# Using Semantic Web Metadata for Advanced Web Information Retrieval

Martin Švihla and Ivan Jelínek

Department of Computer Science, FEE, CTU in Prague, Czech republic
svihlm1@fel.cvut.cz, jelinek@fel.cvut.cz

**Abstract.** The current web is composed mostly of a great amount of hyperlinked (X)HTML documents. Searching such a large information space or even small part of it can be quite a difficult task, since (X)HTML is not a machine-understandable format and the keyword search has many limits.

The semantic web was developed to overcome these disadvantages by adding machine-understandable metadata to web documents, so that computers can "understand" the meaning of information. According to this idea the web information can be processed automatically, which enables deployment of semantic information retrieval, automatic knowledge sharing or intelligent agents.

In this paper we describe a system that enables a semantic web search in a hyperspace annotated by semantic web metadata. We discuss both sides of the the system - the automatic annotation of existing web resources and the semantic search engine.

## 1 Introduction

Current web information retrieval is based on the keyword search that has many well known limitations. The reason is that most of web documents are in human-oriented formats (HTML, PDF, RTF etc.), which are suitable for the presentation, but machines cannot understand the meaning of published information.

The semantic web technologies (RDF[2] and OWL[4]) are capable to describe a meaning of web page information in a machine-understandable way. The semantic web [1] is meant to be an extension of the current web, in which existing web documents are annotated by machine-understandable metadata. Concerning the domain of an information retrieval, such an extension means the web documents could be searched according to the meaning of their content. That means a search engine would "understand" documents on the annotated web and could find all web pages, where "the person working on project X" is mentioned.

However, these possibilities are not used yet. Though the semantic web standards are already deployed, the web is still not annotated by metadata. The metadata generation and processing are still topics of a research.

In this paper we describe the infrastructure that should enable a semantic search in a particular web space. This work has two steps. In the first step the existing web resources have been annotated by metadata so that every relevant web document is described also by the machine-understandable information. In the

Vojtěch Svátek, Václav Snášel (Eds.): RAWS 2005, pp. 85–89, ISBN 80–248–0864–1.

second step the search engine was built. This search engine crawls both HTML pages and RDF documents. Aggregated data are indexed and stored in a knowledge base, which is queried by the end-user by means of the simple web interface with the semantic search capability.

All these issues are detailed in the following sections.

## 2 Dynamic web page annotation

When creating metadata, we suppose the following scenario: there is a web presentation, which is grounded on data from a relational database. The semantic web metadata should be created to extend existing web resources, but the existing web application should not be changed very much.

In our annotation model the maintainer of existing web presentation is responsible for the generation of metadata. The metadata in RDF format are generated from the same database as HTML web pages. For every relevant HTML page one RDF document is generated with the same information content, then these two representations of the same concept are joined together by hyperlinks.

To enable such a generation of metadata we implemented a system called *METAmorphoses*, described in [5]. This system is able to map a database schema into an ontology structure and generate RDF metadata according to this mapping directly from a database.

The process of annotation is following: first an ontology for the knowledge domain is designed, then the mapping from database schema to this ontology is created and finally RDF documents can be produced. These documents are published by means of HTTP protocol with the result that every RDF document has its own URL. When referencing this RDF from corresponding HTML page, we use special hyperlink designed for this purpose:

```
<link rel="meta" type="application/rdf+xml"
href="sewebis/person.rdf?username=svihlm1" />
```

The architecture of such semantic web extension of classical web application is depicted on figure 1.

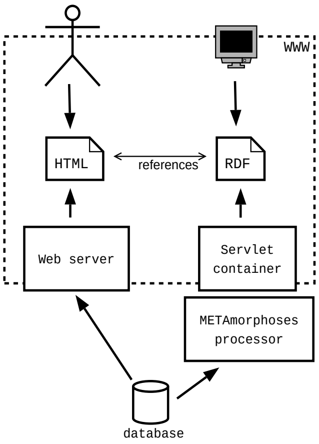
## 3 Crawling web domain

To fetch web content into the search engine index we created two different crawlers, one for crawling the HTML pages and another one for RDF metadata.

These agents cooperate asynchronously. The former crawls HTML pages in a specified domain and searches for hyperlink references to RDF documents. These references are passed to the RDF crawler, the content of HTML pages is not stored.

The RDF crawler is designed to collect the metadata. A basic set of URLs is provided by the HTML crawler, but not all RDF documents must be referenced from HTML. An RDF document can also be referenced from another RDF document (RDFS properties `rdfs:seeAlso` and `rdfs:isDefinedBy`), so that RDF

**Fig. 1.** *METAmorphoses* extends a dynamic web site



agent must also follow these links in order to build a knowledge base. This constitutes defintely an added value as the machine-understandable nature of the metadata allows agents to decide which links to follow according to the meaning of the information. We did not explore these possibilities yet, but our work constitutes a very good environment for research of intelligent agents on the semantic web and we plan to use it in the near future.

## 4 Building search engine index

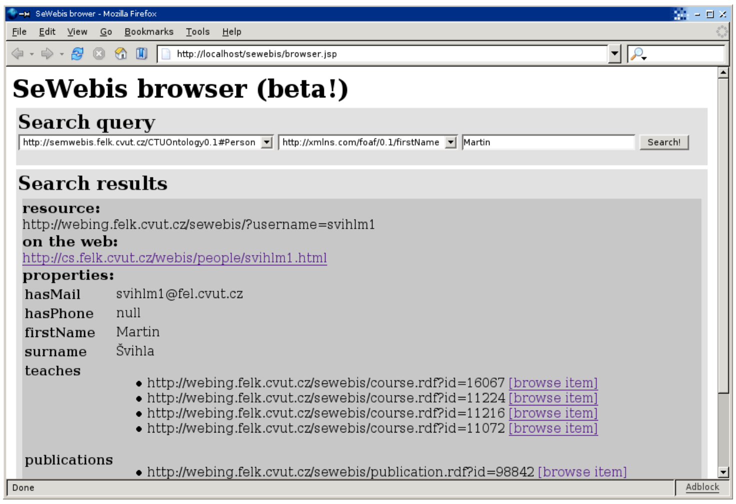
All collected RDF documents are indexed and stored in the knowledge base of the search engine. An index is built so that for every RDF statement in the knowledge base it is possible to track a particular HTML page, which contains an information from the statement.
An RDF document is a set of RDF statements and every statement ($T$ - *triple*) consists of subject ($s$), predicate ($p$) and object ($o$):
$T = (s, p, o)$
To store statements we use the RDF storage system YARS [3], which enables us to mark every triple by a string. We use the URL of the RDF document (*rdf_url*) that contains the marked triple for this purpose. Result is a 4-tuple:
*T=(s, p, o, rdf_url)*
Moreover, to store the information about HTML-to-RDF references another index was created, which is set of ordered pairs *[rdf_url, html_url]*.

Additionally, we store in the knowledge base a set of ontology classes and properties that are used in fetched RDF.
All these data are used later in the semantic search.

## 5 Semantic web search

On the top of the knowledge base we built a web-based user interface (figure 2). When an end-user searches for a particular web page, he or she queries the knowledge base by a meaningful query, which is a simple sentence consisting of subject, predicate and object. A subject and a predicate can be selected from the list of ontology classes and properties in the knowledge base, objects can be only a literal now. An answer for this query is found in metadata, but according to the index structure the search engine is able to assert which web documents are annotated by these particular metadata. Final search result contains not only found RDF statements (in a human-readable format), but also a list of links to classical web resources, as it is common in normal search engines.
This way the end-user searches HTML pages by means of metadata.

**Fig. 2.** Semantic search engine user interface


SeWebis brower - Mozilla Firefox
File
Edit
View
Go
Bookmarks
Tools
Help
http://localhost/sewebis/browser.jsp
Search!
Done
Adblock

SeWebis browser (beta!)

Search query

http://semwebis.felk.cvut.cz/CTUOntology0.1#Person http://xmlns.com/foaf/0.1/firstName Martin

Search results

**resource:**
http://webing.felk.cvut.cz/sewebis/?username=svihlm1
**on the web:**
http://cs.felk.cvut.cz/webis/people/svihlm1.html
**properties:**

| | |
|---|---|
| hasMail | svihlm1@fel.cvut.cz |
| hasPhone | null |
| firstName | Martin |
| surname | Švihla |

teaches

- http://webing.felk.cvut.cz/sewebis/course.rdf?id=16067 [browse item]
- http://webing.felk.cvut.cz/sewebis/course.rdf?id=11224 [browse item]
- http://webing.felk.cvut.cz/sewebis/course.rdf?id=11216 [browse item]
- http://webing.felk.cvut.cz/sewebis/course.rdf?id=11072 [browse item]

publications

- http://webing.felk.cvut.cz/sewebis/publication.rdf?id=98842 [browse item]

## 6 Conclusion and Future Work

In this short paper we described the system that enables a semantic search in a semantically annotated web domain - both web annotation and semantic search engine were discussed.
Though our research is still a work-in-progress, we also deployed the first testing system. We annotated the web information system of the Department of Computer Science at Czech Technical University[1] and we implemented a simple

[1] http://webing.felk.cvut.cz/sewebis/

search engine that semantically searches this web portal. The annotated hyperspace contains over 200 HTML pages about people, publications, projects or courses and we have up to a million statements in our RDF knowledge base. The semantic search user interface is very simple now, but it can be used as a proof-of-concept for the idea of semantic information retrieval.

In the near future we want to improve our search engine so that the more complex semantic queries were possible. We also plan to improve the index structure and examine the combination of semantic search with classical full-text search of HTML resources.

However, the deployed infrastructure is meant mainly as a base for a further work. Main topics of our next research are merging of various information resources on the web, automatic knowledge interchange and semantic intelligent agents.

## 7 Acknowledgements

Research described in the paper was financially supported by the FRV under grant no. 1804/2005 and internal grant of Czech Technical University (IGS) under the external number CTU0507513. It is also a part of "Research in the Area of the Prospective Information and Navigation Technologies" – the grant MSM6840770014 at the Department of Computer Science and Engineering at CTU in Prague.

## References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American, May 2001
2. Beckett, D.: RDF/XML Syntax Specification (Revised). W3C Recommendation 10 February 2004. http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/
3. Harth, A., Decker, S.: Yet Another RDF Store: Perfect Index Structures for Storing Semantic Web Data With Contexts. DERI Technical Report, 2004..
4. Smith, M. K., Welty Ch., McGuinness, D. L.: OWL Web Ontology Language Guide. W3C Recommendation 10 February 2004. http://www.w3.org/TR/2004/REC-owl-guide-20040210/
5. Svihla, M., Jelinek, I.: The Database to RDF Mapping Model for an Easy Semantic Extending of Dynamic Web Sites. To appear in: Proceedings of IADIS International Conference WWW/Internet, ICWI 2005. Lisabon, Portugal (2005).