Chapter 15

XML Data Integration:
Merging, Query Processing and Conflict Resolution

Yan Qi
Arizona State University, Tempe, Arizona, USA

Huiping Cao
Arizona State University, Tempe, Arizona, USA

K. Selçuk Candan
Arizona State University, Tempe, Arizona, USA

Maria Luisa Sapino
University of Torino, Torino, Italy

ABSTRACT

In XML Data Integration, data/metadata merging and query processing are indispensable. Specifically, merging integrates multiple disparate (heterogeneous and autonomous) input data sources together for further usage, while query processing is one main reason why the data need to be integrated in the first place. Besides, when supported with appropriate user feedback techniques, queries can also provide contexts in which conflicts among the input sources can be interpreted and resolved. The flexibility of XML structure provides opportunities for alleviating some of the difficulties that other less flexible data types face in the presence of uncertainty; yet, this flexibility also introduces new challenges in merging multiple sources and query processing over integrated data. In this chapter, the authors discuss two alternative ways XML data/schema can be integrated: conflict-eliminating (where the result is cleaned from any conflicts that the different sources might have with each other) and conflict-preserving (where the resulting XML data or XML schema captures the alternative interpretations of the data). They also present techniques for query processing over integrated, possibly imprecise, XML data, and cover strategies that can be used for resolving underlying conflicts.

DOI: 10.4018/978-1-61520-727-5.ch015

INTRODUCTION

One of the primary motivations behind the development of eXtensible Markup Language (XML) is to create a framework that can support interoperability between businesses and other enterprises. In short time, the simplicity and flexibility of XML leads to many new applications, including peer-to-peer (P2P) applications (Koloniari & Pitoura, 2005; Pankowski, 2008), bioinformatics (Achard, Vaysseixm, & Barillot, 2001), and semantic web (Decker et al., 2000). As we have seen in Chapter titled "XML Data Integration: Schema Extraction and Mapping", the simple, flexible and self-describing data representation of XML provides unique opportunities to support data integration. On the other hand, these same properties, especially the flexibility of the structure of the data and the possibility for each data contributor and user to have their own schemas also introduce many new challenges in the integration process. Figure 1 provides an overview of the major steps underlying the XML data integration process:

- *Schema extraction*: A particular challenge introduced by XML is that not all XML data come with an associated schema. While this enables the use of XML as a flexible messaging and integration medium, when the integration process is schema-aware, it also necessitates a process to extract schemas that can be used during integration.
- *Matching and mapping*: Finding mappings between data components is a common problem in almost all integration domains. XML data can often be represented using trees or tree-like graphs (Goldman & Widom, 1997). This impacts solutions for finding mappings between XML data.
- *XML data/metadata merging*: Once the mappings are discovered, the next step in the process is to integrate the XML data or metadata, depending on whether the system is operating on data- or schema-level. This is often done through a transform-and-merge process.
- *Query processing and conflict resolution*: The results of the merge process, however, may not always be a valid XML data or schema. In these cases, in order to be able to use the resulting merged data in query processing, we either need to apply conflict resolution strategies or develop new query processing techniques that can operate on more relaxed data structures, such as graphs.

In fact, conflict resolution process can be integrated with query processing to support an incremental approach to cleaning the conflicts: as the user explores the integrated data (and conflicts) within the context of her queries, she can provide more informed conflict resolution feedback to the system.

In Chapter "XML Data Integration: Schema Extraction and Mapping" we have discussed the first two bullets in detail. In this chapter, we focus on merging and query processing over integrated XML data, and cover strategies that can be used for resolving conflicts with the user's help. The running example we use in this chapter is from the same domain (*universities and research institutes*) as Chapter "XML Data Integration: Schema Extraction and Mapping".

MERGING

Once the mappings between the sources are discovered through the matching process, the input sources can be merged into a logical "global" view for further use in integrated data processing. The merge process takes as input (a) a set of sources and (b) the mappings among them, and generates an integrated (target) data or schema.