Semantic Integration

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

AnHai Doan
University of Illinois
anhai@cs.uiuc.edu

Alon Y. Halevy
University of Washington
alon@cs.washington.edu

Sharing data across disparate sources often requires solving many semantic heterogeneity problems, such as matching ontologies or schemas, detecting duplicate tuples, reconciling inconsistent data values, modeling complex relations between concepts in different sources, and reasoning with semantic mappings. We refer to this set of problems collectively as *semantic integration*. Over the past two decades, semantic integration has become increasingly crucial to a wide variety of information-processing applications, and has received much attention in the AI, database, WWW, and data mining communities. Until now, however, there has been little cross fertilization across the communities considering the problem.

To assess the current state of research and to draw the communities together, in the Fall of 2003 we organized a workshop on semantic integration at the Second International Semantic Web Conference. The workshop generated significant interest: over 40 research papers and demo proposals were submitted, and 70 leading researchers (from the AI and database communities, government organizations, and industry) attended. We reported on the workshop in this magazine (Volume 25-1, Spring 2004), highlighting many of the discussions and arguments that took place at the workshop.

This special issue follows up on the workshop, and seeks to showcase semantic integration research to the broader community. The issue covers many aspects of semantic integration. The main focus is on ontology-based semantic integration, a topic that has recently received significant attention from AI researchers. However, the issue also discusses matching data tuples and text fragments. In addition, we briefly review semantic integration activities in the database community. For more detail on these activities, we invite the reader to check out a companion special issue in the SIGMOD Record Magazine (Volume 34, December 2004).

One of the most important—and most actively studied—problems in semantic integration is establishing semantic correspondences (also called *mappings*) between vocabularies of different data sources. Given two ontologies, two database schemas, or any other structured resources, how do we determine which concepts are similar or related? Some of the techniques that researchers have applied to answer this question include linguistic analysis of terms, comparison of graphs corresponding to the structures, mapping to a common reference ontology, use of heuristics that look for specific patterns in the concept definitions, and machine-learning techniques.

The first two papers in the issue offer different and complementary solutions to this problem. Michael Grüninger and Joseph Kopena describe using a standard shared ontology for semantic integration in their paper "Semantic Integration through invariants." The authors suggest that the process of semantic integration can be automated significantly if there is an accepted interlingua for expressing common knowledge between ontologies. They show that this approach is feasible in specific domains. In particular, they describe Process Specification Language (PSL), developed at the National Institute for Standards and Technology and endorsed as an International Standard within the International Organization of Standardisation (ISO). PSL is an interlingua for ontologies representing different manufacturing processes. It can be used to share process information among manufacturing systems such as scheduling, process modeling, and process planning. The authors describe an integration architecture where the PSL ontology is at the center and ontologies for specific manufacturing processes are mapped to the PSL ontology. The mappings are specified semi-automatically by presenting ontology developers with a set of questions (in natural language) helping them to map terms in their process-specific ontology to the terms in PSL. The system then generates two-way mappings between the task-specific ontology, such as scheduling, and the PSL interlingua. Note that the generation of these mappings is defined formally and is not based on heuristics. These mappings can be composed to provide mappings between any two task-specific ontologies.

In many settings, however, standard ontologies do not exist. Hence a large body of research has focused on developing semi-automatic methods to discover se-

mantic mappings between disparate ontologies. The paper "Automatic Ontology Matching using Application Semantics" by Avigdor Gal, Giovanni Modica, Hasan Jamil, and Ami Eyal describes one such effort: ontology matching for business applications. They observe that domain semantics are often exhibited in the way business applications are presented to the users and that this consistency can be exploited to improve the accuracy of predicted semantic mappings. For example, when matching ontologies underlying two car-rental reservation systems, they can use the fact that pick-up date comes before drop-off date on each of the reservation forms to match concepts of the ontologies more accurately.

The increasing number of methods available for ontology matching imposes the need to establish a consensus for the evaluation of these methods. There is now a coordinated international initiative to forge this consensus through two events in 2004. The Information Interpretation and Integration Conference (I3CON) at the NIST Performance Metrics for Intelligent Systems Workshop is an ontology alignment (i.e., matching) demonstration competition on the model of the NIST Text Retrieval Conference. This competition focuses on "real-life" test cases and compares global performance of algorithms. The Ontology Alignment Contest of the 3rd Workshop on Evaluation of Ontology-based Tools (EON), at the International Semantic Web Conference (ISWC), targets the characterization of alignment methods with regard to particular ontology features. This contest aims at defining a proper set of benchmark tests for assessing feature-related behaviors.

Matching at the ontology or schema level is only one of several steps in a semantic-integration process. A similar step must be carried out at the *data* level: decide if two given data fragments (e.g., two relational tuples or two paper citations in textual format) match, and if so, how to merge them. Martin Michalowski, Snehal Thakkar, and Craig Knoblock present one such data matching effort in their paper "Automatically Utilizing Secondary Sources to Align Information Across Sources." The authors match entities from various data sources where, for example, names of entities are spelled slightly differently, their attributes, such addresses take different forms, one source can use a full name and another an acronym, and so on. The authors use secondary data sources that are linked to the ones that they are trying to match, to get the additional information necessary to identify identical records. These secondary sources may provide, for instance, longitude and latitude coordinates for an address, thus allowing to match addresses in different forms, listings of officers in a company, allowing to match references to the same officer, and so on.

Most papers in this issue, and indeed most efforts in matching and semantic integration in general, have focused on *structured* artifacts, such as ontologies and data tuples. Real-world data, however, is overwhelmingly in unstructured formats, such as news articles and emails in text, or Web pages in HTML. Managing such data is therefore of pressing concerns, and has received steady attention from several research areas, including natural language processing, question answering, information extraction, and text mining.

In the past few years, however, this attention has been magnified, due in part to the explosion of unstructured data and text on the World-Wide Web, and semantic issues underlying the problem of managing such data have now come to the forefront. The paper "Toward Concept-based Natural Language Processing" by Xin Li, Paul Morie, and Dan Roth discusses some of these semantic integration issues. It focuses on the problem of concept matching in text: given for example two names "JFK" and "John F. Kennedy" in a news article, decide if they refer to the same real-world entity. This problem is similar in some aspects to the tuple matching problem mentioned earlier, but differs in fundamental ways, due to the general lack of structure in text. The authors apply several learning methods to this problem, and offer a principled solution that uses generative models to capture the underlying semantic structure of the application domain. A greater cross fertilization between semantic integration research in the structured data and the text realm is likely to happen in the near future.

Fertile and active as the area of finding mappings between resources is, it is not the only component of semantic-integration research. Once we know correspondences between two sources, we must represent them in a machine-processable way and to use them for specific integration tasks. Researchers have developed a number of ways to represent mappings declaratively. Some examples include representing mappings as instances in an ontology of mappings, defining bridging axioms in first-order logic to represent transformations, and using views to describe mappings from a global ontology to local ontologies.

Once we have correlated the resources, we can tackle the many different tasks that utilize this semantic integration. Prime examples of these tasks includes data transformation from one source to another, merging of ontologies and schemas, robust reading of natural text, query and data mediation in peer to peer settings, and data integration.

Data integration is one of the core application areas that has motivated much research in ontology and schema matching. The paper "Data Integration: A Logic-Based Perspective" by Diego Calvanese and Giuseppe De Giacomo discusses semantic integration issues in this context, using description logics. They present an integration architecture where there is a global ontology that contains the information common to the ontologies that need to be integrated. In this scenario, the global ontology is usually developed after most of the local ontologies have been developed, with the explicit goal of providing common query access to the local ontologies. The authors use the power of description logics to answer queries posed in terms

of the global ontology with data from local ontologies. The paper addresses the perennial trade-offs between expressive power and tractability in computer systems. The authors explain that query answering with common and reasonably expressive description logics is decidable but not tractable. They propose a new subset of description logics called "DL-Lite", which retains a useful subset of primitives but yields good complexity characteristics.

The emerging Semantic Web brought new challenges to the field of semantic integration. Not only the envisioned scale of the Semantic Web is far greater than anything that ontology or data integration researchers have dealt with so far, but also the unique setting of the Web invalidates some of the traditional dataflow assumptions. In their paper, "Ontology Translation for Interoperability Among Semantic Web Services", Mark Burstein and Drew McDermott discuss these challenges. Specifically, the authors consider whether the traditional approach of having mediators to translate between requestors and services will work in an open environment such as the Semantic Web. They discuss a vision of the Semantic Web where web services make the semantics of their services, inputs, outputs, and conditions explicit, thus enabling the automation of matching and composition of services. In such an environment, the authors argue, mappings will need to be published as first-class objects, just as ontologies are in order for different services to use them.

The final paper in the issue, "Semantic Integration Research in the Database Community: A Brief Survey",by AnHai Doan, Natalya Noy, and Alon Halevy, surveys similar research in the database area, discusses future directions, and makes connection between semantic-integration research in the AI and database communities.

As the special issue demonstrates, semantic integration has now become a vibrant research area that spans multiple communities. Its underlying problems are becoming increasingly crucial to a broad range of information processing applications, on the WWW, at enterprises and governments, and at scientific collaboration. Much progress has been made, but many open questions remain, with great payoff potentials. The journey therefore has just barely begun, and much excitement still lies ahead for semantic integration research.