Foundations of XML Data Manipulation

Giorgio Ghelli

Course structure

- Data Model
- Query languages
- XPath
- Type systems, logics, tree automata
- Storing and querying

Structured data

<table>
<thead>
<tr>
<th>ID</th>
<th>Last Name</th>
<th>First Name</th>
<th>Title</th>
<th>Birth Date</th>
<th>Hire Date</th>
<th>City</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Order ID</th>
<th>Customer</th>
<th>Emp ID</th>
<th>Order Date</th>
<th>Required Date</th>
<th>Shipped Date</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Unstructured data

- Sample databases included with Access
  - Microsoft Access provides sample databases that you can use while you’re learning Access.
  - Northwind Traders sample database
  - The Northwind database and Access project (available from the Sample Databases command on the Help menu) contains the sales data for a fictitious company called Northwind Traders, which imports and exports specialty foods from around the world. By viewing the database objects included in the Northwind database. . .

Semistructured data

A syntax for SSD

expr :: value | oid value | oid
definition ::= atomic | { label : expr, . . . , label : expr }

{ Bib : &o1 ( paper : &o12 . . . ),
  book : &o24 ( . . ),
  paper : &o29
  { author : &o52 "Ableboul",
    author : &o96 { firstname : &o243 "Victor",
      lastname : &o206 "Vianu"},
    title : &o93 "Regular path queries",
    references : &o24,
    page : &o25 { first : &o64 122, last : . . . } }
}
Why SSD

- The origin:
  - Data integration
  - Documents
  - Scientific databases
- The interest:
  - Cannot be ignored
  - WWW and bioinformatics

The Data Model

- The information behind the syntax, i.e.: when two pieces of data really differ
- Some alternatives:
  - OEM: SSD as graphs modulo bisimulation
  - XML: ordered trees with node identity (and with pointers)
  - TQL: unordered trees

OEM with bisimulation

- Edge-labelled version
- Bisimulation: generalizes the notion of set equality to labelled graphs:
  - \( \{a: v, b: w\} = \{b: w, a: v\} \)
  - \( \{a: v, a: v, b: w\} = \{a: v, b: w, b: w\} \)
- Exists \( R \subseteq G \times G' \) such that:
  - \( n R m \text{ and } n, l, l' \text{ in } G \Rightarrow \text{exists } m, l, l' \text{ in } G' \text{ with } n' R m' \text{ and conversely} \)
  - \( n R m \text{ and } n \text{ leaf in } G \Leftrightarrow m \text{ leaf in } G' \)

TQL data model

- Edge-labeled trees defined as multisets of label-tree pairs:
  - \( f ::= \emptyset \ | \ a[f] \ | \ f \uplus f \)
  - \( t ::= \emptyset \ | \ \{a: t\} \ | \ t \uplus t \)
- Hence:
  - \( \{a: v, b: w\} = \{b: w, a: v\} \)
  - \( \{a: v, a: v, b: w\} = \{a: v, b: w, b: w\} \)
- The same syntax can be interpreted as node-labeled forests

Ordered children (as in XDM)

- Node-labeled ordered trees of elements
  - \( \text{item ::= <label> value <label> | leaf} \)
  - \( \text{value ::= item *} \)
- Hence:
  - \( \{a: v, b: w\} \neq \{b: w, a: v\} \)
  - \( \{a: v, a: v, b: w\} \neq \{a: v, b: w, b: w\} \)

Binary representation

- Diagram of binary tree representation and nodes.
XML

- Simplification of SGML
- Designed to substitute HTML
- The standard for data exchange and web-services invocation
- Some W3C related standards:
  - XPath/XQuery
  - XML Infoset and XDM
  - XSLT
  - DTD, XSD
- Many other things

XML for data exchange

```xml
<doc>
  <title>Sample databases included with Access</title>
  <subtitle>Microsoft Access provides sample databases.</subtitle>
  <subtitle>Northwind Traders database</subtitle>
  <body>
    <para author="JDM" font="times">The Northwind database contains the sales data for a company called Northwind Traders, which imports and exports specialty foods from around the world. By viewing the database objects included in the Northwind database.</para>
  </body>
</doc>
```

XML as it was designed

- A value is a sequence of nodes
- Parent axis: a node is a pair <tree, path in the tree>
  - {a: {b: w}}/b /= {b: w}
- Node identity: a store is a forest-structured graph <N,E>, and a node is an element of N
  - {a: v} /= {a: v}

Moreover

- Six other types of nodes
- Unordered attributes
- ID – IDREFs to encode pointers
- Namespaces
- Type annotations
Conclusions

- We now know what SSD is
- Questions:
  - How do we describe its structure?
  - How do we manipulate it?
  - How do we store it?

Suggested readings

- Some papers are in the “query folder”
- klarkundSchweintick: general introduction to XML, DTD, XSD, XPath, XQuery, XSLT.
- AbiQuamMo97: OEM and Lorel
- BunDavHill96: UnQL data model
- CoiGheAl06: MicroXQuery data model
- Hiddens: XML data model
- CarGhe03: TQL data model
- [www.w3.org/TR/xpath-datamodel](http://www.w3.org/TR/xpath-datamodel): XQuery/XPath data model (XDM)