

Semantic Integration Workshop at the Second International Semantic Web Conference (ISWC-2003)

AnHai Doan, Alon Halevy, and Natalya F. Noy

■ In numerous distributed environments, including today's World Wide Web, enterprise data management systems, large science projects, and the emerging semantic web, applications will inevitably use the information described by multiple ontologies and schemas. We organized the Workshop on Semantic Integration at the Second International Semantic Web Conference to bring together different communities working on the issues of enabling integration among different resources. The workshop generated a lot of interest and attracted more than 70 participants.

In numerous distributed environments, including today's World Wide Web, enterprise data management systems, large science projects, and the emerging semantic web, applications will inevitably use the information described by multiple ontologies and schemas. Interoperability among applications depends critically on the ability to map between them. Semantic integration issues have now become a key bottleneck in the deployment of a wide variety of information management applications. The high cost of this bottleneck has motivated numerous

research activities on methods for describing mappings, manipulating them, and generating them semiautomatically. This research has spanned several communities (databases, AI, World Wide Web), but unfortunately, there has been little cross-fertilization between the communities considering the problem.

To bring these communities together, we organized the Workshop on Semantic Integration at the Second International Semantic Web Conference on Sanibel Island, Florida. In addition to presenting the state of the art of semantic integration research, we wanted to start a discussion on what semantic integration really is, what different communities bring to the table, and how to develop a common research agenda and outline the next big challenges. Hence, the emphasis on the day of the workshop was on discussion rather than formal presentations.

The workshop generated a lot of interest: There were more than 70 registered participants, twice as many as for any other workshop at the conference. We received more than 40 research papers and demonstration proposals for review. The workshop proceedings contain 19 research papers and 7 demonstration descrip-

tions of semantic integration systems that passed a rigorous peer review of the international program committee.¹ Many workshop participants submitted position statements, which also appear in the proceedings. We invite the reader to browse the workshop proceedings for the content of these papers and statements. In this article, we focus on presentations and discussions that are not part of the proceedings.

The format of the workshop reflected our goal of fostering discussion and active exchange of ideas. We had two excellent invited talks by Philip Bernstein (Microsoft Research) and Eduard Hovy (University of Southern California Information Sciences Institute [USC/ISI]).² There were three panel discussions: (1) one on controversial topics in semantic integration, (2) one on automated techniques for mapping definition and discovery, and (3) one on future research directions. There was a lively poster and demonstration session and, despite a large number of participants, active discussion throughout the day.

Invited Talks

The workshop opened with a talk by Philip Bernstein (Microsoft) on a general framework for model management. The goal of model management is to provide a set of high-level operators for manipulating models of data, rather than the data itself. A *model* is a representation of any structure, such as relational database schema, XML schema, and ontology. In the model-management framework, both models and the mappings between them are first-class objects. Models and mappings are manipulated by operators such as Match, Merge, Diff, Compose, and Extract. Bernstein discussed possible semantics of these operators and specific implementation of the operators and their combinations.

Although Bernstein set out an infrastructure on which integration projects can be built, Eduard Hovy (USC/ISI) reported on results from many practical integration projects that his group has performed. Hovy

argued that it is paramount to experiment with different matching techniques, using different heuristics, sources, and combinations of techniques to understand what works and what does not. He also argued that at the current point in ontology research, what's important is building content of ontologies, doing it bit by bit, and figuring out what works and what doesn't rather than concentrating on different formalisms and formal methods. In the experience of his group, many seemingly naïve and informal techniques can tremendously reduce the load on humans in determining mappings between taxonomies or ontologies.

Panel Discussions

By many accounts, the panel discussions were the high points of the workshop. The main questions discussed at the first panel, *What Are They Smoking? Controversial Issues in Semantic Integration* were whether having formal ontologies will facilitate the task of semantic integration, whether we need standard ontologies, and how we should design schemas and ontologies to facilitate integration. The panel participants were Alon Halevy (University of Washington), Pat Hayes (University of West Florida), Len Seligman (MITRE Corporation), and Christopher Welty (IBM).

The original idea behind much of ontology research was that ontologies provide a common language for computer agents to speak. Thus, one point of view expressed at the panel was that if we can get people to agree to use a small number of ontologies (there was a general agreement that one ontology will never be enough), then the semantic integration problem will go away or at least become much more manageable. In fact, one does not even have to designate specific ontologies as standards: By virtue of being on the semantic web, being usable and used by others, some ontologies will become de facto standards. Examples include DUBLIN CORE and the DAML ontology of time. Clusters of agents and applications will then form around these de facto

standards. Thus, the main challenge might be not integrating ontologies and schemas but, rather, enabling people to find out what is already available and how to use it. Others argued that people will not be able to agree even on a small number of ontologies and schemas, and the semantic integration problem will always be a crucial one. Many referred to the experience of the database community that has been addressing the integration problem for the past 30 years. In fact, database designers are adding new formal constraints in each new schema language, but that alone falls far short from solving the integration problem.

Another issue that generated lots of discussion in the audience was how precise integration methods should be. Will having expressive knowledge representation languages help? In particular, one of the main features of the current web is that it is very tolerant to errors. However, if we are building the semantic web to be resilient, isn't formal knowledge representation the wrong way to go? The AI side of the audience argued that descriptions can be precise and still be formal and allow inference engines to deal with representations. One should distinguish between semantics of the language and semantics of what you say in the language. The statement *A is a class* can have precise semantics and be very imprecise about any properties of *A*. In some sense, use of probabilistic reasoning is a "precise way of doing imprecision." However, UML is a great success story, and it has no formal semantics.

In the second panel, *Let's Get Down to Business: Mapping Definition and Discovery*, we discussed and contrasted current approaches to finding mappings. Panelists included Michael Grüninger (NIST), Jérôme Euzenat (INRIA), Fausto Giunchiglia (University of Trento), Li Xu (University of Arizona), and Bernstein. Panelists presented specific methods they used for finding mappings.³ Most of the methods were based on heuristics and included significant input from users. Grüninger presented one exception to this trend: a method that

relied on structural invariance between the models of the theories being mapped, rather than heuristics, providing a segue into the discussion on how much various techniques presented at the panel rely on specific domain and task assumptions or on specific sources, such as WORDNET. In fact, is there too much of a quest for an absolute for having everything "right"? Conceptualization often depends on the domain: In the transportation domain, donkeys are similar to trucks, and in the food domain, donkeys are more similar to cows. On this question, the panel seemed to agree that reliance on specific domain features and sources is not necessarily a bad thing as long as assumptions are made clear from the beginning.

Another question that figured very prominently at the panel is evaluation of mapping techniques. Should we measure results of specific matching algorithms (or their combinations), or should comprehensive tools that would enable users to integrate schemas and ontologies be the measure of our success? Can we develop general tools that will combine all these algorithms and help people in their everyday tasks? The consensus on this question seemed to be that we really need both.

The third panel, *Where Should We Go from Here?* summarized the issues raised during the day and wrapped up the workshop. The panelists were Mike Uschold (Boeing), Christoph Bussler (DERI), Halevy, and Hovy. The panel and the audience were almost unanimous about the need for developing test suites and benchmark problems to provide data and compare performance of different methods. Participants mentioned several ongoing efforts in this area: AnHai Doan (University of Illinois) is collecting test suites for schema and ontology matching, Halevy is building corpora of schemas for statistical schema matching purposes, and Euzenat announced a workshop to develop standards and benchmarks for ontology alignment (to be held in March 2004). The panel also discussed the need for formal frameworks to compare different semantic

Journal of Artificial Intelligence Research

AAAI Is Pleased to Announce
that subscriptions for the print version of *JAIR*
are now available for purchase from the
AAAI Press web site.

www.aaai.org/JAIR

integration solutions.

There was general discussion that in addition to developing formal frameworks for semantic integration, it is crucial to go ahead, get our hands dirty, and just do it. We should be able to build tools, collect lessons on what has been done and what we have learned, develop good ontologies and schemas, and identify best practices.

Acknowledgments

We thank the organizers of the ISWC-2003 conference for their help. The hard work of the program committee members ensured the high quality of the proceedings. We are very grateful to everyone who has participated in the workshop, making it such an exciting event.

Notes

1. Published electronically at ceur-ws.org/Vol-82.
2. Slides from invited talks are available on the workshop web site.
3. Please see the papers in the proceedings

for the description of different mapping discovery methods that the panelists presented.

AnHai Doan is an assistant professor in the Department of Computer Science, University of Illinois at Urbana-Champaign. He obtained a Ph.D. from the University of Washington at Seattle in 2002. His interests span databases and AI, with a current emphasis on schema and ontology matching, object fusion, autonomic data integration, and machine learning. His research on semantic integration was nominated for the Association of Computing Machinery Best Doctoral Dissertation Award, and his teaching on the same topic earned him a spot on the list of excellent teachers as ranked by the University of Illinois students. His e-mail address is anhai@cs.uiuc.edu.



Alon Halevy is an associate professor of computer science at the University of Washington at Seattle. His research interests span the AI and database communities and address various aspects of information integration and semantic



heterogeneity. In particular, he is interested in applying machine learning and knowledge representation to schema mapping and constructing semantic web applications based on peer data-management systems. He received his Ph.D in computer science from Stanford University in 1993 and his Bs.C in computer science and mathematics from Hebrew University in 1988. His e-mail address is alon@u.washington.edu.

Natalya Noy is a research scientist at the Stanford Medical Informatics at Stanford University. She currently works on automatic and semiautomatic tools for managing multiple ontologies. Her interests include ontology development and evaluation, semantic integration of ontologies, and the accessibility of ontology development to experts in noncomputer-science domains. Her e-mail address is noy@smi.stanford.edu.



Computers & Thought



Edward A. Feigenbaum
& Julian Feldman

EDITORS

ISBN 0-262-56092-5 560 pp., index, softcover

The AAAI Press • Distributed by The MIT Press

Massachusetts Institute of Technology, 5 Cambridge Center, Cambridge, Massachusetts 02142

To order, call toll free: (800) 405-1619 <http://mitpress.mit.edu>

MasterCard and VISA accepted.