Database support for XML

Tore Risch
Information Technology
Uppsala University

What is XML?

- Originally HTML subset of SGML
- HTML *text markup* language
- XML larger subset than HTML
- HTML has predefined markup tags,
  e.g. <a…> …
- XML allows programmer to define and use *user defined* tags
- As in HTML annotation of document elements also allowed:

  User interface

DTD

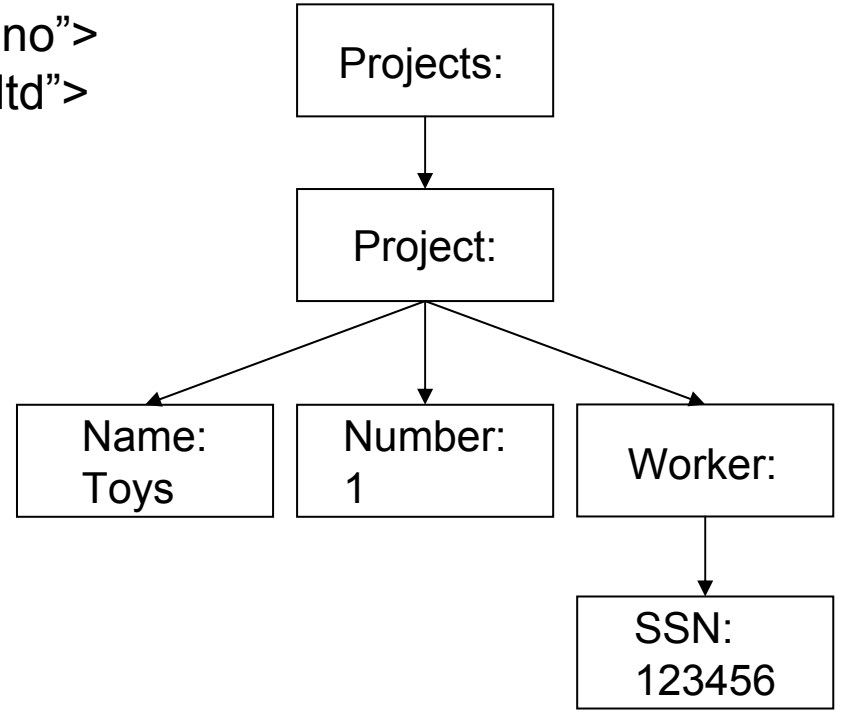
- *Grammar* of XML documents, e.g. allowed tags, can be described (constrained) by *DTD-documents*
- Can be seen as simple *schema*
- Original purpose still document markup
- XML documents don’t require DTDs

XML for data exchange

- There is a lot of need to *exchange* data between systems, e.g. by e-mail or between programs
- DTDs allow to define standard formats (schema) for data to exchange
- E.g. record structures, lists, etc.
- DTDs defined e.g. for various business exchanges
- Still clumsy format

Example XML data

```
<?xml version “1.0” standalone=“no”>
<!doctype projects system “proj.dtd”>
<projects>
<project>
  <name>Toys</name>
  <number>1</number>
  <worker>
     <ssn>123456</ssn>
     ….
   </worker>
   …
 </project>
 <project>
…
</projects>
```



Tree structure!

(projects (project (name “Toys”)(number 1)(worker (SSN 123456))))

Example DTD

```
<!doctype projects [
   <!element projects (project+)>
   <!element project (name, number, workers)
   <!element workers (worker+)>
   <!element worker (ssn)>
]>
```

Schema for XML documents.

Unlike RDBs *sequences* not sets.

'*any*' might be specified for subtree ->no constraints

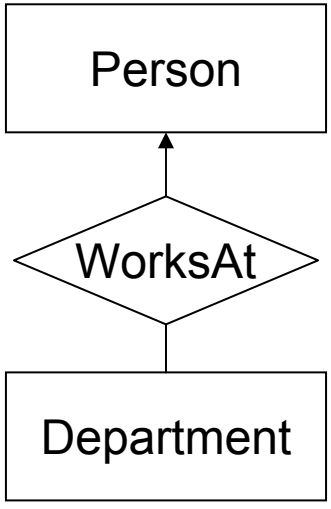
Semistructured data!

Problems with XML for data exchange

- DTDs are voluntary and need not be followed
- Only datatype is *string*
- XML still very common and standard schemas for various application areas have been defined

XML-Schema

- XML-Schema is extension of DTDs
- XML syntax for *schema* i.e. *type definitions* too
- Very rich set of built-in data types
- User defined data types
- Type system for nested record structures
- Can be seen as ‘object-oriented’ schema
- *Oracle XML DB* can translate XML-Schema -> OR schema for subsequent querying




```
<?xml version="1.0" ?>
<xsd:schema xmlns:xsd="http:www.w3.org/2001/XMLSchema">
<xsd:element name="Person"> <xsd:complexType><xsd:sequence>
  <xsd:element name="SSN" type="xsd:integer">
  <xsd:element name="WorksAt" type="Department" maxOccurs="1">
</xsd:complextype></xsd:sequence></xsd:element>
<xsd:element name="Department"> <xsd:complexType><xsd:sequence>
  <xsd:element name="Dname" type="xsd:string">
  <xsd:element name="Workers" type="Person" maxOccurs="unbounded">
</xsd:complextype></xsd:sequence></xsd:element>
```

XML query languages

- *XPath*, based on regular path expression selecting XML substructures of tree structure:

  doc(user.it.uu.se/~torer/doc)/Persons//Worker/SSN

- *XQuery* also allows *joins* and constructing new *document views*, FLWR expressions:

  for $x in doc(user.it.uu.se/~torer/doc)//Person
  where $x/Name eq Toys
  return <ssn> $x/Worker/SSN </ssn>

- Xquery calls XPath as sublanguage
- SQL calls Xquery as sublanguage

Commercial systems

- Commercial systems
  Oracle XML DB
  Microsoft SQL Server 2000 SQLXML
  IBM DB2 XML Extender
- On top of RDB
  XMLType in Oracle
  XML2CLOB in DB2
- Different query models
  Oracle: Translate XMLSchema to tables
  IBM: Embed XPtath expressions as UDFs in SQL

Research issues

- Data storage for XML
  - Tables structured, XML *semistructured*
  - Separate XML DBMS or on top of SQL?
  - Compact representation of XML trees
- How to query XML?
  - Syntax and semantics of query language
  - Efficient processing of XML queries
  - XML algebra
  - Translation to SQL?
- Research area was very hot in early 2000s
  - very many papers

Some papers

Fernandez, et al: SilkRoute: A Framework for Publishing Relational Data in XML. ACM TODS, 27(4), Dec 2002.

Boncz et al: MonetDB/XQuery: A Fast XQuery Processor Powered by a Relational Engine, SIGMOD 2006

Cheng, et al: Twig©÷Stack: Bottom-up Processing of Generalized-Tree-Pattern Queries over XML Documents, VLDB 2006

Halverson, Josifovski, Lohman, Pirahesh, Mörschel: ROX: Relational Over XML, VLDB 2004

Arion et al: Pushing Queries to Compressed XML Data., VLDB 2003 (demo)