Interview with Rudi Studer on "Semantic Technologies"

Rudi Studer is professor at the Institute of Applied Informatics and Formal Description Methods (AIFB) of the Karlsruhe Institute of Technology (KIT) and director at the Karlsruhe Service Research Institute (KSRI). His research focuses on knowledge management, Semantic Web technologies and applications, ontology management, data and text mining, Semantic Web services, service science, and semantic grid. Moreover, Prof. Dr. Rudi Studer is member of the board of FZI Forschungszentrum Informatik and co-founder and member of the scientific board of ontoprise GmbH, Karlsruhe. Prof. Dr. Rudi Studer is engaged in numerous national and international projects, among others in the DFG graduate program "Information Management and Market Engineering (IME)", in the EU Integrated Project NeOn (as technical director), and in the research program THESEUS which is supported by the German Federal Ministry of Economics and Technology. He is vice president of the Semantic Technology Institute International (STI2).

DOI 10.1007/s12599-009-0085-9

Prof. Dr. Rudi Studer
Karlsruher Institut für Technologie (KIT)
Institut AIFB
Englerstraße 11
76131 Karlsruhe
Deutschland

Interview by

Prof. Dr. Ulrich Frank
Chair of Information Systems and Enterprise Modelling
Institute for Computer Science and Business Information Systems
University of Duisburg-Essen
Universitätsstr. 9
45141 Essen
Germany
ulrich.frank@uni-due.de

This article is also available in German in print and via http://www.wirtschaftsinformatik.de: Frank U (2010) Interview mit Rudi Studer zum Thema "Semantische Technologien". WIRTSCHAFTSINFORMATIK. doi: 10.1007/s11576-009-0204-8.

BISE: In recent years so-called semantic technologies have received considerable attention. What are the main achievements of the relevant research?

Studer: It is a challenging task to mention all achievements in such a comprehensive field of research. In regard to the entire research in the field of semantic technologies, however, two developments particularly catch one's eye:

- RDF (Resource Description Framework) and RDFa – the definition of a global standard which made it possible that data exchange on the Web has become reality and will continue to gain importance.[1]
- OWL (2) – a standardized and globally accepted language for representing ontologies. Only this makes a global exchange and reuse of the developed ontologies realistic. Particularly the latest version OWL2 should be highlighted, which was only issued these days by the World Wide Web Consortium as a W3C Recommendation.[2] With this new version OWL has become even more powerful and there are exciting new "profiles" – simplified variants that are tailored to specific applications. The use of the profile QL, for example, makes it possible to fully carry out the inference for answering a query in conventional relational databases (enabled by a clever re-wording of the queries).

In addition, many other breakthroughs could be achieved – such as the progress in the collaborative creation of structured data (particularly embodied in the Semantic MediaWiki (SMW[3]) which has

[1] http://www.w3.org/TR/rdf-concepts/ and http://www.w3.org/TR/2008/WD-xhtml-rdfa-primer-20080620/.
[2] http://www.w3.org/TR/2009/REC-owl2-overview-20091027/.
[3] http://semantic-mediawiki.org/.

been developed here in Karlsruhe by my research group) or the achieved results in automatic reasoning of very large data sets.

BISE: Which of these developments reached business practice?

Studer: There are far more than can be listed here. Data represented in RDF can be found e.g. on millions of Web pages and RDF is the technical basis for embedding metadata in PDF documents. Particularly remarkable is the "Linked (Open) data"[4] movement which uses RDF to publish linked data sets for easy reuse on the Internet (or intranet). Here, we must repeat as a vision what HTML and Google achieved for text documents: it should become possible to easily publish, find, use, and reuse heterogeneously structured data that have been created in a decentralized way – as easily as it is possible today with textual data. In the context of this initiative, in May nearly five billion RDF statements have already been made freely available on the Web – including data about books, countries, companies, genes, and many more. For example, the New York Times has just published their vocabulary used for indexing articles in RDF, making it accessible for everyone.[5]

Also the above mentioned Semantic MediaWiki is currently used for about 200 registered, publicly available websites. Moreover, there are an unknown number of productive intra-organizational uses. With regard to commercial use, semantic search and semantic data integration are certainly the areas that have become most effective in practice. For example, our semantics spin-off ontoprise GmbH at Karlsruhe developed corporate search solutions based on ontologies, metadata, and high-end search technology which have been productively used by customers, such as T-Systems, for many years.

BISE: Unlike other areas of computer science, the transfer into practice seems to work well in the field of semantic technologies. What is the reason?

Studer: Different factors play a role here:

- In many areas semantic technologies can be successfully integrated into an existing software ecosystem. Comprehensive changes are not necessary as e.g. a new semantic search engine is able to improve an otherwise unchanged intranet.
- We usually deal with problems that really "hurt" companies (the search in unstructured text and multimedia data is an example for one such area).
- The necessity for and benefits of semantic technologies are intuitively plausible. Improving search through more background knowledge and a better understanding of the terms entered by the computer is immediately comprehensible.
- The rigorous standardization policy in cooperation with the W3C is an important factor for the industrial acceptance, but also supports concerted research activities.
- In the past decade, Europe could gain an edge through extensive public support for research, for example by the EU with landmark projects such as SEKT, NEON, or NEPOMUK, or by the German Federal Ministry of Economics and Technology and its research program THESEUS.

In areas where we do not find these characteristics, semantic technologies sometimes also have difficulties regarding their transfer into practice.

BISE: In many companies, the lack of data integration reduces the efficiency of information systems. At the same time IT managers are often reluctant to replace legacy applications – and thus: heterogeneous IT environments. Which opportunities do semantic technologies offer to meet this challenge?

Studer: Semantic technologies offer tools and methods at different levels to simplify the management of heterogeneous IT landscapes. Thus, modern tools for semantic integration enable the simple consolidation of data sets from systems with different schemas. In addition, the individual services in an IT environment can be found more easily and can be combined by means of semantic description. The previously mentioned Linked Data approach is another very interesting way to deal with heterogeneous IT landscapes: by means of a stepwise extension of data sources with wrappers providing the data from these systems in semantic formats, the consolidation and use of data from different systems is simplified.

BISE: This is often also connected with a lack of process integration – in companies and in particular in cross-enterprise processes. How can process management benefit from the semantic enrichment of business process models?

Studer: Business process models – especially machine-understandable ones– can with some justification be considered as semantic models in themselves. By enriching them with more expressive semantics and the use of Semantic Web standards, however, a surplus in findability, provability, feasibility, and interchangeability can be achieved:

- *Findability*: Through semantic descriptions of business process components it becomes easier to find appropriate software modules for certain sub-processes or services. For example, an automatic check whether the preconditions of a sub-process are satisfied by the postconditions of the previous step is facilitated.
- *Provability*: Through more powerful formal descriptions of business process components it becomes easier to verify certain properties of the process, especially in the field of GRC (Governance, Risk Management, and Compliance).
- *Feasibility*: Through additional semantic information computers can play a greater role in the automated process execution – for example, they can automatically find and integrate a suitable replacement service in case of the failure of services within a larger process, and thus keep the overall process running.
- *Interchangeability*: The use of standardized modeling languages further simplifies the exchange of business process models.

In the long term, the vision is one of a "Semantic Enterprise" – a unified common, always up-to-date, and collaboratively enhanced digital model of the whole enterprise. In this model approaches merge that have previously been considered separately, such as business process management, business rule management (decision management), ERP, and CRM. The realization of this vision certainly still lies many years ahead (if it is ever fully realized). However, today we can already observe a development in this direction, e.g. by the current convergence of BPM (business process management) and BRMS (Business Rule Management System) software.

BISE: There are two fundamentally different approaches to semantically enrich

[4] http://linkeddata.org/.
[5] http://data.nytimes.com/.

IT artifacts. One approach aims at overcoming the heterogeneity of data structures and their often poorly differentiated semantics through carefully designed, semantically rich reference structures. Thus, a reference semantics is set ex ante in the sense of a *lingua franca*. The other approach assumes that inefficiencies and heterogeneity of factual representations are scarcely to be overcome and instead focuses on reconstructing semantics by cumbersome – and risk-afflicted – analyses. Which approach do you prefer?

Studer: I think this cannot be answered in generalizations – depending on the domain certainly the one or the other approach will be better suited. In many cases, even a combination of the two approaches will be appropriate. The thorough and often manual (and costly) development of reference structures (which may well take place *before* the generation of data) is worthwhile if the reference structures are used very frequently, are of high value or very persistent, or if the cost of failure is very high. An example can be found in the development of a classification of well-known diseases to enable the global integration of disease statistics. The mostly automatic "reconstruction of semantics" is particularly suitable for such domains where a high degree of heterogeneity or dynamics of the data makes the development of reference structures appear too expensive. An example is the handling of data from the system "Google Base" where every user can publish structured data according to his own scheme and where therefore millions of schemata exist.

BISE: There are clear parallels to the research in business and information systems engineering (BISE). This applies not only to shared common research topics, such as business processes, but also to the research objectives. Reference models, for instance, which play an important role in BISE, are comparable with ontologies. Where do you see starting points for a profitable cooperation?

Studer: This is certainly a correct observation and a largely underestimated aspect. With very few exceptions, the focus of work for ontologists is often set on (very powerful) languages and tools, while reference modelers have a great deal of knowledge about domain-specific content, but also meta-knowledge about what "good" models look like. To integrate this domain knowledge in a solid reference ontology, to include the meta-knowledge in the ontology design process and thereby critically question new ontological language constructs in terms of their usefulness and usability on the basis of years of practice in reference modeling, would certainly be rewarding fields.

Basically, the transfer of reference models into ontologies is an interesting idea. For ontology-based systems domain models are indeed not only used in software development, but the software is created to access an explicitly represented domain model. In this way, we achieve a greater reusability on the one hand (the software can be used with different domain models, and a domain model can be used by different software) and may on the other hand also improve and further adapt the software (since the domain model can be changed without programmers and changes to the software being necessary).

BISE: The formal languages used in the field of semantic technologies are largely in the tradition of AI research. They usually allow for deduction, which is a clear advantage compared to common languages of conceptual modeling. At the same time, there is a semantic gap with popular implementation languages. What options do you see to deal with this conflict?

Studer: This is indeed an important and interesting problem. First of all, this gap is hardly more fundamental than that between imperative or object-oriented implementation languages and relational databases – querying a OWL inference engine with SPARQL out of a Java program does not constitute a fundamentally distinct gap to querying a relational database with SQL out of Java or C.

One way to better deal with this conflict is to use declarative languages (such as rule languages) for larger parts of computer programs. Especially in the field of scripting languages for Web pages or in the context of simple applications with many user interfaces (the UI itself is increasingly described declaratively) this seems promising.

Another important possibility is the automatic generation of (wrapper) code from the semantic models – similar to ORM tools (object relational mapping), as they have been known for a long time from the field of relational databases. An example of such a tool is the RDFReactor tool which has been developed in my research group and which can be understood in analogy to ORM tools as object-RDFS mapping.[6]

BISE: The development of ontologies – as well as of reference models – may require an effort that even exceeds the opportunities of major research institutions at universities. At the same time, this is a central research topic that should not be neglected. How can we meet this challenge?

Studer: In the long run, the biggest effort in the development of reference models as well as of reference ontologies has to be made by the experts (or enthusiastic amateurs) in the respective domains themselves; in some cases this may also be a task of BISE or the "applied" computer sciences (geological computer science, medical computer science, ...). The role of basic research in semantic technologies must be seen in providing methods and tools to support these development processes, for which e.g. the (semi-) automated ontology learning from texts or from user interaction may be helpful. As an additional task it is very often necessary to maintain and evolve models during their use. Each tool that has currently been developed for this purpose (such as the tools SMW and Soboleo[7], which have been developed by my research group) will continue to support the distributed collaboration and incremental development and maintenance of such models – so that the effort can be spread over different institutions and longer periods.

BISE: Where do you see the main objectives of future research and which challenges have to be taken into account?

Studer: Major current challenges are very large and "messy" models, the efficient processing of space- and time-related statements, or the population of the semantic web by analyzing unstructured sources. Let me explain:

- Semantic technologies have made great progress in dealing with large data sets[8] – at the same time, however, the usual amount of data has grown considerably and there are still many application problems that cannot be

[6] http://semanticweb.org/wiki/RDFReactor.
[7] http://www.soboleo.com/.
[8] Current triplestores are e.g. able to carry out very simple inferences over some billion RDF statements with response times of about one second.

processed fast enough with prevalent semantic tools. Here, extending the limits is certainly one of the most important objectives.

- Closely connected to the size of the models are internal inconsistencies in these models, which cannot be avoided if a certain size is exceeded and particularly occurs in the case of distributed development or re-use of (sub-)models. To handle this problem in a reasonable way that still remains applicable for very large data sets is one of the important outstanding issues.
- The temporal and spatial dimensions are important qualities in almost every domain which so far cannot be satisfactorily represented by the usual semantic tools. Changing this is currently a core research objective.
- Finally, an obvious idea is to obtain semantically enriched information through automatic analyses of vast amounts of information that are already available on the web in an unstructured form. A recent and very interesting approach is the exploitation of the large redundancies in published data on the Web along with the use of already existing ontologies – leading to the fact that this task does not become more difficult with more data but instead easier to solve. Tom Mitchell gave a well regarded keynote address during the last ISWC on how this might be realized.[9]

These examples show that Semantic Web research still provides exciting and challenging questions despite its many successes – not only for logicians and modelers, but especially in the use and combination of techniques from language processing and information retrieval, machine learning, databases and distributed computing, and many more. Also, the further dissemination of solutions in industrial practice will certainly challenge research with exciting questions in the near future.

[9] Cf. http://rtw.ml.cmu.edu/papers/mitchell-iswc09.pdf.