Betsy Rolland
LIS 540D
Dr. Terry Brooks
June 2, 2006

Information is Perception:
XML Schema Mediation

> While schema heterogeneity is challenging for humans, it is drastically more challenging for programs. A program is given only the two schemas to reconcile, but those schemas are merely symbols. They do not capture the entire meaning or intent of the schemas—those are only in the minds of the designers. (Halevy, 2005, p.5).

Organizations are increasingly storing and exchanging data in the XML format. Frequently, related data will be stored in different XML files with different XML schemas. XML schema mediation, or integration, is a technique for integrating the data from those different schemas. This could mean creating an entirely new schema, bringing a portion of one schema into another or taking pieces of several different schemas and integrating them into another.

Sometimes schema mediation means simply understanding that, for example, the "Name" field in one schema is the same thing as the "Author" field in another and the data can simply be copied over. More frequently, however, the data in one or more XML documents needs to be manipulated somehow. For example, one file may have "Last Name, First Name" in one field while the other file has separate fields for "Last Name" and "First Name." In order to be integrated, either the "Last Name, First Name" data will need to be split or the "Last Name" and "First Name" data will need to be concatenated.

In this way, schema mediation supports the premise that "Information is Perception." Until a human interprets them, the names of the elements and attributes are simply marks on a page. Are <last_name> and <surname> the same thing? Not to a computer, but to a human they are. The computer can only deal with them as separate pieces of data, but a human can make the connections that allow these two elements to be matched.

It is worth noting that the process of XML schema mediation is very similar to schema mediation in the world of relational databases. Most of the research that has been done in that area can be easily transferred to the study of XML schema mediation.

Disparate Data

As the borders of organizations blur and dissolve, it is especially likely that organizations will regularly need to integrate data from their partners – suppliers, customers, manufacturers, shippers – into their own data sources. This requires some sort of standardization or transformation of the data format to integrate the data. "…[F]or rich data to be shared among different groups, all concepts need to be placed into a common frame of reference. XML schemas must be completely standardized across groups, or mappings must be created between all pairs of related data sources" (Halevy, p.1).

Even within an organization, it is likely that different departments and teams will have data that is formatted in vastly different ways. "Since the schemas are independently developed, they often have different structures and terminology. This can obviously occur when the schemas are from different domains, such as a real estate schema and property tax schema. However, it also occurs even if they model the same real world domain, just because they were developed by different people in different real-world contexts" (Rahm, p.335). Different departments in an organization have different information needs and views of the business. This translates into different information structures.

Changes in an organization's structure and mission can also create data integration issues. "There are many reasons why data in enterprises resides in multiple sources in what appears to be a haphazard fashion. First, many data systems were developed independently for targeted business needs, but when the business needs changed, data needed to be shared between different parts of the organization. Second, enterprises acquire many data sources as a result of mergers and acquisitions" (Halevy (2005), p.3). Additionally, poor planning and control contribute to the problem. Organizations fail to plan for future data needs and end up with a hodgepodge of data schemas. New projects aren't generally encouraged to take advantage of existing information structures.

When Schema Mediation Makes Sense

There are several scenarios under which schema mediation can be useful.

- *Integrating information from across the company*

  In most organizations, different departments develop and maintain their own data sources, organized in schemas that make sense to them. These data silos are easier to develop and maintain, but keep the organization from seeing the big picture its data might provide. Data silos are also inefficient and error-prone as different departments maintain copies of the same data. Creating a mediated schema is one way to integrate this information into a single data source that can then be used to make more informed business decisions.

One example of this is when both engineering and marketing departments in a software engineering firm are collecting customer data. The marketing department may focus its data collection efforts on what features customers want, while the engineering department focuses on how well certain features perform. Obviously, integrating this information into one larger view would be immensely useful.

- *Integrating external data into internal data stores*

  Another possible use for schema mediation is integrating information from external data sources, like suppliers and customers. A garden supply business could ask its suppliers to submit their inventory information regularly in XML format. It is unlikely that the data formats for the two organizations will match initially; one party is going to need to somehow transform its data to match the other's.

  In many ways, small organizations have the most to gain from this type of schema mediation. For example, a museum of Native American art that maintains a database of art for native peoples around the world would be best served by taking advantage of information from other museums. Again, it is highly unlikely that the information from the other museums will be structured in the same way as its own.

- *Migrating from one application to another, new view of data*

  Another use for schema mediation is in the process of migrating data from one application to another. A company might use a sales database application to maintain all of its client information, including contacts, client history and order information. If they decide to switch to a new application or even a new version of the existing application, their data will probably need to be migrated. The data will need to be manipulated to fit into the new schema of the new application. Generally, the application developers offer customers an automated way to do this, but if the end users have altered the schema at all to better fit their own information structures, the upgrade process will also need to be altered.

- *Web services*

  Web services return information in XML format. Using the accompanying schema, clients using the web service can mediate between the schemas to integrate the incoming information. "Data mediation tackles the problems of how to bridge between two data models that may have differences in terms of semantics and syntax. This is a significant issue for Web service communication where the server requester and provider are often heterogeneous and autonomous entities with independent data models (Moran, p.3).

- *Querying multiple data sources*

  Some organizations might simply want to be able to search other organizations' databases rather than actually integrating their information into the home database. Mediated schemas are an excellent way to do this. As discussed below, ontologies are often used in this process, as in the Bio-Mediator project here at the UW (Mork, p.1). This approach is also called the "federated" approach and can be characterized as follows:

> The benefit of this work is that it can preserve the autonomy of the particular data warehouses and their applications, and users outside this framework should not be aware of how many data warehouses exist under the framework. When a user poses a global query on the system, the system will decompose the global query and send the obtained sub-queries to the mediators. Then the mediators send those sub-queries to corresponding local data warehouses. After local data warehouses have processed the queries, they send the query results to the mediators. All mediators send the local query results to the federated layer to integrate the result for users. (Tseng, p.211)

By utilizing data stored in their original "home" data sources rather than copying the data to one massive data warehouse, searches always return the most up-to-date information. Also, searching several smaller data sources is generally more efficient than searching one big one.

Each data source will return the requested information in its native format and need to be converted to the querying organization's format in order to be integrated. The federated layer takes care of this by using the mediated schemas to convert all received information into a standard format as expected by the user.

This approach has its drawbacks in that it "suffers from the complexity of the mediators and the communication mechanism among the mediators. It may lead to heavy loading on each local data warehouse and the federated component in this framework. If users pose the same query at different times, the results must be recomputed or re-processed" (Tseng, p.211).

When Schema Mediation Doesn't Make Sense

There are times when schema mediation may not be the best option. If the schema is not published or not well documented, it may be impossible to translate between it and another schema. Similarly, if the target schema has "catch-all" fields where different kinds of data that don't fit anywhere else are dumped or the users who entered the data didn't follow the schema's rules, it is unlikely that any useful, structured data will be able to be gleaned. Also, if there is too little overlap between the two schemas, it may not be worthwhile to try to mediate between them.

Approaches to Schema Mediation

There are two ways to do schema mediation – manually, perhaps using tools or coding XSLT stylesheets by hand, or in an automated fashion where the computer makes suggestions as to what fields should be matched.

No matter what the mechanism, schema mediation consists of three main stages: "conflict analysis, conflict resolution, and schema merging. During conflict analysis, differences in the schemas are identified. In the second stage the conflicts are resolved. Finally, the schemas are merged into a single global schema using the decisions made during the previous stage" (Almarimi, p.2).

Two of the tools available for schema mapping using a GUI interface are Stylus Studio 2006 and Altova MapForce 2006. These tools work by presenting the user with the schemas that need to be integrated and allowing her to integrate as necessary. Fields and attributes can either be copied directly or the dozens of library functions can be used to somehow manipulate the data. Once all relationships have been mapped, an XSLT stylesheet is generated that produces the desired result.

While the tools are simple and straightforward to use, even the small mappings I was generating quickly overwhelmed my screen, and I was unable to really see all of what I was doing. It seems like it would be simpler and cleaner to write the XSLT manually. I could see mistakes being made because the visual of the mapped relationships became distorted by the lines all over the place.

Another popular technique for dealing with data integration is to use ontologies that describe the domain and allow for translation among XML files or between an XML file and the ontology. This method allows for a standardized way of looking at the data and ensures that mediation will be consistent. It also serves as a type of controlled vocabulary.

Again, this is the tactic used by Bio-Mediator. In this data integration system, "[o]ntologies play several important roles …: First, ontologies of genetics and molecular biology can serve as *data sources*. In this role concepts from the ontologies are returned as results of queries. Second, queries are posed against a *mediated schema*, which is an ontology describing the domain of discourse. User queries are expressed using the concepts in the mediated schema to indicate which results to retrieve. Third, each data source is an instance of the *system ontology*. This ontology describes information about the data sources including how often the source is updated and by whom. Finally, we are exploring the use of ontologies as a mechanism for *mapping* data sources to the mediated schema." (Mork, p.1).

Problems in Schema Integration

Integrating dissimilar schemas can present several issues. First among those is the enormous challenge of having a clear understanding of all of the schemas involved. One person's "Name" is another person's "Client," but without extensive documentation of the semantics of the schemas, this may not be clear.

A mediated schema can be quite brittle, and sometimes it "becomes a bottleneck in the process. Mediated schema design must be done carefully and globally; data sources cannot change significantly or they might violate the mappings to the mediated schema; concepts can only be added to the mediated schema by the central administrator" (Halvey, p.1). Further development of the schema and its accompanying business processes is limited by the fact that an organization has so much invested in the current mediated schema. The gain from changing the schema will need to outweigh the pain and expense of recreating the mediated schema and updating all previous documents. This ROI isn't always easy to calculate.

Finally, "In a typical data integration scenario, more than half of the effort (and sometimes up to 80 percent) is spent on creating the mappings, and the process is labor-intensive and errorprone" (Halevy, 2005, p.3). As discussed above, the tools available are manual and are incredibly expensive.

Automated Schema Mediation

The idea of a computer being able to automatically translate between schemas is enticing. But can it work? In their paper "A Survey of Approaches to Automatic Schema Matching," Rahm and Bernstein evaluate and categorize the various approaches to automatic schema matching. They call for more research into developing a generic application of what they call "Match" that works across domains, disciplines and applications.

Interestingly, many of their techniques are drawn from the field of library and information science. One of the more interesting approaches is what they call "name matching." Basically, this involves using controlled vocabularies to make the matches between schemas. The system looks up the schema elements in these controlled vocabularies and attempts to make a match suggestion based on what it finds. "In addition, name matching can use domain- or enterprise-specific dictionaries and is-a taxonomies containing common names, synonyms and descriptions of schema elements, abbreviations, etc." (Rahm, p.340) This obviously requires a substantial effort on the part of the organization, but may be more flexible than the manually generated mediated schemas approach. Natural language processing and information retrieval techniques are also proposed as possible approaches to the Match problem.

One area in which an automated data integration approach could have significant effect is the semantic web. Automation of matching schema elements could allow agents to quickly and easily understand a website's data structures and find the information they need. If the agent, for example, is looking for a used car and knows that car and automobile mean the same thing, it will be more successful at finding the necessary information. However, creating a controlled vocabulary of some sort for a single organization's needs is quite different from creating one for the entire web. Language, culture and worldview would make any efforts toward a universal thesaurus virtually impossible.

In order for automated matching, as it currently exists, to be useful, it needs to use a combination of complementary approaches. "[A] matcher that uses just one approach is unlikely to achieve as many good match candidates as one that combines several approaches" (Rahm, p.343). Using several of the techniques for schema matching is likely to yield a set of matching suggestions that will require the least amount of human intervention.

Alternatives to Schema Mediation

Schema mediation is an extremely expensive process that results in a brittle product. Are there other ways for organizations to accomplish the same goal or to reduce some of the costs?

First, organizations can create organization-wide global schemas that cover all of the core areas of their mission. For example, in any corporation, there are employees. The global company-wide schema can have a section for employee information, which includes all the relevant fields. Any schema that has employee information would duplicate this section, utilizing only the elements it needs for its information structure The company can go further and declare that anywhere in the company, if there is a person's name, it will always consist of <first_name> and <last_name> instead of <full_name>. Addresses will have each line in a separate element, with <city>, <state>, and <ZIP> as three elements rather than one.

In addition, trade groups can help create industry-wide schemas that have the basic industry data in an agreed-upon format. Organizations can agree with their suppliers to utilize the same format for common fields like quantity and cost. Simple steps such as these would alleviate a large portion of schema mediation issues and would reduce the number of data elements that need to be manually mapped. "While it is unlikely that the whole world agrees on such schemas, they can be specified for an enterprise, its trading partners, relevant standards bodies, or similar organizations to reduce the degree of variability" (Rahm, p.341). If these matches are automated, the remaining mappings will be less expensive.

Organizations could also recognize that controlled vocabularies can be a powerful tool in data integration. Creating a thesaurus or ontology for their data would not only ease the pain of data integration, it could potentially have an enormous impact on other areas of their businesses where data needs to be organized and retrieved. These same controlled vocabularies could be used to create useful indexes for the company intranet or used in a content management system, thus leveraging the work required to build the controlled vocabulary.

Finally, organizations could focus on reusing their existing mappings or using libraries of mappings. "Reuse-oriented approaches are promising, since we expect that many schemas need to be matched and that schemas often are very similar to each other and to previously matched schemas. … Schema editors should access these libraries to encourage the reuse of predefined schema fragments and defined terms, perhaps with a wizard that observes when a new schema definition is similar but not identical to one in a library" (Rahm, 341).

While schema mediation is a potential boon to organizations, the current methods for manual data integration are costly, tedious and error-prone. Available GUI-based tools do not seem to improve the process significantly. More automated methods of schema mediation hold the promise of making this process much less painful, though they still have a long way to go before their accuracy can be trusted. Perhaps the most useful things organizations can do are to change their internal processes of creating and storing data such that they use standardized global schemas as much as possible, both internally and within their networks, and to make every effort to re-use their existing mediated schemas.

References:

Almarimi A., Pokorný J. (March 2005). A mediation layer for heterogenous XML schemas. Presented at: iiWAS2004 Information Integration and Web Based Applications & Services, Jakarta, Indonesia. 27.9.-29.09.2004. In: *International Journal of Web Information Systems*, Vol. 1, No. 1, pp. 25-32.

Halevy, A.Y., Ives, Z. G., Suciu, D., Tatarinov, I. (2003). Schema mediation in peer data management systems. *In proceedings of ICDE.*

Halevy, A.Y. (2005). Why Your Data Don't Mix . *ACM Queue, 3(8).*

Moran, M., Mocan, A. (2005). Towards translating between XML and WSML based on mappings between XML Schema and an equivalent WSMO ontology. *2nd WSMO Implementation Workshop (WIW 2005).*

Mork, P., Shaker, R., Tarczy-Hornoch, P. (July 2005). The multiple roles of ontologies in the BioMediator Data Integration System. [*Proceedings of the Data Integration in the Life Sciences Workshop*] [PDF-paper]

Rahm, E., Bernstein, P. A. (December 2001). A survey of approaches to automatic schema matching. *The VLDB Journal The International Journal on Very Large Data Bases*, Volume 10, Issue 4, Pages 334 - 350, DOI 10.1007/s007780100057, URL http://dx.doi.org/10.1007/s007780100057

Tseng, F. S. C., Chen, C. (2005). Integrating heterogeneous data warehouses using XML technologies. *Journal of Information Science* 31: 209-229.