Rivista "Informatica e diritto", XXXVII Annata, Vol. XX, 2011, n. 1-2, pp. 453-476

Linked Open Data for the Italian PA: The CNR Experience

ALDO GANGEMI, ENRICO DAGA, ALBERTO SALVATI
GIANLUCA TROIANI, CLAUDIO BALDASSARRE*

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 ex-

* The Authors belong to the Semantic Technology Laboratory, ISTC-CNR. 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.

plicitly 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)[1].

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 two publications[2], 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 previous experiences[3] we analyze the case of the *Italian National Research Council* (CNR)[4], 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[5], expanding the request for tools that supports *organizational research management* both for internal needs, for opening organizational assets and data to the external world, and for assessing CNR research impact. By *asset* we mean humans, departments, research programs, scientific production (publications, patents), dissemination activities, etc. The objectives pursued by this work include:

[1] A. GANGEMI, V. PRESUTTI, *Ontology Design for Interaction in a Reasonable Enterprise*, in Rittgen P. (ed.), "Handbook of Ontologies for Business Interaction", Hershey, PA, IGI Global, 2007.

[2] A. GLIOZZO, A. GANGEMI, V. PRESUTTI, E. CARDILLO, E. DAGA, A. SALVATI, G. TROIANI, *A Collaborative Semantic Web Layer to Enhance Legacy Systems*, in "Proceedings of the ISWC2007" (Busan, Korea, 2007); C. BALDASSARRE, E. DAGA, A. GANGEMI, A. GLIOZZO, A. SALVATI, G. TROIANI, *Semantic Scout: Making Sense of Organizational Knowledge*, in "Proceedings of EKAW2010", 2010.

[3] C. BALDASSARRE, E. DAGA, A. GANGEMI, A. GLIOZZO, A. SALVATI, G. TROIANI, *Semantic Scout: Making Sense of Organizational Knowledge*, cit.

[4] Cfr. http://www.cnr.it.

[5] A. GLIOZZO, A. GANGEMI, V. PRESUTTI, E. CARDILLO, E. DAGA, A. SALVATI, G. TROIANI, *A Collaborative Semantic Web Layer to Enhance Legacy Systems*, cit.

- 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.

2. Requirements for a Large Organization

2.1. General Aspects

Large research organizations, like universities or public research institutions, are often composed of hundreds of research units spread all over the national territory, covering a wide range of research fields, and within a complex shared infrastructure. We can consider CNR, despite its own peculiarities, as a typical large organization, presenting a fairly complex network of information sub-systems (e.g. accounting, personnel-related, scientific projects and publications, administration documentation, etc.), often maintained by different parties. A number of internal services/procedures (e.g. plan management, contracts repository, activity economic balance etc.) cover a huge amount of applications, sometimes they share the information but more often they hardly integrate and interoperate. In the above cited article[6] the authors explain a possible way to overcome these limitations, by designing an OWL knowledge base with relations among entities from the main classes in the CNR domain. In the other[7] the authors consider the heterogeneity of user groups like administration, researchers, technicians, executives etc., by providing *ad-hoc* mechanisms for fetching the information that fits their working style and daily tasks on top of the knowledge base. In this work we deepen the description of previous work, and also concentrate on the "informative outside world", discussing how data publishing and sharing according to Linking Open Data good practices could be beneficial for the overall research management.

[6] A. Gliozzo, A. Gangemi, V. Presutti, E. Cardillo, E. Daga, A. Salvati, G. Troiani, *A Collaborative Semantic Web Layer to Enhance Legacy Systems*, cit.

[7] C. Baldassarre, E. Daga, A. Gangemi, A. Gliozzo, A. Salvati, G. Troiani, *Semantic Scout: Making Sense of Organizational Knowledge*, cit.

In this work we have considered the following general requirements for a large research organization:

- (open) data integration, sharing and reuse: the system should be able to integrate information form spared sources, considering the privacy level in relation to the data provenance and to the current user. This use case includes a way of linking the data to external sources (can be public data but also private, for example acquired repositories), the task of giving back the information to any unit (sub-system) which is part of the organization, and at the same time to publish public information (but not all the information) on the web, to support transparency and to fulfill Open Data requirements;
- competence finding and subjects coverage: the system should provide support for answering questions like: *is this research field sufficiently covered by the organization?* This use case can cover a demand from outside the organization but also help the management layer;
- impact of research: the system should provide support to analyze the results of research activity, answering to questions like: *what is the impact of the outcomes of a specific research unit?* This requirement is under research at the time of this publication.

2.2. Functional Requirements

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 other resources: requires the capabilities to instantiate relations between entities belonging to the organization, and entities belonging to other knowledge bases. These can be available on the Web (e.g. DBpedia[8]) or can be copyright-protected proprietary repositories (e.g. IEEE publications archives[9]) or, finally, are available on the *deep web*[10], such as the Google Scholar system[11];
6. publish public data in a machine exploitable way, to make it reusable by third-party systems: requires the capability to provide access to (only) the public information by anybody on the web[12];
7. giving back the data to the distributed research units: requires the capability to provide ad-hoc access to the knowledge according to specific privileges by each research unit.

3. METHODOLOGY

Main approach is based on decoupling of data and applications. In common close-world systems, the data layer and the application layer results fom the same requirement analysis, even if they are technically/physically separated. In this case, data do exist before and have is independent from the concrete task. This brings new challenging aspects.

For what we presented so far (i.e. scenario description and user requirements), we can wrap our concerns into two main requests:

– on the one hand we are required to make data interoperable, and

[8] See http://www.dbpedia.org.

[9] See http://www.ieeecss.org/main/electronic-publication-archive/electronic-publication-archive.

[10] M.K. BERGMAN, *The Deep Web: Surfacing Hidden Value*, in "Journal of Electronic Publishing", Vol. 7, 2001, n. 1.

[11] See http://scholar.google.com.

[12] We follow here the Linked Data paradigm, see http://www.linkeddata.org).

– 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. Fig. 1 depicts the five types of methods applied to design and implement the Semantic Scout. The methods are described herewith:

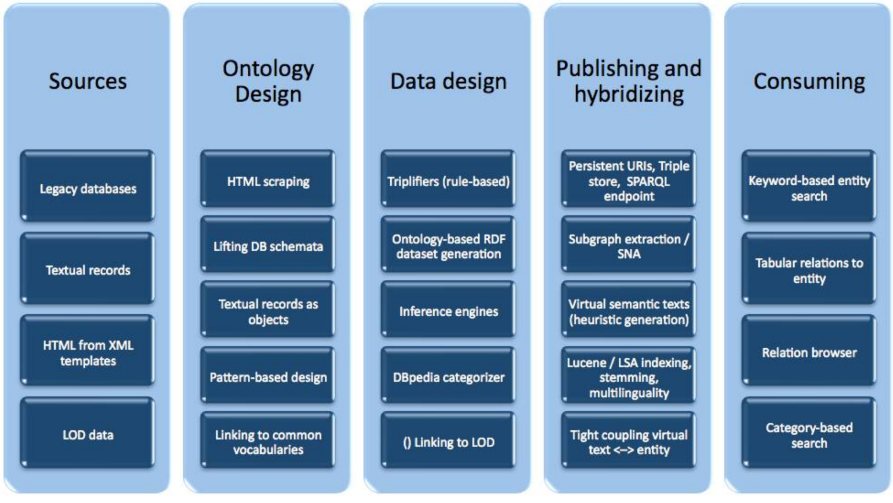

Fig. 1 – Semantic Scout Methodology

Sources. Sources to be reengineered include:

– *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);
– *large textual records* within databases, which deserve to be treated differently, e.g. creating specialized ontology entities to represent them: large textual records are specially important to build *textual representatives* of the entities, and to facilitate the hybridization of ontology engineering and information retrieval techniques;

- *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;
- *Linked Open Data* from the Web[13], to be later linked to organizational ontologies and data.

Ontology Design. The methods used for ontology design include:

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

Data Design. Methods used for data design include:

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

Data Publishing. Techniques for publishing and hybridizing data include:

- *URI schemes, triple stores, SPARQL endpoints* to maintain semantic datasets;
- *subgraph extraction and social network analysis* in order to provide synthetic views over the semantic graph induced by the linked RDF-OWL datasets;
- *heuristical generation of textual representatives* in order to maintain a textual counterpart to key organizational entities, e.g. papers for researchers, official descriptions for departments, etc.;

[13] M. HAUSENBLAS, *Exploiting Linked Data for Building Web Applications*, at http://sw-app.org/pub/exploit-lod-webapps-IEEEIC-preprint.pdf, 2009. Stand 12.5.2009.

- *(multi-)linguistic and indexing techniques*, including LSA indexing, to perform basic and advanced search over textual representatives.

Consuming. Technology for consuming organizational knowledge includes:

- *keyword-based entity search*;
- *table- or matrix-based presentation* of relations and attributes of entities;
- *graphical browsing* of relations among entities;
- *category-based search* of entities.

4. System Architecture

In agreement with the methodology outlined so far, the software architecture is composed of three layers, covering the following requirements:

- Capturing and aggregation: back-end components to acquire and re-engineer the data according to a common ontology, to infer new knowledge and to link the data to others;
- Publishing and hybridizing: components that enable storage and retrieval services;
- Consumption and application: components that contextualize the information with respect to end-user tasks, or web services to fetch the data by remote third-party systems.

Fig. 1 describes how research aspects have been addressed along the methodological dimensions, while in Fig. 2 we have depicted the distribution of the functional components among the architectural layers: an infrastructure of components entirely based on semantic technologies. The system has evolved in the last year with respect to both what as presented previously[14] when we highlighted the importance of following a service oriented application paradigm. In general, such architectures are 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[15]. Our system fits well in this

[14] A. Gliozzo, A. Gangemi, V. Presutti, E. Cardillo, E. Daga, A. Salvati, G. Troiani, *A Collaborative Semantic Web Layer to Enhance Legacy Systems*, cit.; C. Baldassarre, E. Daga, A. Gangemi, A. Gliozzo, A. Salvati, G. Troiani, *Semantic Scout: Making Sense of Organizational Knowledge*, cit.

[15] B. Heitmann, C. Hayes, E. Oren, *Towards a Reference Architecture for Semantic Web Applications*, in "Proceedings of the 1st International Web Science Conference", 2009.

category, even if we found it more effective to describe it through its functional aspects.

Starting from the bottom of Fig. 2 we can see how data is acquired, reengineered and linked from several different sources by the means of a set of predefined processes. The first targets data reengineering (reengineering tools). These processes exploit different technologies and methods, from SQL scripting – in the case of internal relational databases – to web crawling – in the case of Google Scholar data, which is not provided as RDF. Linking tools refer to the set of predefined processes to generate `owl:sameAs` triples to bind stored entities to external repositories. A special case of data linking is the categorizaton by means of Wikipedia categories. The Categorizing tools box includes such components, which interact with the quad store and with DBpedia to generate `dc:subject` links. We have enriched the knowledge base by materializing data from the categorization process, as well as from newly inferred data. We perform enrichment by means of additional components: DL reasoners and SPARQL query engines. A special role is played by the ACL processes set. This includes a set of scripts to configure restricted data to be in specific graphs hidden from anonymous users – exploiting the graph-based access control system of Virtuoso.

All the described components are batch processes which are executed to rebuild the whole data in the quad store.

The second layer of the architecture deals with storage and retrieval. All the functionalities operate over RDF data, hosted in a Virtuoso OpenSource instance. The SPARQL endpoint is the gateway for accessing the data, and is the most common way of exploiting the information by applications. In addition to that, the system includes an entity retrieval search engine, which allow to select entities starting from a traditional full text query. On top of the SPARQL endpoint there is also an entity name resolver, since all entities with a URI within the CNR authority (http://www.cnr.it) can also be retrieved by simply dereferencing the URI, which is computed by processing a specific SPARQL query template.

This last access method is used by the data.cnr.it web site, which is also responsible for the human readable (HTML) version of each resource. This can be considered as the basic way of consuming linked data. The consuming layer contains a set of applications that we can consider to be *clients* of the retrieval services - *entity searcher*, *data.cnr.it browser*, *data.cnr.it query interface*, *graph navigator* and *exploratory browser*, which will be described in Sect. 9.

Next sections from 5. to 9. reflect the organization of methods depicted in Fig. 1, and contain the description of how we achieved the realization of the functional components in Fig. 2.

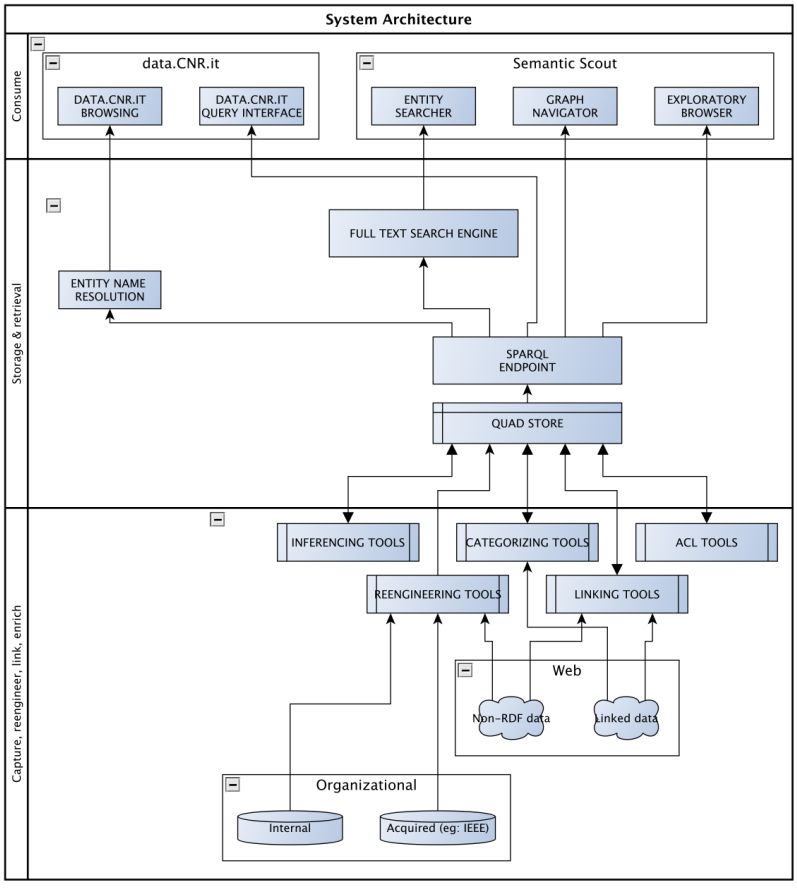

Fig. 2 – Architecture

5. Data Sources

Legacy systems in large organizations often lack of harmony at high level, due to the following reasons:

- spared sources inside the organization; different technologies and access methods;

- overlapping data between internal data management and acquired repositories (for example acquired publication archives);
- overlapping domains of disjoint applications: different applications may not share information when/as they should.

Another viewpoint over data sources shows the wide range of domains covered by organizational data. All those issues become hurgent when bridging all the data to RDF.

In Fig. 2 four sources are shown:

- internal databases, which represent the data managed by the information systems of the organization;
- acquired repositories (in the case of CNR, can be the IEEE archives of publications);
- publicly accessible sources on the web (not-RDF);
- linked Open Data (for example DBPedia categories used as classifier for research subjects).

Internal Data (CNR Databases)

We have elsewhere described[16] how the databases of CNR are hosted and which kind of information is present. For the sake of the current work we have been automated the reengineering process in order to be able to produce up to date information. The current policy provides the execution of the whole reengineering process once a month. Ongoing work includes the possibility of reengineer the different databases asynchronously, according to the management lifecycle of the specific data. The domains that they cover include organizational data, research activities, administration, people and documents and come from a set of not complitely integrated subsystems.

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 con-

[16] C. Baldassarre, E. Daga, A. Gangemi, A. Gliozzo, A. Salvati, G. Troiani, *Semantic Scout: Making Sense of Organizational Knowledge*, cit.

sume the data (e.g. forms in the web portal); this is detailed in section about ontology desing.

Acquired Repositories (IEEE Archives)

Another kind of source comes from outside of the organization and is hosted in local system as copyright-protected data, which usage is allowed under the restrictions of specific contract. One example are the IEEE archives, which CNR acquires as part of its istitutional activity. This data is stored in physical hard-drives as PDF files along with specific XML files hosting metadata. Both the syntactic structure and the vocabulary used is proprietary of IEEE.

Not-RDF Publicly Accessible Data on the Web (Google Scholar Data)

This source is one of the most considered public sources for publications' citation numbers. This is one example of *deep web* source, since it is accessble only by humans and its information can hardly be extracted in machine computable format.

Linked Open Data (ORO Dataset, DBPedia)

This category includes SPARQL endpoints, ie data published according to the linked data best practices. In this article we will meet two exemplary sources: one is the well known DBPedia, used as a categorization criteria, the other is the ORO repository published by the Lucero project[17], which expose the publications done by the Open University.

6. Ontology Design

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

[17] See http://lucero-project.info/lb/about/.

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[18] 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[19].

7. Data Design and Linkage

This activity includes all the preliminary tasks that needed to be solved for publishing the data.

[18] V. Presutti, 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, Springer, 2008.

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

7.1. Reengineering Data

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 data design 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 propose to "embed" *ad-hoc* queries to databases into annotations to the elements of an ontology)[20].

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

1. each dataset must be focused on collecting the instantiation of a single OWL property (i.e. obtaining an property-centric dataset);
2. a network of datasets is preferred to a monolithic collection of data materialized in a single file.

7.2. Inferencing New Knowledge

Some new triples have been generated:

- inverse relations;
- co-authorship relations;

[20] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, A. Poggi, R. Rosati, M. Ruzzi, *Data Integration Through Dl-lite Ontologies*, in Schewe K.-D., Thalheim B. (eds.), "Revised Selected Papers of the 3rd Int. Workshop on Semantics in Data and Knowledge Bases (SDKB 2008)", Vol. 4925 of Lecture Notes in Computer Science, Berlin, Springer, 2008, pp. 26-47.

– top entity (an entity for the CNR individual), well known data properties and basic relations to e.g. CNR departments have been materialized.

7.3. Linking Data

Considering we have now all the data in RDF format, we must accomplish the task of consuming this data as a whole. This task is not simple, and we will not cover here all the aspects: the basic starting point is to link different URIs representing the same real-world entity, in other words, to generate so-called `owl:sameAs` links.

For this task we used the Silk Discovery Framework (SILK2)[21] generating the following link types:

– `owl:sameAs` links between the publications of CNR and the same entries in the IEEE dataset;
– `owl:sameAs` links between the publications of a CNR research unit and the same publication belonging to another research unit (research results are acquired as disjoint sets, one for each research unit).

7.4. Categorizing Entities

A different kind of data linking applies a method that exploits DBpedia knowledge base as a reference subject catalogue[22].

Such method is based on a text categorization system whose goal is to link documents to categories selected from the more than 500,000 categories present in DBpedia, which then provides a rich set of distinctions for the scientific subjects of interest from CNR case study. The system[23] implements a method called *SemioSearch*: it finds the best matching between a text and a synthetic text that is assumed as a *semiotic* representative of a DBpedia category. The synthetic text for a category is generated by means of a customizable SPARQL query that selects all the texts relevant for that category, e.g. the titles, labels and abstracts for the pages with that category. In the data.cnr.it project, we have applied SemioSearch also in order to gener-

[21] See http://www4.wiwiss.fu-berlin.de/bizer/silk.

[22] C. BIZER, J. LEHMANN, G. KOBILAROV, S. AUER, C. BECKER, R. CYGANIAK, S. HELLMANN, *DBpedia – A Crystallization Point for the Web of Data*, in "Journal of Web Semantics", Vol. 7, 2009, pp. 154-165.

[23] See http://wit.istc.cnr.it:8080/wikifierNew.

ate a representative text for CNR researchers, based on a different SPARQL query that extracts the synthetic text from the most relevant literals for a researcher, e.g. publication titles, project names, etc. (Fig. 3).

Fig. 3 – Output of the DBpedia categorizer applied to text describing one of the authors of the paper

The output of the categorizer has been represented in RDF by using the `subject` relation from the Dublin Core vocabulary[24], e.g.:

```
<> dc:subject
dbp:Knowledge_representation                    (1)
<> dc:subject
dbp:Artificial_intelligence                     (2)
```

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 Sect. 9. we show the usefulness of the categorization data for expert finding and semantic browsing of data.

[24] Prefixes resolve as follows: cnr: http://www.ontologydesignpatterns.org/ont/cnr/; dc: http://purl.org/dc/elements/1.1/; dbp: http://dbpedia.org/resource/Category.

7.5. Configuring Access Levels

Not all data from a research organization can be published as Open Data. In our case, most of the internal organizational data have been published as Linked Open Data and accessed by the data.cnr.it access point. At the same time a sub-set of the data belong to the research unit, to the person (it is private personal data) or it has been acquired with limited usage rights. All those cases should be covered by a system which aims to enforce data integration widely, as in the case of semantic web applications.

In our case the following access levels have been considered and applied at a graph based level:

- public graphs: data can be seen by any anonymous agent;
- research unit graphs: only users belonging to the specific research unit can see the data;
- CNR graphs: only authenticated users can see this data (this is the case of the IEEE links to PDF versions of the articles).

In Sect. 8. we describe how this affects the publishing of data.

8. Data Publishing

data.cnr.it is the public access point to the linked data of CNR (Fig. 4). The CNR data cloud consists of several graphs. The core part of the system is based on a Virtuoso OpenSource[25] instance with in parallel a full text Lucene based engine, which expose a search engine web service. The basic access is given by a SPARQL endpoint, on top of it several applications can be build.

As outlined in Sect. 7.5. graphs can have different access policies. To better manage this we have started by publishing property centric graphs. We have generated more then 200 RDF datasets, each of them instantiating the value for a single property. This approach has several benefits. First, the granularity of the publication 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. Another positive aspect is in query optimization, since it is possible to restrict the conditions to the name graph which we knoe holds the relative triples. In perspective, the property-centric organization of datasets support also the lifecycle of reengineered data because it is easier to identify smaller

[25] See http://virtuoso.openlinksw.com/dataspace/dav/wiki/Main/.

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.

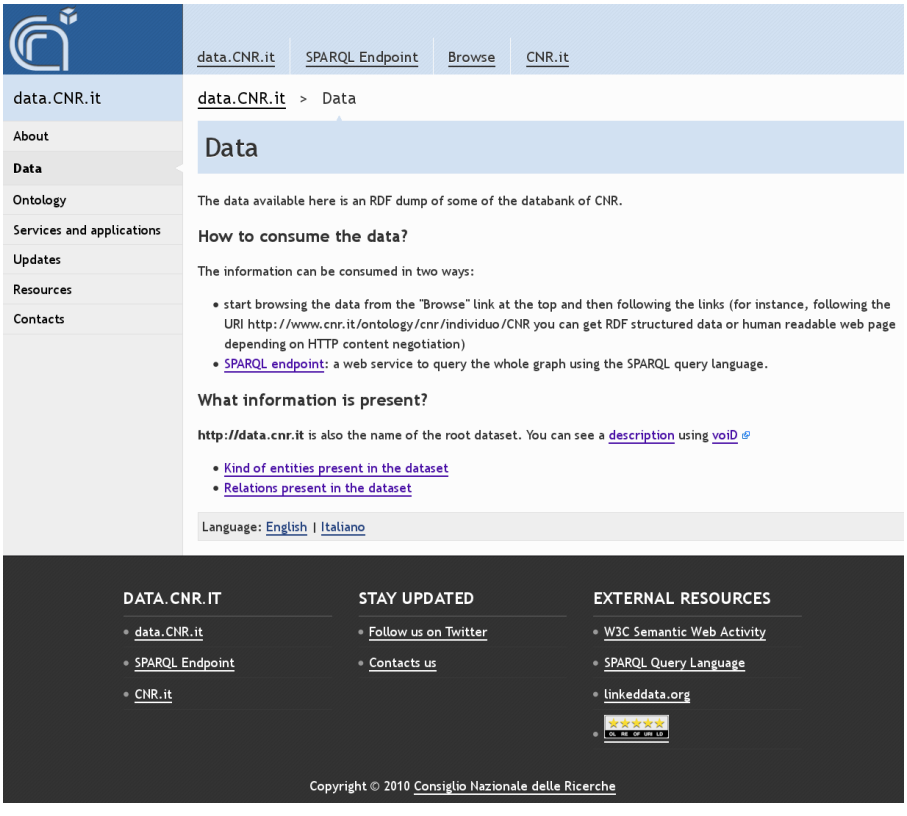

Fig. 4 – data.cnr.it

9. Data Consumption and Applications

Applying linked data techniques to organizational data has the result of dividing the time of data design to the one of application design. The basic way of observing data is provided by the infrastructure of data.cnr.it.

9.1. data.cnr.it

data.cnr.it is the knowledge hub of the CNR: on top of it several applications can be built by exploiting the same harmonized data.

9.1.1. Browsing Linked Data with Entity Resolution

The first way to explore CNR linked data is HTML. The recipe we have used ensures that each entity can be resolved via HTTP protocol. Web of Data principles require the Universal Resource Identifiers to be the unique name of entity: in data.cnr.it all the entities can be browsed by resolving their URI. In Fig. 4 the *browse* link in the top menu points to the entity for the CNR organization: http://www.cnr.it/ontology/cnr/individuo/CNR, which is the top level access point to the network. By using this access point, the HTML version of the resource can be easily browsed by a human.

9.1.2. SPARQL Query Interface

The second way to access CNR linked data is querying. The SPARQL endpoint can be queried using the data.cnr.it query interface, which wraps some user-friendly mechanisms:

- Default namespaces are prepended to the query. Those include predefined prefixes for all the classes and properties used (this is needed because we are dealing with an ontology network, which has many namespaces, see Sect. 6.), and prefixes for entity names (based on the patterns used for generating entity names)
- Autocompletion is enabled for all the schema entities (see Fig. 5).

9.2. Semantic Scout

The third, and most sophisticated way to explore CNR linked data is the Semantic Scout. It includes a set of functional components, on top of the CNR data cloud and the CNR ontology. This section unfolds each tool and explains how and why they fit into the data.cnr.it toolkit. In addition, we present the use cases where they are daily used by CNR people.

9.2.1. Entity Searcher

The Semantic Scout infrastructure includes an information retrieval engine. As described in Sect. 3., starting from a known interaction pattern is beneficial to the users. Fig. 6 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 semantic search performed by

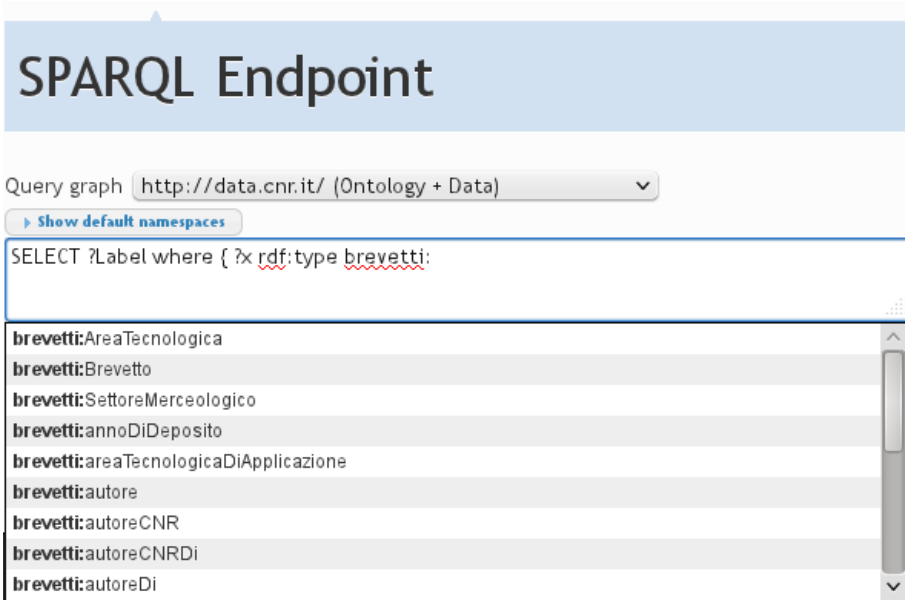

Fig. 5 – data.cnr.it - SPARQL query interface

Fig. 6 – CNR Semantic Scout - Entity Search Engine

the Scout is still performed on documents, but retrieves *entities*. The results of that query actually include a major set of researchers working at the cross-roads between cognitive science, sociology, and computer science[26].

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.

This search design pattern is based on a semiotic assumption: each entity has 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 search modality 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[27].

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 for data management, reasoning, and actual consumption of data.

9.2.2. Graph Exploration

The *Semantic Scout* visualizes a network out of the entities of an organization, besides content objects like documents or images. 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. We have tested several examples of RDF data browsers, and *Graph Explorer*[28] has been our first choice in order to effectively tackle the presentation of, and the interaction with,

[26] For a related experiment on using the Semantic Scout to build a "dream team" for a focused research programme see C. BALDASSARRE, E. DAGA, A. GANGEMI, A. GLIOZZO, A. SALVATI, G. TROIANI, *Semantic Scout: Making Sense of Organizational Knowledge*, cit.

[27] See http://code.google.com/p/semanticvectors/.

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

the datasets. It overcomes the limitations of exposing data in tabular format only, providing an appealing interface for our 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 include some of the CNR-ontology concepts, considered as *hub subjects* of properties from the CNR ontology.

Suppose a user looks for CNR researchers that have competence in 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 certain researcher is involved in the Semantic Web from within CNR. Notably, it allows us to find additional serendipitous information. Fig. 7 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 the related departments. A panel on the right gives a description of the focused node.

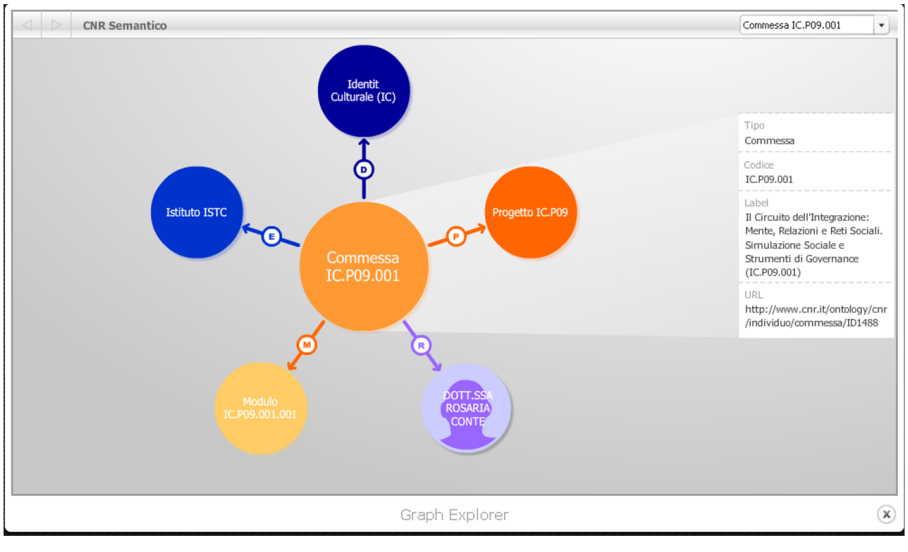

Fig. 7 – CNR Semantic Scout - Graph Explorer

Finding serendipitous information moves the focus on exploring, instead of (purely) searching: exploring works along multiple paths, which could be previously unknown to the user; this is emphasized by graph-based representation.

More recently, we have built a second exploration tool, called *Exploratory Browser* (Fig. 8), which works directly on RDF graphs[29]. The Exploratory Browser introduces some other features. First, the graph is built within a spherical space, which makes closer entities focused, while moving the more distant ones farther: this metaphor is better with respect to the very nature of RDF graphs.

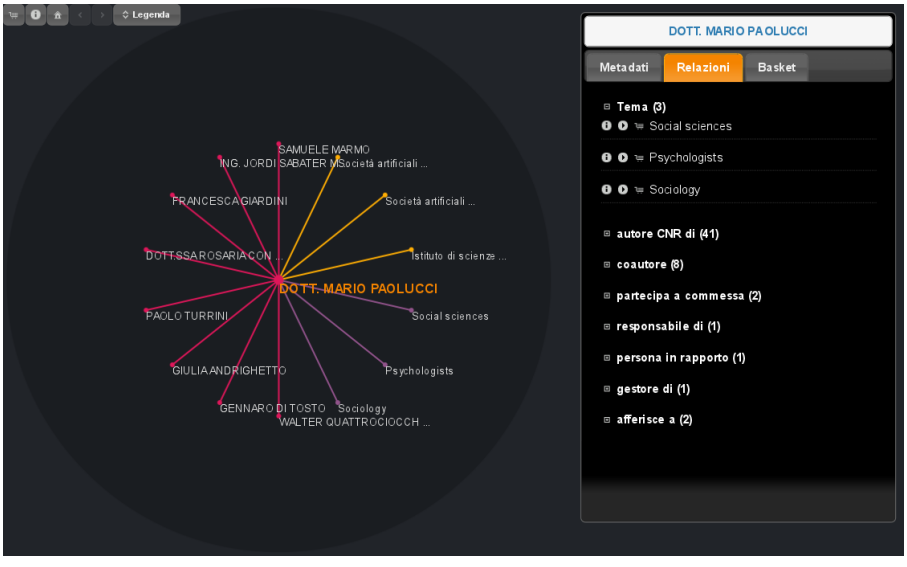

Fig. 8 – CNR Semantic Scout - Exploratory browser

Second, with the Exploratory Browser the user can customize the navigation by adding nodes to the predefined relations showed in the graph. By default, only a set of predefined core nodes are shown in the graphical space in order to limit the tangledness of the graph, but the complete set of properties (including datatype properties) is shown in the right box, which displays all the information about the entity currently focused in the graph. Additional nodes can then be selected in the box and added to the current graph.

Third, user-selected entities can be put in a basket, and then exported as an HTML page that contains all data about them. Ongoing work includes the removal of entities from the graph and a customization layer, to personalize the default behaviour of the explorer.

[29] Graph Explorer needs an intermediate conversion to XML.

10. 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 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. Our work has lead to the creation of the semantic portal data.cnr.it, and the Semantic Scout services.

Visualization and interaction components are key in adopting semantic technologies on a large scale, and we have addressed those issues by experimenting with two different metaphors, one of which specifically designed for the Semantic Scout.

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 in the Linking 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.

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:

- social refinement of the CNR datasets through semantic wikis or content management systems, and social bookmarking,
- enriching the CNR datasets with new relations inferred from linking to other external data, and
- extending the capability of matching offer and public request for CNR competences.