# Peer-Data Management Systems: Plumbing for the Semantic Web

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

The vision of the Semantic Web [Berners-Lee *et al.*2001] offers exciting challenges for research in Artificial Intelligence. In a nutshell, the Semantic Web envisions agents coordinating complex tasks over the Internet. The coordination is facilitated by annotating the data and services on a web destination with rich ontologies, hence enabling semantic interoperation.

In this talk I will discuss some of the challenges raised by the Semantic Web and highlight some dangerous pitfalls. I will illustrate some of the challenges using the Piazza Peer-Data Management System [Gribble *et al.*2001] (PDMS) being developed at the University of Washington.

At a high level, the Piazza project seeks to facilitate wide-scale data sharing among cooperating communities. By *data sharing*, we mean that participants can export and share schemas, base data, and data views constructed through integration over multiple distributed data sources. Topologically, we imagine the cooperating communities to be peer-oriented, self-organizing structures. That is, we assume that (1) each participant is both a client and a server, consuming and producing components of a distributed information database, and that (2) participants join and leave the communities at will. The goal of our research is to design, implement, and deploy the infrastructure on which such data-sharing communities can form and grow.

The Piazza vision is the natural next beyond *information integration* systems. Such systems permit queries over distributed heterogeneous data sources, but they (1) assume that a set of data sources exist and are defined *a priori*, (2) present a uniform interface to the data via a *mediator*, and (3) employ a *single* logical mediated schema at the mediator to specify the semantic mappings between the mediated schema and the schema of the data stored in the sources. In contrast, in Piazza all peers are equal, so a single logical mediated schema is undesirable, impractical, and perhaps even impossible to build.

Some of the challenges we face in Piazza include: (1) managing schema and data heterogeneity, (2) intelligent placement of data on peers in the system to improve performance and availability, (3) efficient propagation of updates to the data, and (4) storage and indexing of descriptions of peer contents. We argue that a PDMS is necessary infrastructure for building a robust Semantic Web.

A crucial element of a PDMS is the ability to map between different models of the same or related domains. A mapping is set of formulae that provide the semantic relationships between the concepts in the models. It is rare that a global ontology or schema can be developed for such a PDMS. In practice, multiple ontologies and schemas will be developed by independent entities, and coordination will require mapping between the different models. There will always be more than one representation of any domain of discourse. Hence, we argue that if knowledge and data are to be shared, the problem of mapping between models is as fundamental as modeling itself.

Currently, whenever mappings are needed (e.g., in information integration systems) they are provided manually in a labor-intensive and error-prone process, which is a major bottleneck to scaling up systems to a large number of sources. I will describe some recent work on using Machine Learning techniques for semi-automatically constructing mappings between domain models. These techniques have been applied initially to relational schema [Doan, Domingos, & Halevy2001] and recently to mapping between taxonomies [Doan *et al.*2002].



## References

[Berners-Lee *et al.*2001] Berners-Lee, T.; Hendler, J.; Lassila, O.; and Web, S. 2001. The semantic web. *Scientific American.*

[Doan *et al.*2002] Doan, A.; Madhavan, J.; Domingos, P.; and Halevy, A. 2002. Learning to map between ontologies on the semantic web. In *Proc. of the Int. WWW Conf.*

[Doan, Domingos, & Halevy2001] Doan, A.; Domingos, P.; and Halevy, A. 2001. Reconciling schemas of disparate data sources: a machine learning approach. In *Proc. of SIGMOD.*

[Gribble *et al.*2001] Gribble, S.; Halevy, A.; Ives, Z.; Rodrig, M.; and Suciu, D. 2001. What can databases do for peer-to-peer? In *ACM SIGMOD WebDB Workshop 2001.*