# Managing Web Data

Dan Suciu
AT&T Labs — Research
suciu@research.att.com

## 1 Introduction

The Web today consists exclusively of HTML documents designed for the human eye. While many of them are generated automatically by applications, it is difficult for other applications to read and process them. This may soon change, due to a series of new standards from the World Wide Web Consortium centered around XML (Extensible Markup Language). XML is designed to express the document *content*, while HTML expresses its *presentation*. In short, XML is a data exchange format, easily understood by applications. It enables data exchange on the Web, both intra-enterprise, across platforms (intranet), and inter-enterprise (internet). The focus of the Web shifts from document management to data management, and topics like queries, views, data warehouses, mediators, which were the domain of databases, become of interest to the Web. However, the new data on the Web differs from traditional relational or object-oriented data: it is schema-less, self-describing, irregular, and heterogeneous. Recent database research has considered such data and called it *semistructured data*.

## 2 Overview of the Tutorial

This tutorial is an introduction to the field of semistructured data, presenting some of the recent work done in the database research community. It also contains a brief presentation of the work done at the W3C on XML and related topics. The tutorial compares and contrasts the two developments, highlighting similarities and differences. Specifically, the tutorial addresses the following three topics.

**Data Models** The semistructured data model is that of a labeled graph. The best known instance is the Object Exchange Model (OEM), a light-weight, self-describing data model. What is specific about semistructured data is that schema components are now part of the data (hence: *self-describing*), a characteristic shared by data expressed in XML. However, XML has roots in document markup and some of its features diverge from the semistructured data model.

**Query Languages** Much of the research effort in semistructured data has focused on query languages, both for data extraction and data transformation. Transformations are especially important in the context of XML, since XML data is extracted and/or integrated from other sources. The tutorial presents four powerful features found in semistructured data query languages: regular path expressions, patterns, constructors (both simple, and complex: Skolem Functions), and structural recursion, and illustrates them in the context of XML-QL, a query language for XML. In a separate development, the W3C is close to standardizing XSL, a language for XML stylesheets which can express certain transformations of XML data.

**Schemas** The work on schemas in the semistructured data community has been driven by two principles: the schema is separated and independent from the data (at an extreme it is extracted from the data), and it is used to improve efficiency (in query optimization, storage, etc). The tutorial presents schema graphs, unary datalog types, and data guides, explaining their relationships. The W3C community considered several schema formalisms for XML (DTD, RDF-schema, DCD): these are hard-wired with the data, and focus on providing semantics to the user.

## 3 Tutorial Notes

http://www.research.att.com/~suciu/tutorial-sigmod99.ps