Editor: Steffen Staab
University of Koblenz-Landau
staab@uni-koblenz.de

The Semantic Web in One Day

York Sure, Pascal Hitzler, and Andreas Eberhart, *University of Karlsruhe*
Rudi Studer, *University of Karlsruhe and Ontoprise, Karlsruhe*

Can you build the Semantic Web in one day? Most likely your answer would be "No, that's impossible." Typical projects in which concrete applications are built last months or years. And the Semantic Web isn't just another application. Primarily, it's an idea that requires realizing multiple applications to become reality—similar to the Web, which took years to build into what it is now.

Let's look at a simpler question: Can you build the Web in one day? If you consider the Web in its entirety, your answer would also be "No." But what if you think of it as magnitudes smaller? You can build HTML pages in minutes and a simple Web shop in a few hours, including, for example, applying for Web space. All this in the same amount of time it takes your kids to set up a wireless LAN, install a Web server, and create a Web site for the home intranet to show family vacation pictures. You can realize a broad range of applications quite easily.

How is this feasible? (Many aspects such as scalability, reliability, availability, security, and so on, must be considered for real-world applications, but for the moment let's emphasize feasibility.) Today's end users benefit from the large scale in which the industry applies Web technologies. The strong demand for simplicity has resulted in technologies that let you quickly set up the basic infrastructure such as hardware and software. By applying off-the-shelf technology, you can build your private Web or your part of the World Wide Web in one day. All it takes is integrating several standard technologies to set up your application.

The key challenge for us, the Semantic Web community, is to push technology in a similar direction. To gain momentum, technologies for building private Semantic Webs or parts of the World Wide Semantic Web must become a commodity and easy to integrate.

To determine just how far Semantic Web technologies have come, we wanted to create a snapshot of what you could do by applying and assembling existing Semantic Web technologies—in one day. Our experiment's main aim was to get a feel for the practical applicability of current research by integrating different technologies into something "up and running." As an added benefit, we learned a lot about the areas in which the Semantic Web's many research directions intersect, such as knowledge representation, natural language processing, and peer-to-peer.

The scenario

We planned this experiment without the participants' prior involvement so that we could measure what could be done in only 24 hours. We also introduced all participants to the task at the same time—right before they started it. Our scenario had four key elements:

- 24 hours,
- teams of three or four people,
- unlimited access to the Web, and
- availability of all tools developed at Karlsruhe.

The teams received a general problem description, which gave them plenty of room for interpretation. The task was to design and create a Web information system concerned with publications, authors, research topics, and so on. During the 24 hours, each team had to perform a project cycle with requirements analysis, specification, implementation, and presentation.

The six teams comprised members of the Institute AIFB at the University of Karlsruhe, the Research Center for Information Technologies (FZI), and the company Ontoprise. The members share an interest in the Semantic Web, but they all have their own competency profiles and contexts in which they develop and apply Semantic Web technologies. The competencies include, for example, logic, machine learning, natural language processing, and software engineering; the contexts range from basic research and prototype development to industrial strength product development. We assembled the teams 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 that contained widely known Semantic Web tools (including ones from the represented groups), some ontologies and text corpora (such as International Semantic Web Conference articles),

1541-1672/05/$20.00 © 2005 IEEE
Published by the IEEE Computer Society

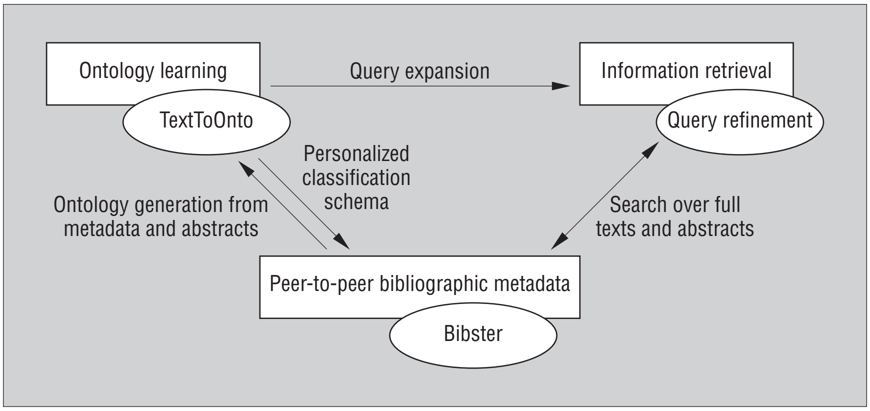


Figure 1. One team integrated TextToOnto, Bibster, and query refinement, as shown in this schematic overview.

and standard Web tools. Additionally, we permitted unlimited access to the resources on personal laptops and the Web; the basic idea was that teams could use available technology without limitations.

The results

The teams came up with completely different ideas and implementations, typically driven by team members' experiences and preferences. We'll highlight one exemplary idea in detail and summarize the others. (For more information on all contributions, including presentations, visit http://km.aifb.uni-karlsruhe.de/projects/swsc.)

A successful prototype

Team The One, comprising Peter Haase, Nenad Stojanovic, Max Völkel, and Johanna Völker, developed a semantic information retrieval system over abstracts and full texts of scientific publications. Their system, which efficiently integrated management of metadata and full texts, enabled personalized, ontology-driven query refinement as well as ontology-based browsing by means of custom-learned ontologies.

The team set up the system by integrating the Bibster and TextToOnto systems, both developed in Karlsruhe (see the sidebar for more information). No one had previously attempted (or even considered) such an integration before. In particular, interoperability wasn't guaranteed, so the team had to establish it on the spot.

Query refinement is based on incrementally and interactively tailoring a query to users' current information needs, whereas search systems typically elicit these needs implicitly by analyzing users' behavior during the searching process.[1] The approach quantifies the gap between users' needs and their queries by measuring several types of query ambiguities, which it uses to rank the refinements. Its main advantage is more cooperative support during refinement: by exploiting the ontology, the method supports finding "similar" results and enables efficient refinement of failing queries.

Figure 1 gives a schematic overview of the achieved integration. The new system feeds publications' metadata and abstracts from Bibster to TextToOnto for automatic ontology generation. The generated ontology in turn provides classification schemata for the bibliographic information; then, the system automatically performs the classification. The generated ontology also provides the knowledge needed for semantic query refinement, which enables intelligent, ontology-driven query answering over full texts and abstracts. As a text corpus, the team used the publicly available CiteSeer.IST database (http://citeseer.ist.psu.edu) and processed over 600,000 abstracts.

After 24 hours (including a reasonable amount of sleep), the group presented a working system.[2] The interplay among the Bibster, TextToOnto, and query refinement techniques yields an intelligent query-answering system that performs semantic searches even though the input comprises only nonsemantic BibTex data and text corpora. In other words, the user must 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 without bothering the user with the tedious process of explicitly providing the necessary metadata. We believe that the interplay between the automatic generation of metadata from raw input and intelligent semantic-reasoning techniques is indeed prototypic for successfully applying semantic technologies.

The ultimate thrill and remaining projects

Although having fun was a high priority, each of the teams who participated in the experiment took the challenge quite seriously and was highly motivated. They had to do the really hard and challenging work.

An ultimate thrill was the demo of the group Semantic Web Odyssey (inspired by the HAL 9000 system from the movie *2001: A Space Odyssey*), who created a system that answered typed-in natural language queries by giving meaningful answers derived from relationships modeled in an ontology. The remaining teams and their projects are as follows:

- *Nightshift* focused on complex query processing with the help of rules and natural language processing.
- *The The* integrated Bibster with the KAON (*Ka*rlsruhe *On*tology) portal, thus making P2P-style community support available through a Web portal.
- *SWSC Candidate* enhanced the Lucene search engine with semantic search capabilities and integrated numerous data sources such as BibTex files, Amazon.com, Wikipedia, and FOAF (friend of a friend) data.
- *Web* showed a first prototype of a Semantic Web browser based on the openly available ontologies. By clicking on objects, users could follow their semantic links.

Our findings

After the teams presented their results, we were surprised to see that the systems that emerged after 24 hours were much more sophisticated and functional than we expected. As we mentioned, we believe that easy, seamless integration of tools and techniques is a prerequisite for Semantic Web technologies' success, but we didn't expect that that integration is already possible to the extent realized in our setting.

On the technical side, syntactic aspects of data integration turned out to be tedious. Often, output from tool A can't be used directly as input for tool B, although both have the same language capabilities. For example, both tools can handle RDF for input and output, but the resulting data is syntactically incompatible to the extent that

TextToOnto and Bibster Semantics-Based Systems

TextToOnto (http://sourceforge.net/projects/texttoonto) is a tool suite that supports constructing ontologies semiautomatically through natural language processing and text mining techniques.[1] The suite 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, and several techniques for learning relations between concepts. Such efforts as the European Union's Semantically Enabled Knowledge Technologies project (www.sekt-project.com) are using and extending TextToOnto.

Bibster (http://bibster.semanticWeb.org) is an award-winning, semantics-based, peer-to-peer application aimed at researchers who want to benefit from sharing bibliographic metadata.[2] Many computer science researchers 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 no work on their part. Bibster supports managing bibliographic metadata in a P2P fashion. It lets you import bibliographic metadata—for example, from BibTeX files—into a local knowledge repository so that you can share and search the knowledge in the P2P system, as well as edit and export the bibliographic metadata. Bibster was developed as part of the EU's Semantic Web and Peer-to-Peer project (http://swap.semanticWeb.org).

References

1. A. Maedche and S. Staab, "Ontology Learning," *Handbook on Ontologies*, S. Staab and R. Studer, eds., Springer-Verlag, 2004, pp. 173–190.
2. P. Haase et al., "Bibster: A Semantics-Based Bibliographic Peer-to-Peer System," *Proc. 3rd Int'l Semantic Web Conf.* (ISWC 2004), LNCS 3298, Springer-Verlag, 2004, pp. 122–136.

the tools can't communicate. The existence of different syntactic formats for some ontology languages, such as OWL, aggravate these difficulties. We had to invest precious time for coding converters to rectify this. So, syntactic data conversion became a major bottleneck, and existing and even established tools were of limited use for this purpose. This finding supports the observation that interoperability among ontology tools has potential for improvement.[3] Given the increasing number of developers and tool users of semantic technologies, we're quite optimistic that the situation will improve significantly in the near future.

Once the teams overcame the syntactic difficulties, the data's semantic content was very easy to integrate. We noted this with satisfaction, because semantic data integration is one of the main added values of semantic technologies. We also observed that code integration of our tools generally turned out to be surprisingly easy.

Of course, the fact that most of the participants were researchers heavily influenced what we were doing. Our ideas, our proposed architectures, and our scenario itself were largely driven by our day-to-day work. Considering that basic Semantic Web technology is still being developed in international research efforts and that sophisticated tools and technologies have hardly found significant industrial applications, we found it quite amazing what experts can achieve in 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 everyday, easy-to-use ingredient of our knowledge society. ■

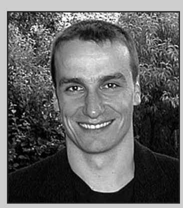
Acknowledgments

We thank our colleagues from AIFB, FZI, and Ontoprise who participated in the experiment for their highly valuable contributions, their collegial competition, and their stamina: 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, and Valentin Zacharias. We also thank the people at Schloss Dagstuhl for keeping our work focused by providing very pleasant surroundings and perfect service.

References

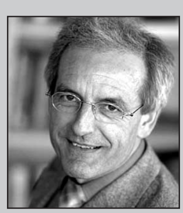
1. N. Stojanovic, R. Studer, and L. Stojanovic, "An Approach for Step-by-Step Query Refinement in the Ontology-Based Information Retrieval," *Proc. 2004 IEEE/WIC/ACM Int'l Conf. Web Intelligence* (WI 2004), IEEE CS Press, 2004, pp. 36–43.
2. P. Haase et al., "On Personalized Information Retrieval in Semantics-Based Peer-to-Peer Systems," *Proc. BTW Workshop: WebDB Meets IR*, W. Mueller and R. Schenkel, eds., Gesellschaft fuer Informatik, 2005, pp. 35–44.
3. Y. Sure et al., "Why Evaluate Ontology Technologies? Because It Works!" *IEEE Intelligent Systems*, vol. 19, no. 4, 2004, pp. 74-81.

York Sure is an assistant professor and project leader at the Institute AIFB, University of Karlsruhe. Contact him at sure@aifb.uni-karlsruhe.de.



Andreas Eberhart, formerly of University of Karlsruhe, is a software architect for virtualization and grid technologies at Hewlett-Packard, Germany. Contact him at andreas.eberhart@hp.com.

Pascal Hitzler is an assistant professor and project leader at the Institute AIFB, University of Karlsruhe. Contact him at hitzler@aifb.uni-karlsruhe.de.



Rudi Studer is head of the Knowledge Management research group at the Institute AIFB, University of Karlsruhe. Contact him at studer@aifb.uni-karlsruhe.de.