Using Semantic Search as a Means of Support for Research Projects

Michael Stencl, Oldrich Trenz, Ondřej Popelka, Jiří Hřebíček

Using Semantic Search as a Means of Support for Research Projects

Michael Stencl, Oldrich Trenz, Ondrej Popelka, and Jiri Hrebicek

Mendel university in Brno, Faculty of Business and Economics,
Department of Informatics, Zemedelska 1, 613 00 Brno, Czech Republic
{michael.stencl,oldrich.trenz,
ondrej.popelka,jiri.hrebicek}@mendelu.cz
http://ui.pefka.mendelu.cz/en/gacr403

Abstract. The present contribution deals with analyzing the issue of semantic searching and the possibilities of its integration as a part of the portal for scientific projects support. The difficulties of integration of the semantic principles into the solution itself, the software support of the portal creation, i.e., solutions that are accessible, and possibilities of further development are discussed here. A part of the solution is the integration and support of XBRL documents being a fundamental part of inter-company communication and information exchange. Last but not least, the contribution discusses the possibilities of employing the principles of semantic searching in portal discussions which are treated on the levels of social networks, and the integration of these discussions as supporting the management of team projects.

Keywords: Searching, semantic search, XBRL, CMS, social networking

1 Introduction

Semantic search and its principles is very often joined with the principles of semantic web and often also Web 2.0 [2], potentially Web 3.0 [5]. Garca-Crespo (2010) then clearly defines the Web 3.0 *"[. . .] as a new version of Web 2.0 in which web has advanced to become what Tim Berners-Lee (2007) has termed the Giant Global Grap."* Development of the Semantic research is connected with the basic needs of information, no data, in actual Web era/age. The amount of actual date accessible for free on the Internet is huge and to get the right information in the right time means to have advantage on all levels including business and research. Retrieving information from heterogeneous data stores give power to development or reincarnation of several machine-learning methods to use them in newly defined hybrid expert systems [6] and not only in experts systems. The Semantic Web its impacting traditional sciences, such as chemical, physical [2], sciences.

The Aim of the article consists in state-of-the-art analysis of semantic search applications as background for the Research project and integration of the Semantic search in the research project phases. The integration will be shown

on application of the approach in the Czech Science Foundation GACR project P403/11/1103 - Construction of Methods for Multi-factorial Assessment of Company Complex Performance in Selected Sectors. The Semantic Web[1] offers another communication channel for the research team at different phases of the project. Integration of the Semantic Search, its tools and approaches aid to reliable re-use of data. It also has potential to bring more efficiency to the resources library productivity. Modern digital libraries offers huge amount of information, but the problem is data retrieval. An actual tool includes traditional browsing or keyword-based search strategies. However current approaches still results in enormous numbers of pages without affecting the right topic. This leads to time-consuming manual browsing/filtering while deciding which of the filtered data is relevant to the topic. The manual browsing hand-in-hand with the diversity of file format used in the resource data store has impact on productivity of retrieving the information and on used information system [7] too.

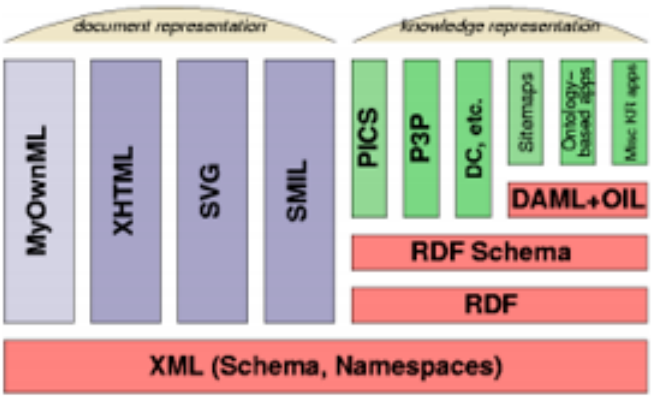

Fig. 1. Descriptive languages [8].

2.1 The Semantic Web and Web 2.0

The fundamental of the semantic web is searching based on the semantic analysis of source data. We are dealing here with amplifying the concept of web pages by supplementary metadata which describes the semantic information of web-sources; and this in such a way that the data are written in a form that can be machine-processed. The metadata subsequently contain a given vocabulary and enable the creation of an adequate relationship between concepts. For in the environment of the common web, it is not possible to establish a single descriptive language containing an established vocabulary (also bearing in mind the specialization of the concrete web), and for this reason we are at present

[1] http://www.w3.org/2001/sw/

witnessing an overlapping of several descriptive languages (Fig.1). In general, we can call this approach decentralized, i.e., it contains in its core all the areas of knowledge.

One competitor or even a successor to the semantic web is the web 2.0, or its successor web 3.0, which enables the centralized treatment of individual services, i.e., processing in one place. This approach greatly improves the capacities of unified administration and the unified application response. In view of the support of community networks and projects (semantic wiki, semantic blogging, semantic desktop), the web 2.0 is, at present, considered to be a suitable extension of the semantic web.

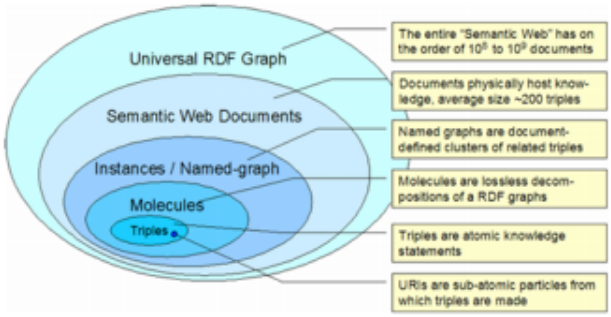

Fig. 2. Search machine [8].

2.2 Semantic Searching and Search Engines

The search engines designed for the semantic and for the classical web are, on the highest application level, very similar i.e., the documents search, data mining from documents, identification of users requests, agents, sequencing and storing results. There is a different configuration, however, on the level of the search core. Here, there is the support of the machine processing on the level of the marked content. If the data is stored using RDF, the support of effective indexing and searching is necessary. In a single document, there are more facts, knowledge and complementing metadata, and all this creates great requirements for the search machine this is shown in Fig.2. If we compare the search chart of the semantic web with the chart created from hypertext links of common web documents, at first glance the two charts are structurally very different. This fact greatly influences the strategy for acquiring and searching documents, and also influences the creation of metrics for their final allocation.

When searching, the search engine compares the identified users requests with the data which have appeared in the indexing already performed. The results of the search engines come as particularly composed documents which, as opposed to ordinary documents, can aggregate data on more levels, and this starting at the level of all RDF data located on the given web, all the way to individual RDF triplets.

3 Results

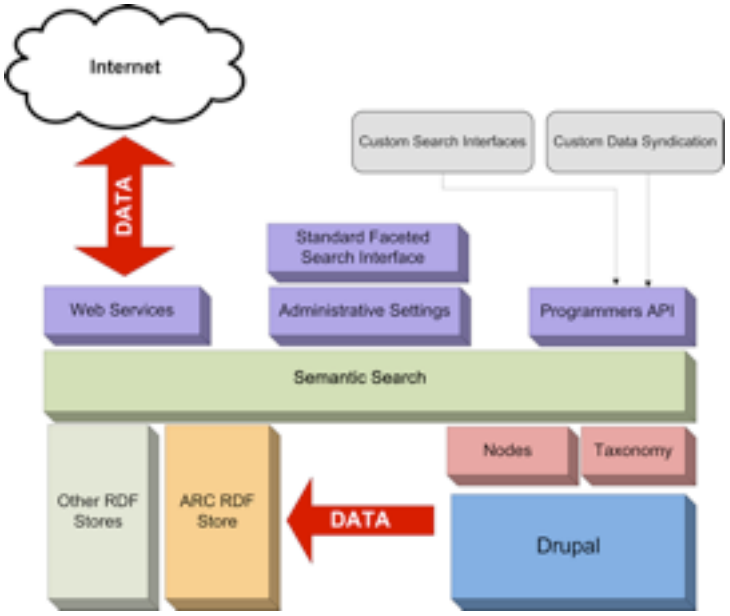

Fig. 3. Drupal Semantic Search module [4].

Good review of Semantic Web applications in, and not only in, digital libraries bring Garca-Crespo (2010). They comment several applications of Semantic Web that provides a complementary vision as a knowledge management environment. The typical research life cycle involves 4 phases:

1. **Planning**; defining the project, involving literature analysis, discussions with a range of experts, arranging funding and resources;
2. **Data preparation**; processing, collecting and describing data;
3. **Analysis**; commenting results, data access, data dissemination, preparation of the knowledge transfer and undertaking the necessary analysis, publications of results over the undertaken experiments;
4. **Research Outcomes**; publications of obtained results, generalizing of the results according to the methods used.

Semantic Web as a research tool could be used across all the phases. Our approach combines several technologies and tools to create qualitatively strong resources for several projects phases.

All the present projects have their own web presentation with basic information about the topic, research team, etc. Our project GACR P403/11/1103 is focused on the construction of methods for multifactor measurement of company performance in chosen economic (CZ-NACE) activities and the creation of a modifiable and broad-spectrum methodology of their putting into practice. Presently we are in the first phase, where the literature analysis takes place.

Table 1. Resource Digital Library (RDL) and Publications digital library (PDL) structure

PID	Authors	Year of Publication	Publisher	Citation (in text)	Key words	Source DBS	Processed by	DOI /URL

As a web project presentation tool we chose the Drupal CMS for the project team background. Outside the standard project annotation and description, Drupal includes the module for Web 3.0[2] that implements the Semantic Search architecture. The module logic of the Drupal is shown on Fig. 3. CMS Drupal uses an RDF store as a search index. The built in store is easy to use. Other RDF stores require Java and configuration. Also dynamically creates default search interfaces, for many searches per site, configurable via admin interfaces. [4] Other possibility to be used as search engine is the Yahoo! SearchMonkey Apps.

Yahoo! SearchMonkey share structured data with Yahoo! Search to display a standard enhanced result (available for certain content types) or the SearchMonkey developer tool could be used to extract data and build apps to display custom enhanced results. [9] For Drupal also speaks presentation of his creator, Dries Buytaert, about the Drupal solution. [3]

But not only the Drupal is involved. Is combined with the Google Docs tools, especially with spreadsheets. The selection where supported because of easy usage and it support the real time multi-user editing option. We use the spreadsheets for several tasks:

1. **Resource digital library (RDL)** it includes results of literature analysis of the state-of-the-art analysis with defined structure (Tab. 1);
2. **Publications digital library (PDL)** key information of the team own publication with defined structure;
3. **Team publication plan**;
4. **Documents share point** information about stored and shared documents containing information for/about project.

First two, the RDL and PDL, also builds the internal resource layer for the Semantic Search engine in CMS Drupal. Google Docs spreadsheets allow usage of standard SQL to query and also its possible to access the spreadsheets data from outside application. The Semantic Search application in our project could be described as on figure 4.

3.1 XBRL

The XBRL format (eXtensible Business Reporting Language) was chosen as s suitable means of information exchange and inter-company communication. It is a worldwide standard for the exchange of commercial information with the

[2] http://semanticsearch.org/

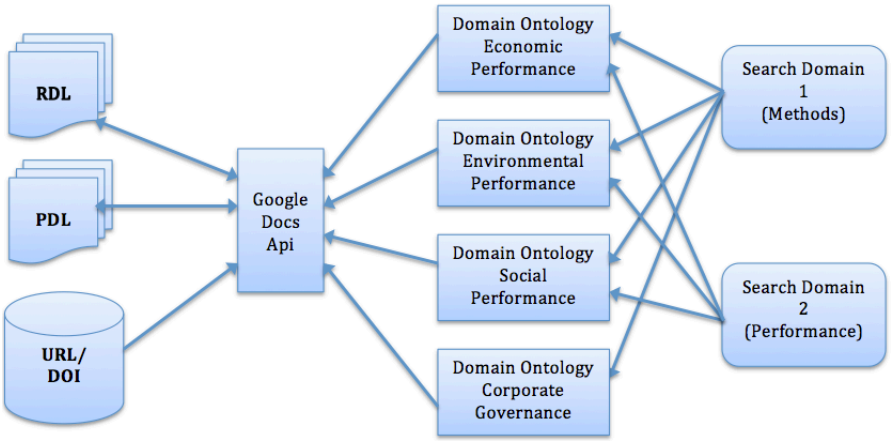

Fig. 4. Project Semantic Search with Google Docs.

support of searching on the level of semantic meaning. An advantage of this standard is the possibility of its simple processing at the level of important world institutions in the state, as well as private sector. This fact is a necessary prerequisite for the integration into the company environment and the willingness to make use of this means of communication for inter-company data exchange.

In our case, the XBRL will be used for acquiring financial documents from individual monitored company subjects. Because the communication will be periodical, the chosen standardized framework appears to be a good choice. The acquired data will subsequently enable us to calculate the indicators describing the present situation of the company, and, based on historical connections, balance sheets about its future aims can be created. The integration will be done on the level of incorporation of the XBRL conversion module into the already existing system; a similar approach to database connection has been applied in the following source. [1]

3.2 Using the social communication means as support of the project solution

Regular meetings are necessary support for the project. The basis is not only the recapitulation of individual steps resolution, but also the coordination of subsequent steps inside each workgroup, because the individual participants are not always available temporally, as well as spatially (teaching, internships, vacations, etc.). It is therefore necessary to select a suitable framework that would enable us to create a work meeting even on the virtual level. Besides the allocation of partial tasks by means of the calendar application on the projects portal we have discussed the possibility of using the Skype technology for leading alternative meetings of the project. This approach has proven to be very effective, not only because it enables the storage of mutual text communication, but also because it enables the implementation of semantic search in this communication.

4 Conclusion

The present contribution has discussed the possibilities of integration of new technologies, as a suitable support for the leading and solving of projects. Namely,

it has dealt with the integration of semantic search on several portal levels, used for the solution of scientific projects, and also the employment of new communication means from the areas of social and community networks for the support and leading of projects, and the execution of work meetings. It has been shown that the use of these technologies brings about the simplification of communication among the group of solvers, and above all, it increases the effectiveness of the project solution.

Acknowledgments. This work has been supported by the project No P403/11/1103 Construction of Methods for Multi-factorial Assessment of Company Complex Performance in Selected Sectors. The Grant Agency of the Czech Republic (GACR).

5 The References Section

References

1. Altova, XBRL Mapping. http://www.altova.com/mapforce/xbrl-mapping.html
2. Burzagli, L. Como, A. Gabbanini, F.: Towards the Convergence of Web 2.0 and Semantic Web for E-Inclusion. COMPUTERS HELPING PEOPLE WITH SPECIAL NEEDS : Lecture Notes in Computer Science. 6179, 343–350. ISSN 0302-9743. (2010)
3. Buytaert, D.: Drupal, the semantic web and search. http://buytaert.net/drupal-the-semantic-web-and-search
4. Drupal Semantic Search Group, http://drupal.org/project/semantic_search
5. Garca-Crespo, A. Gmez-Berbs, J. M. Colomo-Palacios, R.: Digital libraries and Web 3.0. The CallimachusDL approach. Computers in Human Behavior. 1. ISSN 0747-5632. (2010) http://www.sciencedirect.com/science/article/B6VDC-50T94B0-7/2/d0dbc819f8781a98d1c09074eab38800.
6. IBM Watson, urlhttp://www-03.ibm.com/innovation/us/watson/index.html
7. Trenz, O.: The impact of document format on productivity of information systems. Acta of Mendel University of agriculture and forestry. 6, 177–186. ISSN 1211-8516. (2007)
8. Sklenr, V.: Web 2.0 vs. smantick web. [In Czech] In Inforum 2007 : Sbornk z konference INFORUM 2007 [online]. Praha : Inforum, 2007 [cit. 2011-04-1]. http://www.inforum.cz/pdf/2007/sklenak-vilem.pdf> (2007)
9. Yahoo! SearchMonkey Definition, http://developer.yahoo.com/searchmonkey/