

# Collective Information Integration using Logical and Statistical Methods

Lise Getoor  
Computer Science Department  
University of Maryland, College Park  
getoor@cs.umd.edu

Renée J. Miller  
Computer Science Department  
University of Toronto  
miller@cs.toronto.edu

Data integration systems are software systems that permit the transformation, integration, and exchange of structured data that has been designed and developed independently. The often subtle and complex interdependencies within data can make the creation, maintenance, and use of such systems quite challenging. We have available a robust arsenal of tools and mechanisms for reconciling semantic differences in how data is represented including views, mappings, and transformation languages. There is also a growing maturity in our knowledge of how to create [MHH00, PVM<sup>+</sup>02] and use these mechanisms in such tasks as query answering, [Hal01], data exchange, [FKMP03], data integration, [Len02], and data sharing. [AHT03, KAM03].

Given this solid foundation in the tools and modeling structures needed for data sharing, we see several remaining challenges. First, the techniques used for learning or deriving the metadata required for integration vary greatly. We have sophisticated methods for learning matches (aka record linkage) between data entities [BG04], between schema components [DR02], and between ontologies [ELB<sup>+</sup>04]. Yet these methods have not been adequately integrated into cohesive matching systems.

Second, we have sophisticated methods for learning mappings (logical expressions relating schemas and ontologies) [PVM<sup>+</sup>02]. Many of these rely on using data design principles and logical inference over schemas and ontologies. While they use the output of matchers, the mapping learning process has not been well integrated with matching. In particular, we conjecture that the integration of these logical inference and learning techniques with statistical learning methods will improve the quality and perhaps the degree of automation that can be achieved in learning the metadata required for semantic integration and interoperability.

Our challenge is centered on developing more robust (and integrated) systems for learning and managing the metadata necessary to achieve semantic integration and sharing of data. A new theory of integration learning and adaptation based on *schema quality* is required, along with new scalable algorithms for acquiring and maintaining the metadata required for data integration and sharing in heterogeneous environments.

The methods required will necessarily build on state-of-the-art methods in probabilistic reasoning, machine learning and data management and will provide a new theory of integration learning and adaptation that exploits both logical and statistical relationships in the data. This will require a *collective* approach to information integration in which matches and mappings depend on each other and are found using a combination of logical and probabilistic inference.

## References

- [AHT03] D. Suciu A. Halevy, Z. Ives and I. Tatarinov. Schema Mediation in Peer Data Management Systems. In *Proc. of the Int'l Conf. on Data Eng.*, 2003.
- [BG04] Indrajit Bhattacharya and Lise Getoor. Iterative Record Linkage for Cleaning and Integration. In *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 2004.
- [DR02] H. H. Do and E. Rahm. COMA - A System for Flexible Combination of Schema Matching Approaches. In *Proc. of the Int'l Conf. on Very Large Data Bases (VLDB)*, pages 610–621, 2002.
- [ELB<sup>+</sup>04] Jirtme Euzenat, Thanh Le Bach, Jeszs Barrasa, Paolo Bouquet, Jan De Bo, Rose Dieng-Kuntz, Marc Ehrig, Manfred Hauswirth, Mustafa Jarrar, Ruben Lara, Diana Maynard, Amedeo Napoli, Giorgos Stamou, Heiner Stuckenschmidt, Pavel Shvaiko, Sergio Tessaris, Sven Van Acker, and Ilya Zaihrayeu. State of the art on ontology alignment. Technical Report 2.2.3, Knowledge Web NoE, 2004.
- [FKMP03] R. Fagin, P. G. Kolaitis, R. J. Miller, and L. Popa. Data Exchange: Semantics and Query Answering. In *Proc. of the Int'l Conf. on Database Theory (ICDT)*, pages 207–224, 2003.
- [Hal01] A. Y. Halevy. Answering Queries Using Views: A Survey. *The Int'l Journal on Very Large Data Bases*, 10(4):270–294, 2001.
- [KAM03] A. Kementsietsidis, M. Arenas, and R. J. Miller. Mapping Data in Peer-to-Peer Systems: Semantics and Algorithmic Issues. In *ACM SIGMOD Int'l Conf. on the Management of Data*, pages 325–336, 2003.
- [Len02] M. Lenzerini. Data Integration: A Theoretical Perspective. In *Proc. of the ACM Symp. on Principles of Database Systems (PODS)*, pages 233–246, 2002.
- [MHH00] R. J. Miller, L. M. Haas, and M. Hernández. Schema Mapping as Query Discovery. In *Proc. of the Int'l Conf. on Very Large Data Bases (VLDB)*, pages 77–88, Cairo, Egypt, September 2000.
- [PVM<sup>+</sup>02] L. Popa, Y. Velegrakis, R. J. Miller, M. A. Hernandez, and R. Fagin. Translating Web Data. In *Proc. of the Int'l Conf. on Very Large Data Bases (VLDB)*, pages 598–609, Hong Kong SAR, China, August 2002.