



XML and Data Integration

Elisa Bertino • University of Milan • bertino@dsi.unimi.it

Elena Ferrari • University of Insubria • Elena.Ferrari@uninsubria.it

XML is rapidly becoming a standard for data representation and exchange. It provides a common format for expressing both data structures and contents. As such, it can help in integrating structured, semistructured, and unstructured data over the Web. Still, it is well recognized that XML alone cannot provide a comprehensive solution to the articulated problem of data integration.¹⁻³ There are still several challenges to face.

XML Characteristics

XML is designed to provide a document markup language that is easier to learn and use than SGML⁴ and semantically richer than HTML. In XML (<http://www.w3.org/TR/REC-xml>), a tagged element identifies a content portion of a document or a complex data object. Tags are defined by applications that are thus able to convey the semantics of the tagged contents. Elements can be nested – that is, an element may consist of other elements and can have attributes associated with it. Attributes provide additional information on elements, thus increasing their semantics.

Additionally, a document type definition can be associated with an XML document. DTDs describe the structure of a set of similar documents and are thus instrumental in promoting standardization of application documents and data objects. To overcome the lack of DTDs for modeling complex data, the World Wide Web Consortium (W3C) has developed an XML schema specification that provides functionality above and beyond DTDs. XML schemas support data types and the modeling of relationships and constraints.

XML clearly separates contents from presentation: XML models only content, whereas cascading style sheets, XSL (Extensible Stylesheet Language, <http://www.w3.org/TR/2001/REC-xsl-20011015/>), or XHTML handle presentation separately.

Data Integration Requirements

To integrate heterogeneous data sources requires more than a tool for organizing data into a common syntax. Data integration is a complex activity that involves reconciliation at various levels.

- *Data models.* Data sources can greatly differ with respect to the structures they use to represent data (for example, tables, objects, files, and so on). Reconciliation of heterogeneous data models requires a common data model to map information coming from the various data sources.
- *Data schema.* Once we have agreed upon a common data model, the problem arises of reconciling different representations of the same entity or property. For example, two sources may use different names to represent the same concept (“price” and “cost”), or the same name to represent different concepts (“project” to denote both the project an employee is working on and a project for which an employee is the reviewer), or two different ways for conveying the same information (“date of birth” and “age”). Additionally, data sources may represent the same information using different data structures. For instance, consider two data sources that represent data according to the relational model, where both sources model the entity Employee but the first uses only one table to store employee information while the other spreads this information across more than one table. The need thus arises for tools to reconcile all these differences.
- *Data instances.* At the instance level, integration problems include determining if different “objects” coming from different sources represent the same real-world entity and selecting a source when contradictory information is found in different data sources (for instance, different birth dates for the same person).

How XML Can Help

Given these issues concerning data integration, it is important to understand whether and how XML can help the integration process and where XML is not enough.

Data Model Reconciliation

XML provides a quite natural way of structuring data, based on hierarchical, graph-based representations. Such models have the advantage of being simple, standard, and well accepted, but also powerful enough to represent structured, unstructured, and semistructured information. Thus, XML works well as a common data model over the Web, and a variety of tools have been marketed to support database contents publishing in XML. In addition, several research efforts have investigated the use of graph-based data models for managing semistructured data and for integrating heterogeneous data sources (see Widom³ for details on the Lorel Project).

Schema Reconciliation

The names and meanings of XML tags are arbitrary, which makes it critical to agree on a set of standardized domain-specific tags and schemas. There have been several efforts toward this end (for example, Biztalk [<http://www.biztalk.com/>] and Oasis [<http://www.oasis-open.org/>]). An alternative is to use the W3C's XSLT (Extensible Stylesheet Language Translation) for defining mappings among heterogeneous tags, but the mappings can be cumbersome when the number of tags is high.⁵

Standardization of tags and schemas would greatly simplify the task of data integration, but we also need a way to describe the semantics of elements and attributes. For instance, suppose that two data sources use the same element, "price," to describe the amount of money required to buy a specific item. Questions that still need to be answered include which currency specifies the price, whether the price includes taxes, shipping expenses, and so on. To answer such questions requires some sort of metadata that describes the semantics of

the information to be integrated. Metadata can be expressed using XML itself. For example, the XML fragment

```
<Price>12000
<Currency>ITL</Currency>
</Price>
```

can be used to attach currency information to price. More-complex semantic information can require alternative representations. The reconciliation process can also account for information about context, such as meanings of a particular name or a specific value that depend on the context in which the information occurs. For instance, an employee's numeric identifier can convey relevant information to readers that know the numbering conventions in place at the organization – that is, both the information and the subject share the same context. For the purpose of data integration, it is thus important to devise techniques and architectures that allow the interchange and integration of context information.

Instance Reconciliation

Metadata play a crucial role in instance reconciliation, determining how to deal with similar or contradictory information. For instance, metadata lets us attach a time stamp or quality level to the data being integrated, and such information can be used in solving conflicts among information stored at different sources.

What Is Still Needed

The problem of data integration is not new. It has been studied extensively long before XML came to the stage. Although XML can greatly help the task of data integration and reduce the work of reconciling heterogeneous data sources, it is not enough to address the articulated issue of data integration.

A key element of data integration is a language for specifying the semantics associated with data content. The W3C has recognized this need. Both the Resource Description Framework (RDF, [syntax\) and Semantic Web working groups are addressing issues related to the representation of semantic aspects of Web data. The Semantic Web working group's primary goal is to define architectures, models, and standards for providing a machine-readable description of the semantics of Web resources. The research is still in its infancy, and several issues require further investigation. The most relevant issues for data integration are](http://www.w3.org/TR/REC-rdf-</p>
</div>
<div data-bbox=)

- developing a formal foundation for Web metadata standards;
- developing techniques and tools for the creation, extraction, and storage of metadata;
- investigating the area of semantic interoperability frameworks; and
- developing semantic-based tools for knowledge discovery.

The development of suitable tools for XML-based integration of heterogeneous sources is also important. Those tools must support integration – automated as much as possible – at all three data levels: model, schema, and instances. □

References

1. A. Halevy, "More on Data Management for XML," white paper, available online at: <http://www.cs.washington.edu/homes/alon/widom-response.html>.
2. L. Seligman and A. Rosenthal, "XML's Impact on Databases and Data Sharing," *Computer*, vol. 34, no. 6, June 2001, pp. 59-67.
3. J. Widom, "Data Management for XML: Research Directions," *IEEE Data Eng.*, vol. 22, no. 3, Sept. 1999, pp. 44-52; available online at <http://www-db.stanford.edu/widom/xml-whitepaper.html>.
4. Standard Generalized Markup Language (SGML), ISO 8879, Int'l Organization for Standardization, available at <http://www.iso.ch/iso/en/ISOOnline.openerpage>.
5. C.H. Goh et al., "Context Interchange: New Features and Formalisms for the Intelligent Integration of Information," *ACM Trans. Office Information Systems*, July 1999.

Elisa Bertino is a full professor of database systems at the University of Milan, Italy.

Elena Ferrari is professor of computer science at the University of Insubria, Italy.