, November 2007.
4. A. Gliozzo, A. Gangemi, V. Presutti, E. Cardillo, E. Daga, A. Salvati, and G. Troiani. A Collaborative Semantic Web Layer to Enhance Legacy Systems. In *Proceedings of the ISWC2007, Busan, Korea*, 2007.
5. Michael Hausenblas. Exploiting linked data for building web applications. http://swapp.org/pub/exploit-lod-webapps-IEEEIC-preprint.pdf, 2009. Stand 12.5.2009.
6. Benjamin Heitmann, Conor Hayes, and Eyal Oren. Towards a reference architecture for SemanticWeb applications. In *Proceedings of the 1st International Web Science Conference*, 2009.

7. V. Presutti and A. Gangemi. Content Ontology Design Patterns as Practical Building Blocks for Web Ontologies. In *Proceedings of the 27th International Conference on Conceptual Modeling (ER 2008)*, Berlin, 2008. Springer.