The Semantic Web in One Day

York Sure[1], Pascal Hitzler[1], Andreas Eberhart[1,a], Rudi Studer[1,2,3]

[1] Institute AIFB, University of Karlsruhe, Karlsruhe, Germany,
{sure,hitzler,eberhart,studer}@aifb.uni-karlsruhe.de

[2] FZI, Research Center for Information Technologies at the University of Karlsruhe, Germany

[3] Ontoprise GmbH, Karlsruhe, Germany

The Challenge

Can you build the Semantic Web in one day?

Most likely your first answer will be "No, that's impossible". Typical projects, in which concrete applications are built, have durations of months or years. And "The Semantic Web" is not yet another application. Primarily it is an idea which requires the realization of multiple applications to become reality, similar to the Web. Building the Web as it is now took years.

Why bother the question?

Let's try answering a simpler question first. Can you build the Web in one day? If you think of the Web in its whole dignity your answer is also negative. What if you think of it magnitudes smaller? You can build html pages in minutes. You can build a (simple) web shop in a few hours, including e.g. the application for web space. At the same time your kids are able to set up a WLAN, install a web server and create a website for the home intranet showing the family vacation pictures. A broad range of different applications can be realized quite easily.

Why is this feasible? Today's end users benefit from the large scales in which web technologies are applied. The strong demand for simplicity resulted in technologies which allow you to set up the basic infrastructure very fast, i.e. hardware and software. By applying off-the-shelf technology you can build your own private Web or your own part of the World Wide Web in one day. All it takes is to integrate several standard technologies to set up your application.

The key challenge for us, the Semantic Web community, is to push technology into a similar direction. To gain momentum, technologies for building private Semantic Webs or parts of the World Wide Semantic Web must necessarily become a commodity and easy to integrate. Many aspects such as scalability, reliability, availability, security etc. have to be considered for real-world applications, but for the moment let's put emphasis on feasibility.

We, members of the Institute AIFB at the University of Karlsruhe, the Research Center for Information Technologies (FZI), and the company Ontoprise, on the occasion of a meeting in conclave, took the chance to make a snapshot of what we can do by applying and assembling existing Semantic Web technologies – in one day. The main aim of our experiment was to get a feeling for the practical applicability of current research work by integrating different

[a] Now Hewlett Packard, Walldorf, Germany, andreas.eberhart@hp.com

technologies into something “up-and-running”. As a side effect we learned a lot about the intersections of the many different directions which Semantic Web research amalgamates such as knowledge representation, natural language processing, peer-to-peer, etc.

The Setting

“24 hours, teams of three or four people, unlimited access to the Web and availability of all tools developed in Karlsruhe” summarizes the key elements of our setting.

To measure what can be done in 24 hours, the experiment was planned by the authors of this article without any prior involvement of the participants. All participants were introduced to the task at the same time, right before the start. During the 24 hours each team had to perform a project cycle with requirement analysis, specification, implementation and, finally, presentation.

The teams were formed from members of the three mentioned institutions in Karlsruhe. They share the interest in Semantic Web, but each member has its own competency profile and context in which she develops and applies Semantic Web technologies. The competencies e.g. include logic, machine learning, natural language processing or software engineering, the contexts range from basic research and prototype development to industrial strength product development. The teams were assembled more or less randomly by following a few simple heuristics such as “bring people with different profiles and working contexts together”.

Each team received a starter-pack CD which contained widely known Semantic Web tools including the ones from our groups, some ontologies and text corpora (e.g. ISWC articles), but also standard Web tools. Additionally, unlimited access to the resources on the personal laptops and the Web were granted. The basic idea was that available technology can be used without limitations.

The problem description given to the teams was rather general in style. They were supposed to design and realize some kind of web information system concerned with publications, authors, research topics, etc. On purpose there was much room for own interpretations of the problem description.

We would like to emphasize the fact that the teams had to do the real hard and challenging work in this experiment. Even though having fun had a high priority, all teams took the challenge quite seriously and were highly motivated. You can already guess what happened, one team called itself “Nightshift”.

The Results

The teams came up with completely different ideas and implementations, typically driven by the experiences and preferences of the team members. Presenting all of them here is out of the scope, but we will highlight exemplary the idea of one group in more detail. The ideas of the other groups are briefly summarized. Further information, especially presentations and descriptions from all contributions, can be found at [1].

The system which we will briefly present was developed by the team “The One” which consisted of Peter Haase, Nenad Stojanovic, Max Völkel, and Johanna Völker. They developed a kind of semantic information retrieval system over abstracts and full texts of scientific publications, which allows for personalized ontology-driven query-refinement, ontology-based browsing by means of custom-learned ontologies, and featured an efficient integrated management of metadata and full texts.

The system was set up by integrating the Bibster and TextToOnto systems with semantic query refinement technology, all of which have been developed in Karlsruhe and are briefly described below. Integration such as the one described had not been attempted or even been considered before. In particular, interoperability was not guaranteed and had to be established on the spot.

TextToOnto [2,3] is a tool suite supporting the semi-automatic construction of ontologies by natural language processing and text mining techniques. It provides the ontology engineer with a variety of algorithms for different ontology learning tasks. In particular, TextToOnto implements various relevance measures for term extraction, algorithms for taxonomy construction as well as several techniques for learning relations between concepts. It is currently being used and extended e.g. in the European Union (EU) SEKT project [4].

Bibster [5,6] is an award-winning semantics-based Peer-to-Peer application aiming at researchers who want to benefit from sharing bibliographic metadata. Many researchers in computer science keep lists of bibliographic metadata, preferably in BibTeX format, that they must laboriously maintain manually. At the same time, many researchers are willing to share these resources, assuming they do not have to invest work in doing so. Bibster supports the management of bibliographic metadata in a Peer-to-Peer fashion: it allows importing bibliographic metadata, e.g. from BibTeX files, into a local knowledge repository, to share and search the knowledge in the Peer-to-Peer system, as well as to edit and export the bibliographic metadata. It was developed as part of the EU SWAP project [7].

Query Refinement [8] is based on incrementally (step-by-step) and interactively tailoring a query to the current information needs of a user, whereas these needs are implicitly elicited by analysing the user's behaviour during the searching process. The gap between a user's need and his query is quantified by measuring several types of query ambiguities, which are used for ranking of the refinements. The main advantage of the approach is a more cooperative support in the refinement process: by exploiting the ontology, the approach supports finding "similar" results and enables efficient refinement of failing queries.

Figure 1 gives a schematic overview of the achieved integration. Metadata and abstracts of publications are fed from Bibster to TextToOnto for automatic ontology generation. The generated ontology in turn provides classification schemata for the bibliographic information, and the classification is performed automatically. The generated ontology also provides the necessary knowledge for semantic query-refinement which enables intelligent ontology-driven query-answering over full texts and abstracts. For the text corpus, the publicly available citeseer database [9] was used, and over 600.000 abstracts processed.

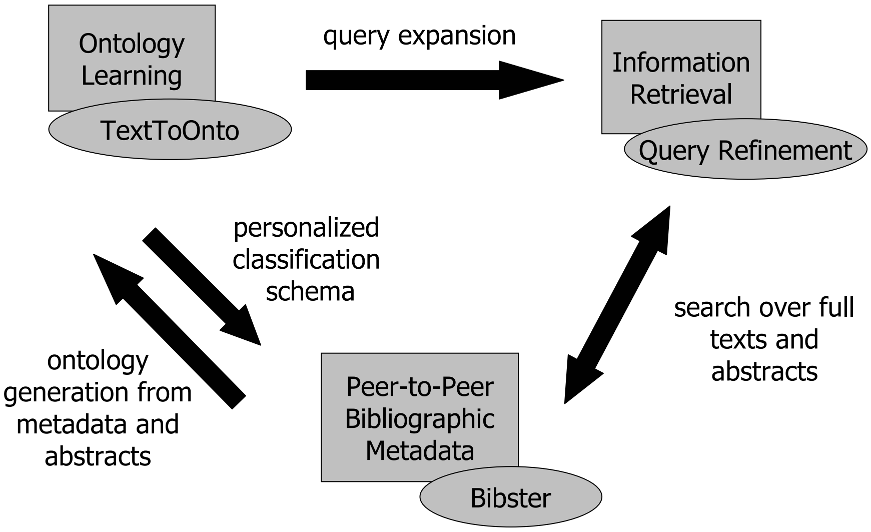
Ontology Learning
TextToOnto
query expansion
Information Retrieval
Query Refinement
personalized classification schema
ontology generation from metadata and abstracts
search over full texts and abstracts
Peer-to-Peer Bibliographic Metadata
Bibster

Figure 1: Integration of TextToOnto, Bibster and Query Refinement

After 24 hours (including a reasonable amount of sleep), the group presented a working system, which has been described in detail in [10]. The achieved interplay between Bibster, TextToOnto, and query refinement techniques yields an intelligent query-answering system which performs semantic searches although the input consists of non-semantic BibTex data and text corpora only. In other words, the user has to provide only BibTeX entries, while the system autonomously performs a semantic analysis of the input data, generates a suitable ontology, and classifies the input data accordingly. Queries posed to the system are also processed intelligently over the generated semantic metadata, taking query refinement techniques into account. Thus semantic technologies allow for intelligent query answering over the input data while the user is not bothered with the tedious process of providing the necessary metadata explicitly. We believe that the achieved interplay between the automatic generation of metadata from raw input and intelligent semantic reasoning techniques is indeed prototypic for successful applications of semantic technologies.

In total we had six groups participating in the experiment. The group “Nightshift” focused on complex query processing with the help of rules and natural language processing. The group “The The” integrated the peer-to-peer system Bibster with the KAON portal, thus making peer-to-peer style community support available via a Web portal. The group “SWSC Candidate” enhanced the search engine Lucene with semantic search capabilities and integrated numerous different data sources such as BibTex files, Amazon, Wikipedia and FOAF data. The “Web” group showed a first prototype of a Semantic Web Browser based on the ontologies openly available. By clicking on objects it was possible to follow their semantic links. An ultimate thrill was the demo of the group “Semantic Web Odyssey” who was inspired by the HAL 9000 system from the movie “2001 – A Space Odyssey” and created a system which answered typed in natural language queries by giving meaningful answers derived from relationships modeled in an ontology. Indeed, we had fun.

Conclusions

After the final presentations of the results we were surprised to see that the systems which emerged after 24 hours were much more sophisticated and functional than what had generally been expected at the start of the experiment. As already noted in the introduction, we believe that the easy and seamless integration of tools and techniques will be a prerequisite for the success of semantic web technologies. However, we did not expect that integration is already possible to the extent realized in our setting. So our main conclusion from the experiment is that applying semantic technologies is already much more feasible than we thought!

On the technical side, syntactic aspects of data integration turned out to be very tedious. Often, output from tool A cannot be used directly as input for tool B, although both have the same language capabilities – e.g. both tools can handle RDF for input and output, but the resulting data is syntactically not compatible to the extent that the tools cannot communicate. These difficulties are aggravated by the fact that there exist different syntactic formats for some ontology languages, e.g. for OWL. We had to invest precious time for coding converters in order to rectify this. As a conclusion, syntactic data conversion turned out to be a major bottleneck, and existing and even established tools are only of limited use for this purpose. This is basically in line with some observations made earlier, viz. that interoperability among ontology tools has potential for improvement [11]. Given increasing amount of developers and tool users of semantic technologies we are quite optimistic that the situation will improve significantly in the near future.

On the other hand, once the syntactic difficulties had been overcome, the semantic content of the data turned out to be very easy to integrate in practice. We noted this with satisfaction, since semantic data integration is one of the main added values of semantic technologies. We also observed that code integration of our tools actually turned out to be surprisingly easy in general.

Of course, the fact that most of the participants were researchers had a large influence on what we were doing. Our ideas, our proposed architectures and our setting itself were largely driven by our day-to-day work. Considering the fact that basic Semantic Web technology is still being developed in international research efforts, and that sophisticated tools and technologies have as yet hardly found significant industrial applications, we found it quite amazing what experts can achieve within only 24 hours. As formal or informal standards become established, and real Semantic Web applications begin to appear, systems will converge and interoperability will increase. Our experiment showed that Semantic Web technology bears the potential of becoming an every-day and easy-to-use ingredient of our knowledge society.

Acknowledgements

We thank all our colleagues from AIFB, FZI and Ontoprise who participated in the experiment for their highly valuable contributions, their collegial competition and their stamina in this endeavor: Andreas Abecker, Sudhir Agarwal, Ernst Biesalski, Stephan Bloehdorn, Saartje Brockmans, Philipp Cimiano, Christian Drumm, Marc Ehrig, Stephan Grimm, Peter Haase, Jens Hartmann, Mark Hefke, Steffen Lamparter, Stefan Mandel, Boris Motik, Daniel Oberle, Lars Schmidt-Thieme, Hans-Peter Schnurr, Björn Schnizler, Wolfgang Sperling, Ljiljana Stojanovic, Nenad Stojanovic, Bernhard Tausch, Christoph Tempich, Denny Vrandecic, Max Völkel, Johanna Völker, Valentin Zacharias.

We thank the people at Schloss Dagstuhl for keeping our work focused by providing a very pleasant surrounding and perfect service.

References

[1] The Semantic Web in One Day, for further details see http://km.aifb.uni-karlsruhe.de/projects/swsc.

[2] Alexander Maedche and Steffen Staab, Ontology Learning. In: S. Staab and R. Studer (eds.), Handbook on Ontologies, Springer Verlag, Heidelberg 2004.

[3] http://sourceforge.net/projects/texttoonto/

[4] http://www.sekt-project.com/

[5] Peter Haase, Jeen Broekstra, Marc Ehrig, Maarten Menken, Peter Mika, Michal Plechawski, Pawel Pyszlak, Björn Schnizler, Ronny Siebes, Steffen Staab and Christoph Tempich, Bibster - A Semantics-Based Bibliographic Peer-to-Peer System. In: Sheila A. McIlraith and Dimitris Plexousakis and Frank van Harmelen, *Proceedings of the Third International Semantic Web Conference, Hiroshima, Japan, 2004*, volume 3298 of LNCS, pp. 122-136. Springer, November 2004.

[6] http://bibster.semanticweb.org/

[7] http://swap.semanticweb.org/

[8] Nenad Stojanovic, Rudi Studer and Ljiljana Stojanovic, An Approach for Step-By-Step Query Refinement in the Ontology-based Information Retrieval. In: Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2004), Beijing, China, September 2004.

[9] http://citeseer.ist.psu.edu/

[10] Peter Haase, Nenad Stojanovic, York Sure and Johanna Völker, On Personalized Information Retrieval in Semantics-Based Peer-to-Peer Systems. In W. Mueller and R. Schenkel, *Proceedings of the BTW-Workshop "WebDB Meets IR"*. 2005.

[11] York Sure, Asun Gomez-Perez, Walter Daelemans, Marie-Laure Reinberger, Nicola Guarino and Natasha Noy, Why Evaluate Ontology Technologies? Because It Works! In: IEEE Intelligent Systems 19 (4): 74-81. July 2004.