The first step in interoperating among multidatabases is semantic integration: Producing attribute correspondences that describe relationships between attributes or classes in different database schemas. Dynamic integration requires the ability to automatically extract database semantics, express them as metadata, and match semantically equivalent data elements to produce attribute correspondences. This process can not be “pre-programmed” since the information to be accessed is heterogeneous. In this paper we present an architecture supporting dynamic integration. We first overview a tool, Semint, for automated semantic integration that helps database administrators generate attribute correspondences. We then introduce a novel framework for dynamic integration and a query language for multidatabase systems that uses Semint as part of a complete semantic integration service. Our framework supports dynamic integration as well as incremental integration. We show the advantages of our framework in an environment where full integration is not desired or complete knowledge of the databases to be integrated is unavailable.

Applications in a wide variety of industries require access to multiple heterogeneous databases due to company mergers, the introduction of new database technology, or integrating information across departments. There are two ways to integrate existing information systems: reengineering and interoperating. In reengineering, the application logic, data definition, and data from the old systems are transferred into a new system. The advantage of reengineering is that the new systems are easier to maintain. However, this process is expensive and complicated.

The second option is to allow the existing systems to interoper as a multidatabase system by putting a new, standard interface on existing databases. This preserves the data and applications from the existing databases, yet allows access to the data from new multidatabase applications. This transforms the existing databases into open databases. Interoperability allows computing resources to be shared, thereby giving organizations a new resource and a uniform, enterprise-wide and database-independent view of data.

The goal of integrating interoperable multidatabases is to provide an enterprise-wide information system where users can access information through an unified interface without knowing the details of each participating component database. This gives a computer paradigm of a large number of information sources that are heterogeneous in semantics and format. Automatically extracting and understanding the semantics of information and format conversion are important issues to information integration and such metadata should be based on the contents of the information available.

Techniques essential to information integration include extracting semantics, transforming formats, identifying attribute correspondence, resolving heterogeneity, multidatabase query processing, and data integration. In order to answer queries in multidatabase systems, three distinct processes need to be performed by the user, database administrator, and/or system as shown in Figure 1. The Schema Integration process includes a possible schema transformation step, followed by correspondence identification, and an object integration and mapping construction step [Parent and Spaccapietra, 1995]. In Query Processing, global queries are reformulated into subqueries, the subqueries are executed at the local sites,
and their results are assembled at a final site. The Data Integration process is complimentary to Query Processing, i.e., it determines how the results from different local databases should be merged and presented at the final site.

The concept of dynamic integration rests on the premise that knowledge (metadata) needed to integrate information is provided by or can be extracted from databases directly. This allows computer “mediators” to handle the heterogeneity so end-users and non-computer experts can access useful data. The formats of data and its semantics presentation are not expected to be standardized, developing schemas to integrate this data is a mediation task [Wiederhold, 1992]. An automatic process for extracting and transforming data semantics is also an important issue as the volume and variation of accessible data increases.

**The Problem**

We are concerned with one area in interoperable multidatabases: dynamic integration. We argue that attribute correspondence identification is a primary bottleneck in integration and query processing of multidatabases. For example, users of heterogeneous databases may issue a query that joins two relations in different databases; to do this they must know what attributes in the relations can be used as a join key. If the attribute correspondences between the databases are known, finding a join key is a simple process. However, if these databases have not been integrated or users are not familiar with the attribute correspondences, multidatabase queries cannot be issued. Attribute correspondences are also essential to data integration — merging results from component databases.

We argue that the solution to this problem is to automate the process of attribute correspondence identification. Human effort is still necessary in database integration unless the semantics of data can be captured completely and the techniques of artificial intelligence is mature to act as an domain expert. Domain knowledge is a necessary part of any semantic integration system; we cannot count on the available metadata completely and accurately describing the semantics of the data. We categorize the approaches to attribute correspondence identification, based on when human effort is required, as follows:

**Manual integration:** Human effort is needed to collect metadata from databases that captures the semantics of data. The semantics may be embodied within a database model, a conceptual schema, application programs, or the minds of users. Humans need to be involved in collecting metadata and matching corresponding attributes. The manual integration is tedious and time consuming. Some corporate experiences of manual integration were described in [Ventrone and Heiler, 1994].

**Semiautomated integration:** In this approach, tools are used to collect metadata and generate candidates for corresponding attributes. DBAs (Database Administrators) need to be involved in checking and confirming the recommendations of the tools. Multidatabase queries can not be issued until attribute correspondences are generated. An example of this kind of tool is described in [Li and Clifton, 1994,Li and Clifton, 1995].

**Dynamic integration:** Users can issue queries before attribute correspondences are generated. The attribute correspondence identification process is carried out dynamically. Collecting metadata and generating candidate attribute correspondences are automated. Users need to check multidatabase query results rather than attribute correspondences. As users are generally “domain experts” on data stored in databases (or at least the data they are interested in), as opposed to experts in schema design, the query results are more meaningful to users than attribute correspondences.

An interface providing semiautomated or dynamic integration needs to utilize available metadata that can be automatically extracted from databases. An automated tool for identifying attribute correspondences can help to open this bottleneck, allowing dynamic integration and query process-
Related Content

View Materialization in a Data Cube: Optimization Models and Heuristics
[www.igi-global.com/article/view-materialization-data-cube/3372?camid=4v1a](www.igi-global.com/article/view-materialization-data-cube/3372?camid=4v1a)

Applying UML and XML for Designing and Interchanging Information for Data Warehouses and OLAP Applications
[www.igi-global.com/article/applying-uml-xml-designing-interchanging/3305?camid=4v1a](www.igi-global.com/article/applying-uml-xml-designing-interchanging/3305?camid=4v1a)

Managing Uncertainties in Image Databases
[www.igi-global.com/chapter/managing-uncertainties-image-databases/8036?camid=4v1a](www.igi-global.com/chapter/managing-uncertainties-image-databases/8036?camid=4v1a)

Querical Data Networks
[www.igi-global.com/chapter/querical-data-networks/20764?camid=4v1a](www.igi-global.com/chapter/querical-data-networks/20764?camid=4v1a)