University of Washington
Department of Computer Science & Engineering

Managing XML and Semistructured Data

Lecture 1: Preliminaries and Overview

Prof. Dan Suciu

Spring 2001

In this lecture

- Goals of the course
- Prerequisites
- Resources
  - textbooks
  - research papers
- Overview of the course

Goals of the Course

Purpose:

- Foundations of semistructured data
- Issues in semistructured data management
- Glimpse at current XML standards and technology

Prerequisites

- A graduate course in database systems
- Logic
- Programming languages
- Complexity theory
- Algorithms and data structures

Textbooks

- **Data on the Web: from Relations, to Semistructured Data and XML**,
  Abiteboul, Buneman, Suciu
  - For foundations
- **W3C homepage**, www.w3.org
  - For current standards
- **Professional XML Databases**,
  Kevin Williams
  - For current XML technologies

Other Useful Texts

- **A first course in database systems** (2 vols)
  Ullman, Widom and Garcia-Molina
- **Data and Knowledge based Systems** (2 vols)
  Ullman
- **Foundations of data bases**
  Abiteboul, Hull Vianu
- Proceedings of SIGMOD, VLDB, PODS conferences.

Papers: Data Models

- XML, Java, and the future of the Web by Jon Bosak, Sun Microsystems.
- W3C XML Query Data Model Mary Fernandez, Jonathan Robie.
- Adding structure to semistructured data by Buneman, Davidson, Fernandez, Suciu, in ICDT 97
- Object Exchange Across Heterogeneous Information Sources Y. Papakonstantinou and H. Garcia-Molina and J. Widom, Data Engineering 95

Papers: Query Languages

- A formal semantics of patterns in XSLT by Phil Wadler.
- XQuery: A Query Language for XML Chamberlin, Florescu, et al.
- XML-QL: A Query Language for XML by Deutsch, Fernandez, Florescu, Levy, Suciu, in WWW8.
- Catching the boat with Strudel VLDBJ 2001.
- UnQL: A Query Language and Algebra for Semistructured Data Based on Structural Recursion Buneman, Fernandez, Suciu.VLDBJ 2000
- The Lorel Query Language for Semistructured Data by Abiteboul, Quass, McHugh, Widom, Wiener, in International Journal on Digital Libraries, 1997.

Papers: Schemas

- MSL: A Model for W3C XML Schema by Brown, Fuchs, Robie, Wadler, in WWW10, 2001.
- Keys for XML by Buneman, Davidson, Fan, Hara, Tan, in WWW10, 2001.
- Subsumption for XML Types by Kuper and Simeon, ICDT'2001.
- Extracting Schema from Semistructured Data Nestorov, Abiteboul, Motwani. SIGMOD 98

Papers: Query Analysis, Typechecking

- Optimizing Regular Path Expressions Using Graph Schemas Fernandez, Suciu, ICDE'98.
- XDuce: A typed XML processing language by Hosoya and Pierce
- Regular Expresssion Pattern Matching for XML by Hosoya and Pierce (in POPL 2001)
- Typechecking for XML TransformersMilo, Vianu, Suciu.

Papers: Indexing

- Index Structures for Path Expressions by Milo and Suciu, in ICDT'99.

Papers: Publishing

- Efficiently Publishing Relational Data as XML Ducments by Shanmugasundaram, Shekita, Barr, Carey, Lindsay, Pirahesh, Reinwald in VLDB'2000
- SilkRoute: Trading between relations and XML by Fernandez, Suciu, Tan R, in WWW9, 2000
- Efficient Evaluation of XML Middle-ware Queries in SIGMOD'2001

Papers: Compression

- XMILL: An Efficient Compressor for XML Data by Liefke and Suciu, in SIGMOD'2001

Overview

- Semistructured Data
  - Model
  - Syntax
  - Comparison with relational data

Overview

- XML
  - Motivation
  - Syntax:
    - Basic stuff: elements, attributes, content
    - Esoteric stuff: PIs, entities, CDATA, comments
  - DTDs
  - Data model (XQuery)
  - Miscellaneous: Name spaces, XPointer, XLink

Overview

- Query Languages
  - Lorel extends OQL
  - UnQL structural recursion, patterns
  - StruQL Skolem Functions
  - XML-QL everything for XML
  - Quilt/Xquery the standard
  - XSL the standard
  - XDuce a general-purpose language

Overview

- Schemas
  - Theory: lower bound, upper bound
  - XML-Schema
  - "XML-Schema are regular tree languages"
  - Constraints (keys for XML)

Overview

- Query analysis
  - Query pruning
  - Query containment

Overview

- XML Publishing from Relational Databases
  - Virtual XML publishing: SilkRoute, Microsoft's XDR
  - Materialized XML publishing: Experanto, SilkRoute, Microsoft's "for XML"

Overview

- Indexes
  - Indexes for ss data: data guides, T-indexes
  - Indexes for XML: we are still waiting for them...

Overview

- Miscellaneous
  - XML compression (Xmill)