XML Data Mediation and Collaboration: A Proposed Comprehensive Architecture and Query Requirements for Using XML to Mediate Heterogeneous Data Sources and Targets

Paul Benjamin Lowry
University of Arizona
Paul.Lowry@CMI.Arizona.EDU

Abstract
This paper discusses the potential of using XML technologies as mediators for accessing and processing structured, relational data and semi-structured data over the Internet.

Because significant gaps exist in all the reviewed XML data mediation approaches, a more comprehensive framework is presented. Some of the potential benefits of this architecture include: provides a clean abstraction of heterogeneous data for end users, applications, and Internet applications; supports all major data formats; uses one standard query language for all XML data queries; thus, optimizes and simplifying query writing; can be implemented in phases; consistent query language regardless of data source; returns all data results in XML format; provides different views of data according to different application needs; improves Internet spidering; continues to support all necessary DBMS functions; provides a platform for future growth; and encourages coherent and comprehensive data management and data modeling with organizations.

1. Introduction

XML is a markup language for structured and semi-structured data that was developed by the WWW Consortium (W3C) [1] and is supported by major industry rivals such as IBM, Sun, Oracle, and Microsoft [2]. The purpose of XML is to allow anyone to freely describe structured and unstructured data and their relationships [1]. Both HTML and XML are derivatives of the meta-language, SGML, which has supported similar markup concepts for over two decades [3]. However, XML is not based on HTML [3] XML is a simplified version of SGML without its overall complexity [4].

XML has several benefits and features that allow it to act as a mediator for accessing many forms of data over the Internet. First, XML is an open standard, which is platform independent, license free, vendor neutral, and has strong vendor support [1, 3, 5-7]. Second, XML’s syntax is very simple, yet strongly enforced [8]. Third, XML promotes automated data reuse [6, 7]. Fourth, XML uses meta-data, which allows it to be self-describing [1]. For further elaboration on how these benefits can facilitate electronic commerce and collaboration refer to [9].

2. Potential benefits of XML Mediation

XML is a strong candidate to mediate structured and unstructured data on the Internet; because XML’s data structure is flexible, it supports advanced views, and it supports intelligent agents and applications integration. Specific examples of XML’s benefits of acting as a data mediator between disparate data sources are discussed as follows.

XML’s data structure supports data mediation; because XML tags have no set meaning [3] XML completely separates the data from presentation [10] and XML allows internationalization and media independence [3]. Additionally, XML data structures are user-definable [10], which allows significant extensibility: data structures can be modeled to virtually any level of simplicity or complexity depending on the complexity of the data to be captured [3]. XML’s separation of data from presentation and its flexibility promotes the use of any level of complexity for device-dependent or independent views [3]. This promotes universal data access [11] from alternate user interface metaphors [7]. Furthermore, because XML is an open standard that is self-describing, it facilitates the use of intelligent agents for straightforward access to XML applications [12]. Thus, XML allows structured queries to extend search capabilities on Web pages [12]. As XML is used more throughout the Internet, XML will make the Internet more fully accessible to agents and other automated processes [13]. Finally, XML also