# Information Integration Research: Summary of NSF IDM Workshop Breakout Session

September, 2003,
Seattle, Washington

Alon Halevy and Chen Li*

January 24, 2004

*With contributions from Philip Bernstein, Kevin Chang Jayavel Shanmugasundaram and Mike Uschold.

# 1 Executive Summary

Information integration systems provide users a uniform interface to a multitude of heterogeneous, independently developed data sources. They free the user from having to locate the data sources, interact with each one in isolation and manually combine data from multiple sources. The applications of information integration systems range from management of data in large enterprises, data sharing amongst government agencies and large scientific projects (e.g., biological research and astronomy), and integration of data sources on the World-Wide Web.

In the past few years we have seen significant progress on many aspects of data integration, including languages for mediation between data sources, query processing techniques for data integration, and the construction of wrappers to data sources. In addition, recent commercial activities have produced tools that efficiently process queries across multiple data sources. To credit our community, it is noted that in many cases the commercial advances were carried out by members of the research community. The commercial experience points us in several important research directions, including managing semantic heterogeneity in a more scalable fashion, the use of domain knowledge in various parts of the system, and transforming these systems from query-only tools to more active data sharing scenarios.

The breakout group behan by discussing the current trends that change the landscape of information integration needs today. In particular, we noted that current challenges are fueled by the substantially increased scale of the number of sources being integrated, the trend towards sharing data via web services, the reliance of corporations on business intelligence to remain competitive, and the new phases of research in the life sciences that require information integration.

We then turned to note the main technical challenges facing the community today. Among others, these challenges include the development of tools for reconciliation of schemas (and ontologies), developing more flexible architectures for data sharing that do not rely on a central logical schema, the need to support updates and active aspects of such systems, the use of domain knowledge for various aspects of data integration (e.g., assisting users to pose queries against a multitude of sources), supporting flexible querying and summarization and explanation of results, and dealing with uncertain, inconsistent and unstructured data in a principled fashion.

Finally, we noted several recommendations. In terms of research style, we recommend to encourage the development of benchmarks (and, in particular, we recommend that funding agencies should make all of its own data available as a test case). We encourage both theoretical research on foundations of information integration as well as system and toolkit building. Finally, we noted that further progress in this area will be significantly accelerated by combining expertise from Database Systems, Artificial Intelligence, and Information Retrieval.

# 2 Background and Motivation

Information integration systems provide users a uniform interface to a multitude of heterogeneous, independently developed data sources. They free the user from having to locate the data sources, interact with each one in isolation and manually combine data from multiple sources. The applications of information integration systems span many realms. Large enterprises often have dozens of data sources whose integration results in better business processes, customer care and creation of external portals. On the World-Wide Web (WWW) there are currently hundreds of thousands of structured databases hidden behind web forms. Here too, querying across web sources can yield very valuable services. Similarly, in government, sharing data across (or within) agencies is deemed to be very valuable. Finally, large scientific projects are predicated on the ability to share data

across research groups, a prime example being medical and biological research.

In the past few years we have seen significant progress on many aspects of data integration. The community has developed flexible architectures for data integration, powerful methods for mediating between disparate data sources, tools for rapid wrapping of data sources, methods for optimizing queries across multiple data sources, and more. This progress has been the direct result of funding from NSF and other agencies in the fields of Database Systems, Artificial Intelligence and Information Retrieval.

In parallel, in the last few years we have seen significant activity in the commercial world. In addition to several startup companies selling products for information integration, some of the big players in the data management industry (e.g., IBM, BEA) have started offering such products. The commercial developments *were directly influenced by the research results.* The vast majority of the data integration products were developed by members of the research community, either in the form of company creation or by commercializing technology from industrial research labs. For the first time, the experience from the commercial world provides significant feedback on directions for research in information integration. In a nutshell, current products do a reasonable job at processing queries efficiently over multiple data sources, even employing an XML data model. However, much more work is needed on the higher-levels of the system, including managing semantic heterogeneity in a more scalable fashion, the use of domain knowledge in various parts of the system, transforming these systems from query-only tools to more active data sharing scenarios, and easy management of data integration systems.

Information integration has been a subject of research for more than a couple of decades now. In fact, the challenges raised by reconciling heterogeneity across multiple sources is one of the longest standing problems for the database research community [BDD$^+$89, SSU91, AAB$^+$03]. We began the discussion of the breakout group by highlighting what is different about information integration today compared to the past. First, we noted that the emergence of the WWW and related technologies completely changed the landscape: the WWW provides access to many valuable structured data sources at a scale not seen before, and the standards underlying web services greatly facilitate sharing of data among corporations. Instead of becoming an option, data sharing has become a necessity. Second, business practices are changing to rely on information integration – in order to stay competitive, corporations must employ tools for business intelligence and those, in turn, must glean data from multiple sources. Third, recent events have underscored the need for data sharing among government agencies, and life sciences have reached the point where data sharing is crucial in order to make sustained progress. Fourth, personal information management (PIM) is starting to receive significant attention from both the research community and the commercial world. A significant key to effective PIM is the ability to integrate data from multiple sources. Finally, the group noted that information integration is an inherently hard problem that cannot be solved by just a few years of research.

In the remainder of this report we describe our discussion on the specific challenges for the research community and put forth several recommendations.

## 3 Current Challenges

The long-term goal of information integration research is to build systems that are able to provide seamless access to a multitude of independently developed, heterogeneous data sources. These systems should have the following capabilities:

- integrate sources at scale (hundreds of thousands of sources),

- support automated discovery of new data sources,
- be easy to configure, manage and maintain,
- protect data privacy,
- incorporate structured, semi-structured, text, multimedia, and data streams, and possibly inconsistent data,
- provide flexible querying and exploration of the sources and the data,
- adapt in the presence of unreliable sources, and
- support secure data access.

### Specific Challenges

Some of the specific challenges the community faces at the moment are the following:

**Reconciling heterogeneous schemas/ontologies:** as noted time and time again, the fundamental problem in any data sharing application is that systems are heterogeneous: they have different ways of representing data and/or knowledge about the world, and they use different representation mechanisms (e.g., relational databases, legacy systems, XML schemas, ontologies). In order to share data between heterogeneous sources we need a *semantic mapping* between their respective representations. The process of generating these mappings (often referred to as schema matching or ontology alignment) is inherently a human-centered task. Hence, research in this area develops tools that aid a human designer and increases their productivity. Except for developing new matching techniques, the key challenges here are incorporating a variety of methods in principled fashion, building systems that *improve* over time, and scaling up to large schemas.

**Data-sharing with no central control:** Traditional data-integration systems use a centralized mediation approach, in which a centralized mediator, employing a *mediated schema*, accepts user queries and reformulates them over the schemas of the different sources. However, mediated schemas are often hard to agree upon, construct and maintain. For instance, labs conducting bioinformatics research are willing to share their experimental results with each other, but they do it in an ad-hoc fashion, without any central control. A similar scenario is found in data sharing among government agencies. We need to develop architectures that enable large-scale sharing of data with no central control, either at the logical or physical levels.

**On-the-fly integration:** Currently, data integration systems rely on relatively static configuration with a set of long-lived data sources. On-the-fly integration refers to scenarios where we want integrate data from a source immediately after discovering it, and we may use a source only a few times for a particular set of tasks. The challenge is to significantly reduce the time and skill needed to integrate data sources.

**Source discovery and deep-web integration:** Over the past few years, the Web has deepened dramatically - a significant and increasing amount of information is hidden behind the query interfaces of searchable databases. The potential of integrating data from a large number of data sources is enormous. The challenges include discovering sources automatically, integrating them appropriately, supporting efficient query processing of user queries, and developing mass collaboration techniques for the management, description and rating of such sources.

**Management of changes in data integration:** in the simplest form, data integration systems need to be able to handle updates to the data sources through the integrated view. Such updates need to be in accordance with policies set forth by the component systems (and the interfaces they support). As a next step, we would like data integration systems to be a basis for distributed collaboration between a set of data producers. More generally, we see the need for more "active" information integration systems. For example, in scientific data sharing, users may wish to be notified (without repeatedly issuing a query) whenever new data of a certain kind becomes available in any one of the distributed data stores. Similar issues arise in the case of distributed data streams, such as streaming stock quotes. Active information integration systems are thus broadly similar to publish/subscribe systems, but have to address additional aspects such as heterogeneity of data sources and large-scale distribution.

**Combining structured and unstructured data:** we often see the need to integrate data from structured sources (e.g., databases, XML documents) and unstructured text (e.g., web pages). While there has been progress in limited settings (e.g., adding keyword search capabilities to database systems), there are fundamental issues in this area that need to be addressed. In particular, understanding the variety of queries combining structured and unstructured data needs to be investigated along with the semantics of such queries. In addition, we need to develop languages that are appropriate for such queries (taking into consideration their usability) and efficient methods for processing them.

**Managing inconsistency and uncertainty:** when integrating data from multiple sources, the data may be inconsistent or uncertain. We need to develop methods for managing such data, explaining inconsistencies and uncertainties to users, and for guiding them through the process of reconciling different data sets.

**The use of domain knowledge:** Domain knowledge can play a key role in the usability of data integration systems. Knowledge can be used to guide the user to find appropriate sources of data, to facilitate the formulation of queries, when the terminology is complex, for formulating results in more succinct ways and explanations for the results, and in expressing the semantic mappings between data sources. We need to develop methods for constructing domain knowledge in such a way that it can be used for these tasks, and ensuring that the knowledge evolves as new sources are added and as the semantics of the data sources evolve.

**Interface integration and lineage:** data sources provide access to the data through a particular set of visualization techniques that enable users to browse and drill down into the data. Hence, integrating data from multiple sources requires that we combine the different visualizations in sensible ways. We need to develop the principles that enable us to state such combinations. In addition, we need techniques for tracing the origins (lineage) of data when answers are produced from multiple sources.

**Security and privacy:** in many scenarios owners of data will be reluctant to share their data without the appropriate security policies in place and ensuring data privacy, when required. Research is required for specifying and implementing processes for ensuring security and privacy.

# 4 Recommendations

The discussion of the group resulted in the following recommendations:

1. We recommend to encourage research on information integration in the fields of Database Systems, Artificial Intelligence and Information Retrieval. Of particular importance are works

that span two or more of these fields. The work required involves both single-PI grants and larger-scale efforts.

2. Researchers should be encouraged to create toolkits for data integration that can be shared among researchers. These toolkits should remove the need for implementing an entire data integration system from scratch for every project, and will facilitate large-scale collaborations.

3. The work required in the field spans the spectrum of theoretical work on the foundations underlying different aspects of information integration, and work that develops and evaluates information integration systems. In the past, the theoretical work has had considerable impact (e.g., work on mediation languages and reformulation techniques), and the systems work led to both commercial products and the unveiling of subsequent challenges for data integration.

4. It is crucial that we develop a set of benchmarks on which we can compare different techniques in information integration. An example of such a benchmark would include a collection of heterogeneous schemas, on which one can evaluate techniques for dealing with heterogeneity, automated schema matching and on reformulation algorithms.

5. Agencies funding information integration should make its own data available as public data sources (to the extent compliant with the privacy requirements). NSF can then pose specific information integration challenge tasks.

6. Many of the computing grand challenges (e.g., those proposed by the CRA) require an information integration substrate. Funding agencies should try to ensure that the efforts on the grand challenges and on information integration are aware of each other and coordinated to the extent possible.

7. While workshops in information integration occur on a pretty regular basis, both at Database and Artificial Intelligence conferences, we should encourage more targeted one-time workshops in this area. In particular, bringing researchers together from multiple fields and connecting computer scientists with researchers from other fields of science is highly encouraged.

# References


[AAB+03] Serge Abiteboul, Rakesh Agrawal, Phil Bernstein, Mike Carey, Stefano Ceri, Bruce Croft, David DeWitt, Mike Franklin, Hector Garcia-Molina, Dieter Galwick, Jim Gray, Laura Haas, Alon Halevy, Joe Hellerstein, Yannis Ioannidis, Martin Kersten, Michael Pazzani, Mike Lesk, David Maier, Jeff Naughton, Hans Schek, Timos Sellis, Avi Silberschatz, Mike Stonebraker, Rick Snodgrass, Jeff Ullman, Gerhard Weikum, Jennifer Widom, and Stand Zdonik. The lowell database research self assessment. *CACM, to appear*, 2003.

[BDD+89] Philip A. Bernstein, Umeshwar Dayal, David J. DeWitt, Dieter Gawlick, Jim Gray, Matthias Jarke, Bruce G. Lindsay, Peter C. Lockemann, David Maier, Erich J. Neuhold, Andreas Reuter, Lawrence A. Rowe, Hans-Jörg Schek, Joachim W. Schmidt, Michael Schrefl, and Michael Stonebraker. Future directions in dbms research - the laguna beach participants. *SIGMOD Record*, 18(1):17–26, 1989.

[SSU91] Abraham Silberschatz, Michael Stonebraker, and Jeffrey D. Ullman. Database systems: Achievements and opportunities. *CACM*, 34(10):110–120, 1991.