Semantic Integration Workshop at the 2nd International Semantic Web Conference (ISWC-2003)

http://smi.stanford.edu/si2003/

AnHai Doan
University of Illinois at
Urbana Champaign, IL, USA
anhai@cs.uiuc.edu

Alon Halevy
University of Washington
Seattle, WA, USA
alon@cs.washington.edu

Natalya F. Noy
Stanford University
Stanford, CA, USA
noy@smi.stanford.edu

In numerous distributed environments, including today's World-Wide Web, large scientific projects, enterprise data management, and the emerging Semantic Web, applications will inevitably use information described by multiple schemas and ontologies. Interoperability among applications depends critically on the ability to map between them. However, today, matching between schemas and ontologies is still largely done by hand, in a labor-intensive and error-prone process. As a consequence, 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 semi-automatically. This research has spanned several communities (Databases, AI, WWW), but unfortunately, there has been little cross-fertilization among the communities considering the problem.

To bring these communities together, we organized the Semantic Integration workshop at the Second International Semantic Web Conference, in October 2003. 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, how we develop a common research agenda, and what the next big challenges are. Hence, the emphasis on the day of the workshop was on discussion rather than formal presentations.

The workshop generated significant 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 demo proposals for review. The workshop proceedings (published electronically at http://ceur-ws.org/Vol-82) contain 19 research papers and 7 demo description of semantic-integration systems which passed peer review of the international program committee. Many workshop participants submitted position statements, which also appear in the proceedings. This report focuses on the 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 Phil Bernstein from Microsoft Research and Eduard Hovy from Information Science Institute at USC. (Slides from these talks are available on the workshop homepage.) There were three panel discussions: (a) controversial topics in semantic integration, (b) automated techniques for mapping definition and discovery, and (c) future research directions. There was a lively poster and demo session and, despite a large number of participants, active discussion throughout the day.

Invited Talks

The workshop opened with a talk by Phil Bernstein on model management. Dr. Bernstein discussed his vision and recent work on *model management.* Model management offers programmers a set of high-level operations for manipulating *models* of data and *mappings* between models. A model is a representation of any meta-data structure, such as relational database schema, XML schema, ontology, and so on. Examples of operators include Match, Merge, Diff, Compose, and Extract. Dr. Bernstein discussed possible semantics of these operators and some specific implementations, and argued that a model management system provides a platform in which semantic integration tasks can be performed.

Eduard Hovy, the head of the Natural Language group at ISI, described several practical projects that his group has performed on learning and matching ontologies. Dr. Hovy argued that besides developing formal methods, it is paramount to "get our hands dirty": to experiment with different matching techniques, using different heuristics, sources, and

combinations of techniques to understand what works and what does not. In the experience of his group, even many seemingly naïve and informal techniques, when employed appropriately, can tremendously reduce the load on humans in determining mappings between 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 was moderated by Alon Halevy (University of Washington), and the panelists were 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 on using a small number of ontologies (there was a general agreement that one ontology will never be enough), then semantic integration will become a much more manageable problem. 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 DAML ontology of time. Clusters of agents and applications will then form around these de-facto standards. Thus, the main challenge may 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 semantic integration problem will always remain a crucial one. Furthermore, even if standards exist, we still need to map between local schemas and ontologies and the standard ones. Some participants referred to experience of the database community that has been addressing the integration problem for the past thirty 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. Len Seligman (MITRE) cited a Department of Defense effort to generate 12,000 "standard" data elements, most of which ended up not being used in any system.

Another issue that generated much discussion in the audience was how precise should integration methods 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. If we are building the Semantic Web to be error tolerant as well, isn't formal knowledge representation the wrong way to go? The AI side of the audience argued that descriptions can be imprecise while 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 while being very imprecise about any properties of A. In some sense, use of probabilistic reasoning is a "precise way of doing imprecision." On the other hand, UML is touted by some as a great success story and it has no formal semantics.

Another question that was actively discussed at the panel was whether we are done with research on mappings? Is it all in the user interface now? We have indeed accomplished much, and many of the current techniques can greatly help the mapping process (as the invited talk by Dr. Hovy attested). However, there is a general consensus that in a sense, we only just began, and that numerous semantic integration opportunities and challenges opened up with novel paradigms, such as model management, and with the new data sharing architectures, such as peer to peer, web services, and Semantic Web. Furthermore, the field needs firm experimental grounding.

In the second panel, *"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 Phil Bernstein. Panelists presented specific methods they used for finding mappings (see the proceedings for mapping discovery methods that panelists presented).

Most of the methods employ heuristics and include significant input from users. M. 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 do various techniques presented at the panel rely on specific domain and task assumptions, or 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, while 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 panel was moderated by Mike Uschold (Boeing), and the panelists were Christoph Bussler (DERI), Alon Halevy, and Eduard Hovy. The panel and the audience were almost unanimous on the need for developing test suites and benchmark problems to provide data and to compare performance of different methods. Participants mentioned several ongoing efforts in this area: AnHai Doan is collecting test suites for schema and ontology matching; Alon Halevy is building corpora of schemas for statistical schema matching purposes; Jérôme Euzenat announced a workshop to develop standards and benchmarks for ontology alignment (to be held in March 2004).

Another issue that raised discussion is the need to exploit domain knowledge to aid the matching process. Such domain knowledge can be obtained from experts, schema corpora and multiple ontologies in the domain, and even from the mass of users. It is clear that techniques from the knowledge representation as well as statistical learning communities will be relevant here.

The panel also discussed the need for formal frameworks to compare different semantic integration solutions.

There was general discussion that in addition to developing formal frameworks for semantic integration, it is crucial to just 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 the best practices. It is crucial that we share our semantic-integration lessons in more active ways, so that we have access to the best practice documentation, standards (when reasonable), and design tools when building new models and ontologies. This sharing and cooperation should help us significantly advance the development of the entire area of semantic integration.

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. In light of the great interest in the topic, the SIGMOD Record will dedicate a special issue to Semantic Integration in 2004.