

# Enabling Ad hoc Information Integration

Craig A. Knoblock  
University of Southern California  
knoblock@isi.edu

## Position Statement

A critical, yet unsolved, problem in the world of information integration is the ability to rapidly and flexibly build ad hoc integration applications. There is an incredible amount of information available: on the Internet, within organizations, or held by individuals. In many cases we can derive tremendous value by integrating the diverse sources of information in novel ways, but this is often quite challenging and typically requires building special-purpose applications. The need for the ability to rapidly and flexibly integrate diverse sources stems from the fact that it is impossible to anticipate and build all of the integrated applications that will be required in the future. For example, scientists by the very nature of their work cannot anticipate which sources of data will need to be integrated or how they will need to integrate the available data. By developing integration frameworks and tools that support this rapid and flexible construction of ad hoc integration applications, we will empower scientists, engineers, and even individuals to both solve their own problems and to develop tools that can be shared with others.

In order to achieve this goal, there are a large number of capabilities required. In the remainder of this paper I briefly outline four of the crucial technologies required to make this vision a reality. These includes tools for rapidly defining integrated applications, the ability to automatically construct semantic models of new sources of data so that they can be integrated into a larger framework, the ability to accurately link information across sources, and ability to organize, extract, and integrate multimedia sources.

**Tools for Rapidly Defining Integrated Applications** In general, it is difficult for users to anticipate exactly which sources they will want to integrate or even how they will want to combine and display the integrated data. In addition, even if such integrated applications exist, each user will typically want to tailor the applications to their own needs. This means that any approach that takes the labor-intensive path of programming a special-purpose application will not scale and will not meet the needs of the users. A great example of the need and power of such tools is the huge interest in building mashups<sup>1</sup> using GoogleMaps. Entire communities on the web have now devoted themselves to constructing such applications, and this enterprise was largely spawned by the fact that Google published the application program interface (API) for GoogleMaps. Constructing a mashup with GoogleMaps still requires programming, but they have greatly simplified the task. See [5] for one approach to defining new integrated applications.

**Semantic Integration** A significant obstacle in constructing an integrated application is understanding and relating one source to another. A simple example might arise in integrating a patient database with the research reports on cancer patients, where an integration system might need to understand that one source separated the name into first name and last name fields, which the other source represented names as the last name followed by first name and separated by a comma. A more complex example might involve integrating sources providing weather data and the need to understand not just that a source is returning a temperature, but whether the temperature represents the current temperature, the forecasted temperature, or the historical average. In order to support the rapid integration of sources requires the ability to construct

---

<sup>1</sup>A *mashup* is an application that uses content from more than one source to create a completely new service.

models of sources that describe the semantic types of the inputs and outputs [7] and a semantic description of the information provided by a source [3]. These capabilities are essential to quickly relate a new source of data to already known sources. Schema matching in the database community has received a great deal of attention over the years, but the more general problem of semantic integration has been largely ignored. See [9] for a set of articles on the general topic of semantic integration.

**Record Linkage** Another significant challenge in integrating data across sources is resolving the naming inconsistencies that inevitably arise. In just about any integration application, a user will want to relate information in one source to another source. The problem is that in order to relate them requires some method of determining which records in one source refer to the same real-world entity as the records in another source. There has been some work in the data mining community in the area of record linkage [2, 8], but much of this work addresses the problem on a pairwise, offline basis and there has been little or no work on the problem of how to seamlessly integrate a record linkage capability directly into an integration system. The only example of a more dynamic record linkage capability is in the Whirl system [4], but that provided only a simple linkage capability based on information retrieval techniques.

**Integration of Multimedia Sources** The integration problems today go far beyond the ability to integrate the data across several databases. There are many diverse types of information, including text documents, audio, video, maps, imagery, etc. An integration system needs to be able to organize, extract and integrate data from these diverse sources, handle the issues that arise in dealing with complex objects, and support the diverse ways in which this information is structured. Consider two critical application areas today: bioinformatics [6] and geospatial data integration [1]. In both areas, the information to be integrated or combined involves complex source types, specialized operations for processing data, and very large repositories of information. Yet, these are representative of the pressing integration problems that need to be solved.

## References

- [1] Peggy Agouris and Arie Croitoru, editors. *Next Generation Geospatial Information: From Digital Image Analysis to SpatioTemporal Databases*. A.A. Balkema Publishers, New York, 2005.
- [2] Mikhail Bilenko and Raymond J. Mooney. Adaptive duplicate detection using learnable string similarity measures. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003)*, pages 39–48, Washington, DC, 2003.
- [3] Mark James Carman and Craig A. Knoblock. Inducing source descriptions for automated web service composition. In *Proceedings of the AAAI 2005 Workshop on Exploring Planning and Scheduling for Web Services, Grid, and Autonomic Computing, Technical Report WS-05-03*. AAAI Press, 2005.
- [4] William W. Cohen. Data integration using similarity joins and a word-based information representation language. *ACM Trans. Inf. Syst.*, 18(3):288–321, 2000.
- [5] Craig A. Knoblock, Pedro Szekely, and Rattapoom Tuchinda. A mixed-initiative system for building mixed-initiative systems. In *Proceedings of the AAAI Fall Symposium on Mixed-Initiative Problem-Solving Assistants*, 2005.
- [6] Zoe Lacroix and Terence Critchlow, editors. *Bioinformatics: Managing Scientific Data*. Elsevier, 2005.
- [7] Kristina Lerman, Anon Plangrasopchok, and Craig Knoblock. Automatically labeling the inputs and outputs of web services. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06)*, Boston, MA, 2006.
- [8] Steven N. Minton, Claude Nanjo, Craig A. Knoblock, Martin Michalowski, and Matthew Michelson. Heterogeneous field matching method for record linkage. In *Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM '05)*, 2005.
- [9] Natalya F. Noy, AnHai Doan, and Alon Y. Halevy. Semantic integration. *AI Magazine, Special Issue on Semantic Integration*, 26(1), Spring 2005.