Requirements for XML Document Database Systems

Airi Salminen
University of Jyväskylä
http://www.cs.jyu.fi/~airi/

Frank Wm. Tompa
University of Waterloo
http://db.uwaterloo.ca/~fwtompa/

Outline

1. Motivation
2. The Data Model
3. Data Definition
4. Data Manipulation
5. Conclusion

- Quantities of persistent XML data increasing

- Traditional traditional database systems as such are not suitable

- Traditional SGML systems as such are not suitable

- There is a mixed, heterogeneous collection of implemented systems

  Need for better understanding of the requirements for XML database systems

**Need for understanding
the requirements for**

The data model

Data definition capabilities

Data manipulation capabilities

Needed

1. A universal conceptual model

2. Capability to model document collections as well as enterprises (described in the documents)

3. Well-defined equivalence

Four models for XML data from W3C

XML Information Set

XPath 1.0

DOM 1.0

XQuery 1.0 and XPath 2.0

XML Infoset	XPath 1.0	DOM 1.0 Level 2	XQuery 1.0 & XPath 2.0
Document	Root	Document	Document
Element	Element	Element	Element
Attribute	Attribute	Attr	Attribute
PI	PI	PI	PI
Comment	Comment	Comment	Comment
Character	Text	Text	Text
Namespace	Namespace		Namespace
Document Type Decl.		DocumentType	
Unparsed Entity		Entity	
Unexpanded Ent. Ref.		EntityReference	
Notation		Notation	
		CDATASection	Reference
		DocFragment	
11	**7**	**12**	**8**

Information omitted from XML Infoset

- The content models of elements
- The grouping and ordering of attribute declarations
- The document type name
- White space outside the document element
- Whether characters are represented by character references
- The difference between the two forms of an empty element
- The difference between CR, CR-LF, and LF line termination

Information omitted from XML Infoset

- The order of attributes within a start-tag
- The order of declarations within the DTD
- Comments in the DTD
- The location of declarations (whether in internal or external subset or parameter entities)
- The boundaries of general parsed entities
- The boundaries of CDATA marked sections
- The default value of attributes declared in the DTD
  ...

Equivalence

"XQuery 1.0 and XPath 2.0 Functions and Operators" document will define equality between values and equality between nodes

- all data is not in the model
- semantics of value equality is still unclear

Canonical XML defines logical equivalence of two documents within a given application context

- based on the XPath 1.0 data model, all data is not in the model
- does not define equivalence within different application contexts

Needed capabilities to define

1. Data types
2. Document types
3. Data collections
4. Document type collections
5. Multiple levels of validity
6. Entities and URIs
7. Namespaces
8. Document indexing
9. User roles and access rights

Needed capabilities for

1. Queries

2. Transformations

3. Document assembly

4. Update

Queries in terms of

all data in documents

document collections

schemas

other metadata

Extensive lists of requirements in Maier (1998) and in the "XML Query Requirements"

Transformations for

rendering

integration support

schema evolution

views

Document assembly

Support for the specification of the assembly process

Update

Insertion, deletion, replacement of

whole documents
parts of documents
files
URIs
style sheets
schemas
and other information units

Referential integrity

Requirements for the data model, data definition, and data manipulation were discussed

Top priorities

A complete data model

DDL and DML to include all components of the model

Mechanisms to manage collections of schemas along with documents

Thank you!