WEASEL: Vodafone R&D Corporate Semantic Web

Juan José Valverde[1], Carlos Buil[2], José Manuel Gómez-Pérez[2]

[1]Vodafone, www.vodafone.com
juan-jose.valverde@vodafone.com
[2]Intelligent Software Components, www.isoco.com
{cbuil,jmgomez}@isoco.com

Abstract The 2006 Gartner emerging technology curve highlights the relevance of the Corporate Semantic Web as one of the most promising IT areas in the next five years. The work presented herein describes WEASEL, an initiative funded by Vodafone to apply and evaluate such technology in the context of a large multinational company. This scenario comprises a number of heterogeneous web sites containing unstructured and related, but physically decoupled, information which needs common models that provide unified ways of representing information across the different sources, i.e. ontologies. Three main milestones were defined for WEASEL: the creation of a domain ontology, the extraction of information from the different sources and its semantic annotation and aggregation, and the creation of a new web site containing a semantic search engine which provides natural interfaces for retrieving the aggregated information. WEASEL concluded with its evaluation by Vodafone.

1. Introduction

Vodafone is constantly in the quest of technologies with the potential of playing a key role in the development of new, highly appealing services for its customers. For a company like Vodafone, it is critical to provide the best user interface facilities, especially from mobile terminals, that allow natural and simple access to internet services. For Vodafone it is equally important to enhance user access both to pre-existing and new internet services. The goal is in any case to maximize service usefulness and user satisfaction.

Semantic Web technologies can certainly contribute to these goals. Such contribution is twofold: i) semantic annotation facilitates the aggregation of meaningful information from different web sites, allowing to generate new and more powerful services which expose richer contents to users; ii) the background of the Semantic Web in the field of user interfacing and information retrieval provides the necessary means to intuitively access such contents with the highest degree of usability. Exploiting taxonomic structure of domain ontologies and natural language query interfaces are key in this direction.

The 2006 Gartner emerging technology curve (Figure 1), also known as hype, highlights the relevance of the Corporate Semantic Web as one of the most promising IT areas in the next years. In this direction, project WEASEL, an initiative funded by Vodafone, intends to demonstrate the value of the Semantic Web by increasing service functionality and usability, and allowing better integration processes for the creation of new internet services. The domain of application of WEASEL is the Vodafone R&D corporate web, in particular, the PROGRESS project reporting database and other sites, decoupled from PROGRESS, containing additional and related information. WEASEL

should ultimately improve the way people access information while enhancing the quality of the retrieved data.

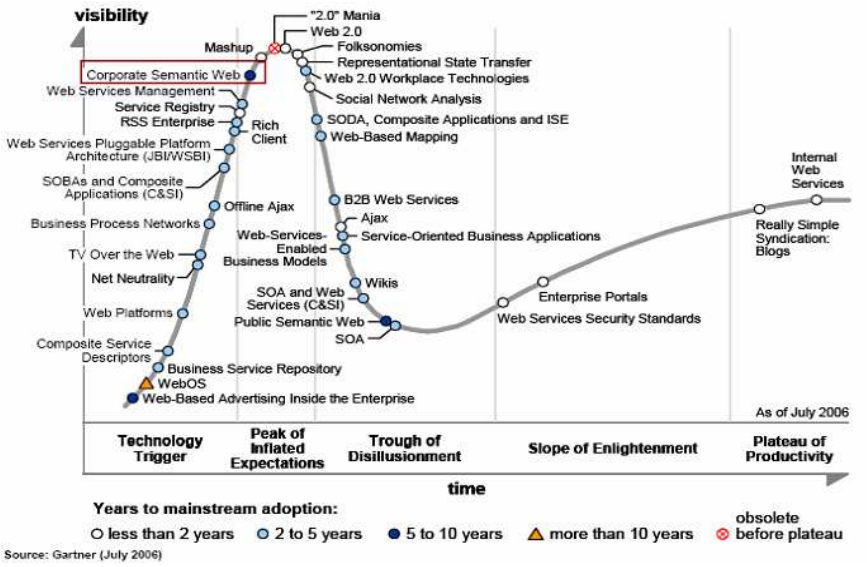

Figure 1: Gartner hype 2006

Three main milestones were defined within the project. The first milestone is the creation of a domain ontology which describes the information contained in PROGRESS and the related sites. The second milestone is the extraction of information from the different sources and its semantic annotation and aggregation. Finally, the third milestone is the creation of a new web site containing a semantic search engine which provides natural interfaces for retrieving the annotated information. The overall process is reflected in Figure 2.

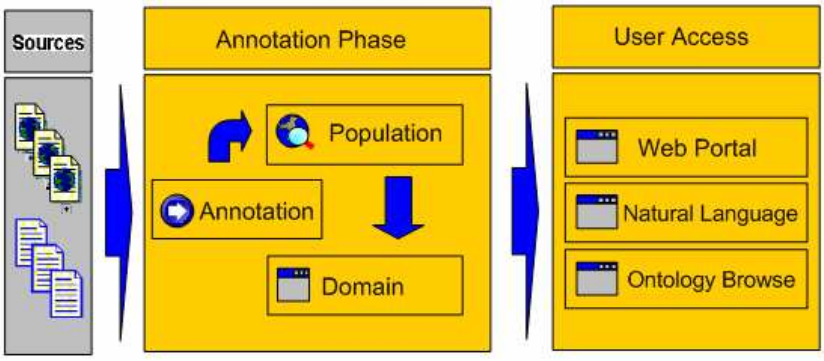

Figure 2: Overall WEASEL process: ontology development, annotation, and information access

The remaining of this paper is structured as follows. Section 2 describes the sources of information exploited by WEASEL and the challenges that they present, in particular PROGRESS and the associated web sites of the Vodafone R&D corporate web. Section 3 presents the Vodafone R&D ontology developed in order to describe the domain of the target information. Section 4 describes our approach towards extracting, aggregating, and semantically annotating such information. Section 5 presents the information retrieval techniques applied in WEASEL. Finally, section 6 concludes the paper and evaluates the work presented herein.

2. PROGRESS and Related Information Sources

PROGRESS is an internal database used to store information related with Vodafone R&D projects. The database contains all the project information available including descriptions, associated reports, project plans, meetings, people involved, etc. The database is accessed using a web interface which allows keeping track of current projects as well as maintaining a repository of those already completed. Querying is enabled by means of a simple keyword-based search engine.

Since the Vodafone R&D group is split into four different locations (UK, Germany, Spain and the Netherlands), this repository has become instrumental to share knowledge among its approximately eighty members about the ongoing work (something which becomes crucial to solve cross-technology related issues). One of the main problems of PROGRESS is scalability. The amount of information stored in the database is getting too large, hence difficulting the retrieval of relevant information during querying.

Other web sites in the corporate web contain relevant information which extends that of PROGRESS. However, relating both sources of information is not a straightforward task. The main difficulty is that the same types of information can be arranged and represented heterogeneously, depending on the site. Thus, a common semantic model is necessary that allows uniquely describing all these information, as shown in section 3.

Figure 3: Aggregating different sites in the Vodafone R&D corporate web

Vodafone R&D projects, although focused in a particular technology, usually need from other areas of knowledge for its accomplishment. Mobile internet projects, for example, tend to need from Radio experts in order to optimize access from mobile terminals. In other occasions, old projects of one team can present solutions for issues found in current projects of another team. In those situations getting the right information from PROGRESS becomes critical for projects to succeed.

For example, if Vodafone R&D Germany needs to contact companies with expertise in the area of Semantic Web, information about projects contained in PROGRESS should be cross-related with information about external companies contained in other site of the R&D corporate web. This way, it would arise that iSOCO, company expert in Semantic Web, has worked in the context of project WEASEL for Vodafone R&D

Spain. However, being completely decoupled, it would not be possible to automatically relate these information sources. WEASEL approaches this problem aggregating disparate information sources by means of semantic annotation (Figure 3).

3. Conceptualization of the Domain: The Vodafone R&D ontology

As mentioned in previous sections, heterogeneous knowledge environments like this need common semantic models that provide unified ways of representing information across the different sources. This problem is faced in the area of Semantic Web by means of ontologies. Ontologies are defined as [1, 2, 3, and 4] formal, explicit specifications of shared conceptualizations.

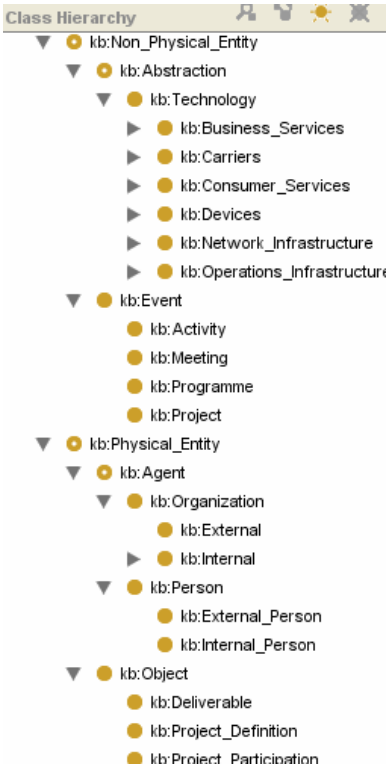

Figure 4: Class hierarchy of the Vodafone R&D ontology

The ontology representing the domain of WEASEL (Figure 4) describes the knowledge contained in the Vodafone R&D corporate web. This ontology specifies concepts like project, project phase, project plan, employee, company, technology, meeting, place, etc., i.e. all the knowledge of interest for the R&D department. The lifecycle of this ontology follows the methodology defined in [5].

The taxonomy is divided in two main high-level concepts: *non physical entity*, representing concepts like e.g. meetings, projects, and technologies, and *physical entity*, for concepts like e.g. persons and organizations. *Person* and *organization* are divided in external and internal entities. All the materials produced by projects, like e.g. reports, proposals, and plans, are classified as *objects*. *Technologies* are structured following a previously existing classification in Vodafone.

As sections 4 and 5 will show, both the semantic annotation and aggregation process and the different techniques for information are based on this ontology.

4. Semantic Annotation and Aggregation

Knowledge Parser® (KP) offers a software platform that combines different technologies for information extraction, driven by strategies which define, according to a domain ontology (see previous section), the most accurate extraction technology for each source type. KP has been successfully applied in a good number of scenarios like e.g. the development of a semantic search engine for the international relation sector [6]. WEASEL follows the same approach.

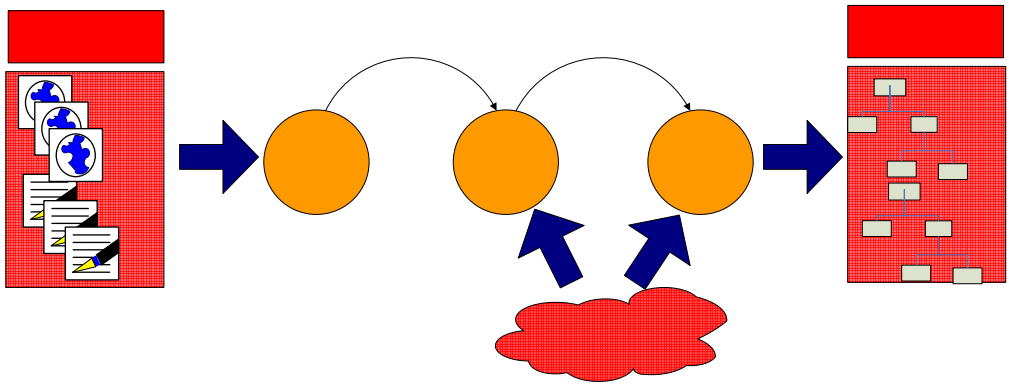

Figure 5: Annotation of unstructured sources using KP

Annotating unstructured sources with KP (see Figure 5) is a process comprising three main stages: i) extracting information from the source, ii) assigning the extracted information to the appropriate ontology entities, and iii) its coherent insertion in the ontology. The first stage deals with document processing and interpretation. The second part adds semantics to the extracted information, according to domain information specified in the ontology. The last stage is in charge of populating the ontology, i.e., creating instances of the concepts contained in the domain ontology with the data previously extracted from the sources and interpreted. A detailed description of the annotation process can be found in [6].

In WEASEL, as a result of this process, the unstructured information previously contained in decoupled locations of the Vodafone R&D corporate web, like e.g. PROGRESS, has been aggregated and related according to the semantics of the R&D ontology. Thus, such information has been endowed with two fundamental properties: i) it is well-structured and ii) it follows a formal, explicit specification, which is common to all the information sources. These two characteristics allow subsequent exploitation of the data in the form of enhanced information retrieval through advanced user interfaces based on both ontology-based information browsing and natural language querying.

5. Information Retrieval

Once the Vodafone R&D information was aggregated and semantically described, it is ready to be exploited, hence satisfying the third milestone to be addressed by WEASEL, which consisted of facilitating access to such information. The resulting semantic search engine was endowed with two user interfaces for information retrieval, both grounded on the ontology. The first interface allows using the ontology taxonomy as graphical means to browse this information. The second interface allows users to make queries expressed in free text which are then processed by the search engine, producing an intuitive explanation of the results. Next, we describe both interfaces.

5.1. Ontology-based information browsing

This interface allows users to browse the ontology and access its instances, i.e. the previously annotated information. The WEASEL web portal directly displays the class hierarchy tree structure of the ontology, enabling users to navigate it and search the occurrences of each concept. Figure 6 shows, on the right, this interface and, on the left, the results produced by the system after a query. Each attribute of the selected concept has an associated text field which can be used to enter information constraining the

search. This kind of search accepts regular expressions, where blank fields are interpreted as free variables, following the "*" wildcard analogy.

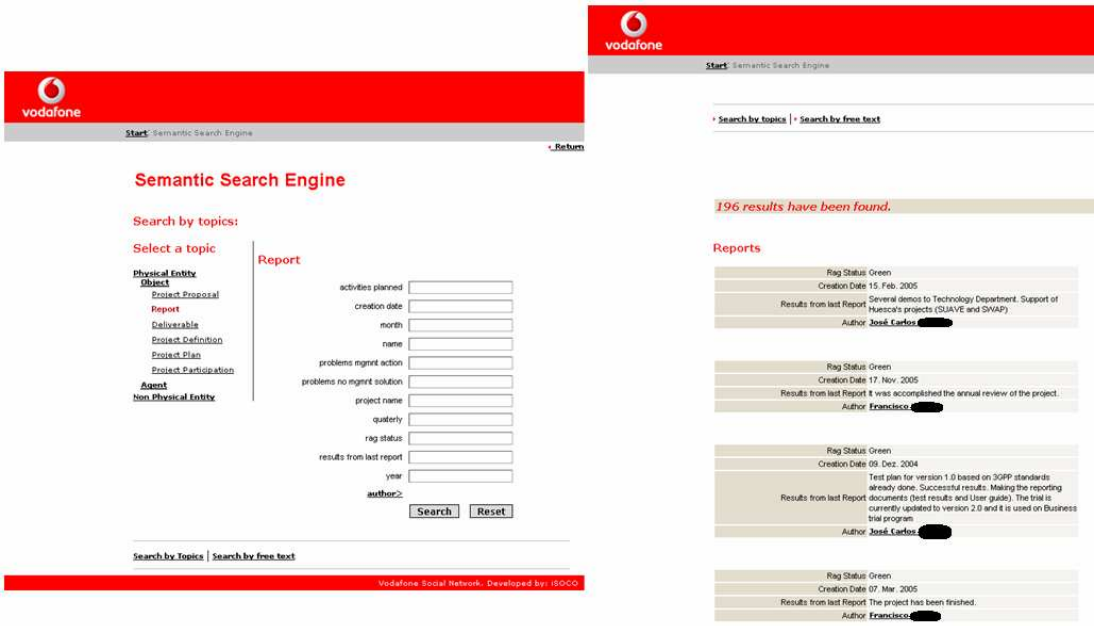

Figure 6: Search by concept and sample result s

5.2. Natural Language Query Interface and Answer Explanation

Figure 7 shows the GUI for the natural language query and answer explanation facilities of WEASEL (left and right of the figure, respectively). In the example of the figure, the user asked about projects related with the Semantic Web and received as answer information about WEASEL.

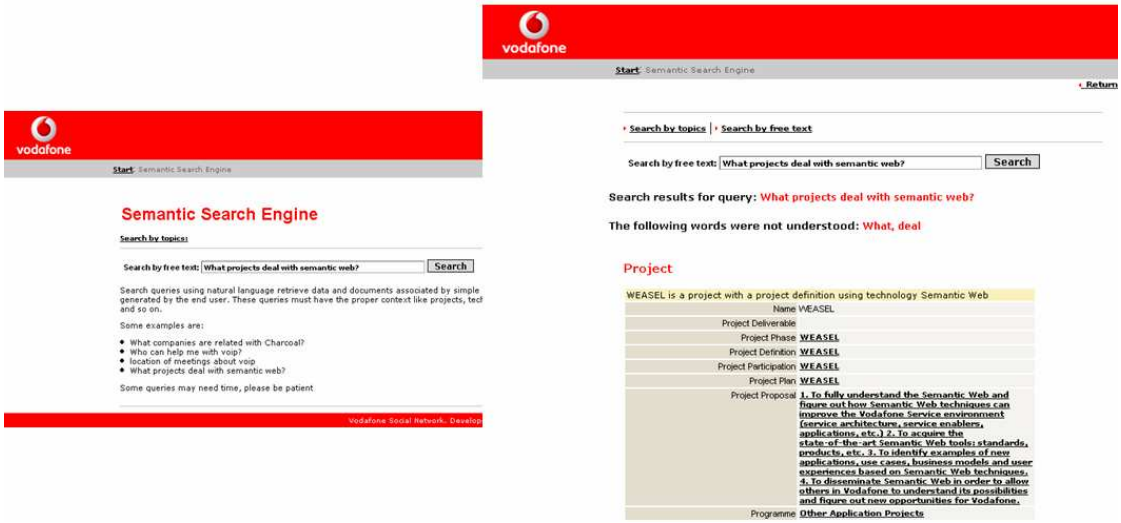

Figure 7: GUI for natural language interface and answer explanation

The natural language querying interface follows a three-staged process: i) the query entered by the user is interpreted, extracting the relevant terms according to the ontology and applying techniques like lemmatization and spell checking, ii) this set of terms is processed in order to automatically build a query, implemented in an ontology query language like SPARQL [7], which is executed against the ontology, and iii) the resulting information is displayed and explained to the user.

This kind of interfaces must provide very high measures of precision and recall, with the goal of minimizing false positives and negatives. However, this is not enough. The main problem here is not to return suitable results but to find the best way to describe these results, which usually are not trivial, or expected, by the human being making the question. Thus, users need feedback from the search engine, explaining its understanding of the query. In WEASEL, this explanation is generated from the SPARQL query itself. The usefulness of answer explanation systems [8] can be illustrated by e.g. the query *who knows about voip?*

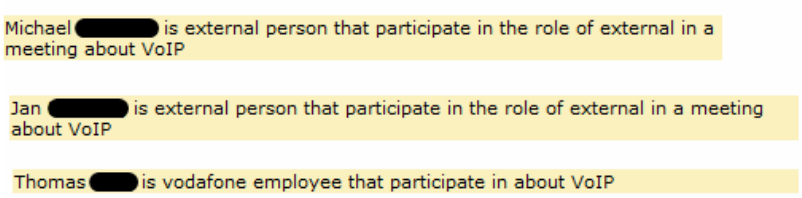

Figure 8: NLP query interface answer explanation

The answers produced by the system are all correct, but for different reasons. Figure 7 shows answer explanations justifying that Michael and Jan are knowledgeable of VoIP because both are external persons who participated in meetings about VoIP. On the other hand, Thomas is a Vodafone R&D member who has participated in work related with VoIP, and therefore a valid answer, too. It can be noted that answer explanation in WEASEL exploits not only the knowledge contained in the ontology about the resulting terms but also the relations between them.

6. Evaluation and Conclusions

The evaluation of WEASEL drafted several conclusions. First, the lack of a large corpus of standard ontologies, which can be used as a repository for the development of Semantic Web applications in a vast variety of different domains, usually requires the construction of ad-hoc ontologies for each particular case. Additionally, these ontologies rarely become standards themselves as they only cover the domain from the perspective of each particular application, preventing future reuse for a more general field of application.

Second, though powerful, the semantic annotation technology used in WEASEL certainly lacks the flexibility required to quickly adapting to changes in the layout of information sources. This is really an annoyance for dynamic environments where the way information is structured changes frequently, but not in cases where the layout keeps constant even if information itself changes. Additionally, as a minor drawback, the annotation process is not completely automatic and requires being supervised by humans.

Third, despite the previous difficulties, WEASEL has proven that the Semantic Web technology applied towards i) aggregating different information sources under the umbrella of a shared conceptual model, i.e. the ontology, and ii) providing natural querying interfaces can greatly enhance user experience. In environments like the Vodafone R&D corporate web, the use of Semantic Web can really push forward what users can get from information systems both in terms of ease of use and usefulness of the results.

In this paper we have described the application of Semantic Web technologies to the R&D department of the corporate web site of a large company like Vodafone in the

context of project WEASEL. We have built an ontology, conceptualizing the domain, which served as the basis both for the semantic aggregation and annotation of R&D information and its retrieval, by means of interfaces focused on user natural interaction. Finally, we have provided the results of the evaluation performed by Vodafone and a summary of their conclusions.

References

1. T. R. Gruber. A translation approach to portable ontology specifications. Knowledge Acquisition, 5:199–220, 1993.
2. N. Guarino. Formal ontology, conceptual analysis and knowledge representation. International Journal of Human-Computer Studies, 43(5/6):625–640, 1995. Special issue on The Role of Formal Ontology in the Information Technology.
3. W. N. Borst. Construction of Engineering Ontologies. PhD thesis, University of Twente, 1997.
4. G. van Heijst, et al. Using explicit ontologies in KBS development. International Journal of Human-Computer Studies, 46(2/3):183–292, 1997.
5. Fernández-López M., Gómez-Pérez A., Juristo N. (1997) *METHONTOLOGY: From Ontological Art Towards Ontological Engineering*. Spring Symposium on Ontological Engineering of AAAI. Stanford University, California, pp 33–40.
6. Rodrigo, L., Benjamins, V.R., Contreras, J., Patón, D. J., Navarro, D., Salla, R. Blázquez, M., Tena, P., Martos, I. A Semantic Search Engine for the Internacional Relation Sector. ISWC 2006.
7. SPARQL Query Language for RDF http://www.w3.org/TR/rdf-sparql-query/
8. McGuinness, D., Pinheiro da Silva P. Inference web:Portable and shareable explanations for question answering. In Proceedings of the American Association for Artificial Intelligence Spring Symposium Workshop on New Directions for Question Answering. Stanford University, March 2003.