XML and Databases

Spring Semester 2017

Date: Thursday, 08:15-9:45, BIN-2.A.10

Lecturer: Dr. Can Türker (Functional Genomics Center Zurich)

Email: tuerker@fgcz.ethz.ch

WWW: http://www.ifi.uzh.ch/dbtg/teaching/courses/xmlanddatabases.html
Google Indicator (February 13, 2017)

<table>
<thead>
<tr>
<th>Database</th>
<th>940'000'000</th>
<th>W3C</th>
<th>60'000'000</th>
</tr>
</thead>
<tbody>
<tr>
<td>XML</td>
<td>427'000'000</td>
<td>ETH</td>
<td>38'500'000</td>
</tr>
<tr>
<td>FIFA</td>
<td>353'000'000</td>
<td>UBS</td>
<td>32'900'000</td>
</tr>
<tr>
<td>SQL</td>
<td>224'000'000</td>
<td>DB2</td>
<td>19'500'000</td>
</tr>
<tr>
<td>UEFA</td>
<td>160'000'000</td>
<td>FGCZ</td>
<td>42'500</td>
</tr>
<tr>
<td>Zürich</td>
<td>98'900'000</td>
<td>Can Türker</td>
<td>23'700</td>
</tr>
</tbody>
</table>

⇒ XML and Databases: Both terms seem to be highly relevant!

In this lecture you will learn more about XML and its interplay with databases!
Lecture Summary

- **Description:** Today, the W3C standard XML is widely used as document format for exchanging data over the Internet. While the generation of XML data is easy, the management of XML data requires systems that are able to efficiently store, query, and process XML data. With other words, database technology is required for handling XML data. The goal of this lecture is to teach the interplay between XML and databases. The following aspects are studied in detail: semi-structured data model of XML, query languages (XPath, XQuery) for declarative access to XML data, XML processor technologies, mapping between XML and databases including efficient storage and index structures for XML data. A further central concern of this lecture is to show the practical relevance of all presented concepts by demonstrating how they are realized in major database management systems such as Oracle, IBM DB2, Microsoft SQL Server, and PostgreSQL.

- **Goal:** Achieve deep understanding of XML and its interplay with database technology

- **Prerequisite:** Databases (Bachelor level), i.e., basic knowledge in databases

- **Recommended:** MSc Studies

- **Documents:** Before each lesson, presentations and demo scripts are provided for download via the lecture’s home page [http://www.ifi.uzh.ch/dbtg/teaching/courses/xmlanddatabases.html](http://www.ifi.uzh.ch/dbtg/teaching/courses/xmlanddatabases.html)
Topics

1. XML and Databases
2. XML Language
3. XML Processors
4. XML Query Languages
5/6. Mapping between XML and Database
7. SQL/XML
8. XML Indexing
9. XML Update
10. XML Support in Database Systems

Diagram:
- XML Application
- XML Document
- XML Databases
- Access
- Process
- Output
- Create
- Transformation
- Generate
- Query & Search
- Storage
- Update
- Presentation
<table>
<thead>
<tr>
<th>Week</th>
<th>Date</th>
<th>Chapter</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>23.02.2017</td>
<td>Introduction and Motivation</td>
</tr>
<tr>
<td>9</td>
<td>02.03.2017</td>
<td>XML &amp; XML-DTD</td>
</tr>
<tr>
<td>10</td>
<td>09.03.2017</td>
<td>XML Schema &amp; XML Processors</td>
</tr>
<tr>
<td>11</td>
<td>16.03.2017</td>
<td>XML Processors</td>
</tr>
<tr>
<td>12</td>
<td>23.03.2017</td>
<td>XML Query Languages</td>
</tr>
<tr>
<td>13</td>
<td>30.03.2017</td>
<td>XML Query Languages</td>
</tr>
<tr>
<td>14</td>
<td>06.04.2017</td>
<td>XML Query Languages</td>
</tr>
<tr>
<td>15</td>
<td>13.04.2017</td>
<td>Mapping between XML and Databases</td>
</tr>
<tr>
<td>16</td>
<td>20.04.2017</td>
<td>No lecture!</td>
</tr>
<tr>
<td>17</td>
<td>27.04.2017</td>
<td>Mapping between XML and Databases</td>
</tr>
<tr>
<td>18</td>
<td>04.05.2017</td>
<td>SQL/XML</td>
</tr>
<tr>
<td>19</td>
<td>11.05.2017</td>
<td>SQL/XML</td>
</tr>
<tr>
<td>20</td>
<td>18.05.2017</td>
<td>XML Indexing &amp; XML Update</td>
</tr>
<tr>
<td>21</td>
<td>25.05.2017</td>
<td>No lecture!</td>
</tr>
<tr>
<td>22</td>
<td>01.06.2017</td>
<td>XML Support in Databases Systems</td>
</tr>
<tr>
<td>24</td>
<td>15.06.2017</td>
<td>Examination</td>
</tr>
</tbody>
</table>
Literature

- M. Klettke, H. Meyer: XML & Datenbanken. dpunkt Verlag, 2002
- S. Abiteboul, P. Buneman, D. Suciu: Data on the Web: From Relations to Semistructured Data and XML. Morgan Kaufmann, 1999
Chapter 1

Introduction and Motivation

Motivation of XML via
1) HTML
2) Databases
3) Information Retrieval
HTML vs. XML

```html
<html>
<head>
<title>The Tragedy of Hamlet, Prince of Denmark</title>
</head>
<body>
<h1>The Tragedy of Hamlet, Prince of Denmark</h1>
<h2>Dramatis Personae</h2>
<p>CLAUDIUS, king of Denmark.</p>
<p>HAMLET, son to the late, and nephew to the present king.</p>
<h2>Scene Description</h2>
<p>SCENE Denmark.</p>
<h2>Subtitle</h2>
<p>HAMLET</p>
<h2>ACT I</h2>
<h3>SCENE I. Elsinore. A platform before the castle.</h3>
<p>FRANCISCO at his post. Enter to him BERNARDO</p>
<p>BERNARDO: Who's there?</p>

```xml
<?xml version="1.0" encoding="ISO-8859-1"?>
<play>
<title>The Tragedy of Hamlet, Prince of Denmark</title>
<personae>
<title>Dramatis Personae</title>
<persona>CLAUDIUS, king of Denmark.</persona>
<persona>HAMLET, son to the late, and nephew to the present king.</persona>
</personae>
<scndescr>SCENE Denmark.</scndescr>
<playsubt>HAMLET</playsubt>
<act>
<title>ACT I</title>
<scene>
<title>SCENE I. Elsinore. A platform before the castle.</title>
<stagedir>FRANCISCO at his post. Enter to him BERNARDO</stagedir>
<speech><speaker>BERNARDO</speaker>
<line>Who's there?</line></speech>

...
Motivation via HTML

- **Web provides information embedded in HTML documents**
  - **Positive:**
    - Platform-independent, producer-independent
    - Always accessible from anywhere
    - Easy access/download via HTTP protocol
    - Applications can easily generate HTML documents
  - **Negative:**
    - HTML document processing is cumbersome within applications since HTML primarily describes layout but not structure of documents
    - HTML provides only a fixed set of structuring elements, which can neither be updated nor extended by users
    - Only simple, text-based queries supported

- **XML tackles the disadvantages of HTML without giving up its advantages**
  - Data transfer is based on Internet protocols such as HTTP
  - Structuring elements can be user-defined
  - Structuring elements determine not only layout but also structure
  - User-defined structuring elements support arbitrary complex queries
XML as Web Standard

- **XML documents**
  - Generated by applications
  - Consumed by applications
  - Used for data exchange

- **Data description and exchange as main scenarios for XML usage**
  - Weakly structured data such as Shakepears dramas available as text documents
  - Strongly structured data such customer/order data coming from a relational database

- **Power of XML goes far beyond HTML**
  - XML vs. classic data models vs. information retrieval models: How they relate?
  - From semantic to semi-structured data models: A schema capable of describing all objects?
Motivation via Databases and Semantic Data Models

- **Semantic data models distinguish between**
  - Schema and instance
  - Type and variable

  ![Diagram showing data types](image)

- **Object-oriented databases focus on atomic, abstract data types**
- **Relational databases build upon constructed data types**
- **Schema definition implicitly determines elementary integrity constraints**
Simple Data Modeling Example (1)

In our mini world, called PC, we distinguish persons and cars. Cars are described by their model, fabrication year, and owner. Persons are characterized by their name and address. Every car belongs to exactly one person while a person can own several cars.
Suppose there are following instances representing two persons and a car

The relational representation would looks as follows:

<table>
<thead>
<tr>
<th>Person</th>
<th>Id</th>
<th>Name</th>
<th>Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>Hugo</td>
<td>Zürich</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>Franz</td>
<td>Basel</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Car</th>
<th>Id</th>
<th>Model</th>
<th>Year</th>
<th>Owner</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
<td>VW</td>
<td>2013</td>
<td>2</td>
</tr>
</tbody>
</table>
Self-Describing Data

• Refers to data with integrated meta-data which annotates (describes) the data

• Remember: databases strictly separate between meta-data (schema) and data (instance)
  – Advantage: logical independence as basis for optimization
  – Disadvantage: instance without schema complicates data exchange

• Data should be self-describing! But how?

• Idea: Introduce a „semi-structured“ data model
  – Use a descriptor-oriented database schema
  – Uniform for all applications ⇒ hence often named as „meta schema“
Simple Descriptor-Oriented Data Model

- Each object has an id and a set $C$ of descriptors $D$ describing the characteristics of the object.
- Each descriptor $D$ has a name $N$, a type $T$, and a value $V$.

Fixed schema ("meta schema"):
- fixed names $Id$, $C$, $D$, $N$, $T$, $V$
- application-independent

- Each object can be described by its characteristics.
- Name and type information are part of the instance: Self-Description!
Simple Descriptor-Oriented Data Model: Relational Representation

<table>
<thead>
<tr>
<th>Entity</th>
<th>Id</th>
<th>Name</th>
<th>Type</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>1</td>
<td>Name</td>
<td>String</td>
<td>Hugo</td>
</tr>
<tr>
<td>Person</td>
<td>1</td>
<td>Address</td>
<td>String</td>
<td>Zürich</td>
</tr>
<tr>
<td>Person</td>
<td>2</td>
<td>Name</td>
<td>String</td>
<td>Franz</td>
</tr>
<tr>
<td>Person</td>
<td>2</td>
<td>Address</td>
<td>String</td>
<td>Basel</td>
</tr>
<tr>
<td>Car</td>
<td>3</td>
<td>Model</td>
<td>String</td>
<td>VW</td>
</tr>
<tr>
<td>Car</td>
<td>3</td>
<td>Year</td>
<td>Integer</td>
<td>2013</td>
</tr>
<tr>
<td>Car</td>
<td>3</td>
<td>Owner</td>
<td>Integer</td>
<td>2</td>
</tr>
</tbody>
</table>

Add further descriptor for each entity to maintain the class information.
Simple Descriptor-Oriented Data Model: Advantages and Limitations

+ Each object described by variable set of characteristics (descriptors)
  - No null values
  - No forced introduction of sub types when more characteristics are known
  - Each individual can be described individually

- Object names (classification) disappear: must explicitly be defined as additional descriptor
  - Which type of an object is the object with the id=1? Person or Car?
  - No comprising context (class)
XML Representation of Relational Data: Variant 1

<Person>
  <Row>
    <Id>1</Id>
    <Name>Hugo</Name>
    <Address>Zürich</Address>
  </Row>
  <Row>
    <Id>2</Id>
    <Name>Franz</Name>
    <Address>Basel</Address>
  </Row>
</Person>

<Car>
  <Row>
    <Id>3</Id>
    <Model>VW</Model>
    <Year>2013</Year>
    <Owner>2</Owner>
  </Row>
</Car>

</PC>
XML Representation of Relational Data: Variant 2

Person

Id Name Address
1 Hugo Zürich

Person

Id Name Address
1 Hugo Zürich

Car

Id Model Year Owner
3 VW 2013 2

<Person>
  <Id>1</Id> <Name>Hugo</Name> <Address>Zürich</Address>
</Person>

<Person>
  <Id>2</Id> <Name>Franz</Name> <Address>Basel</Address>
</Person>

<Car>
  <Id>3</Id> <Model>VW</Model> <Year>2013</Year> <Owner>2</Owner>
</Car>

</PC>
XML Representation of Relational Data: Variant 3

```
<PC>
  <Row> <Person> <Id>1</Id> <Name>Hugo</Name> <Address>Zürich</Address> </Person></Row>
  <Row> <Person> <Id>2</Id> <Name>Franz</Name> <Address>Basel</Address> </Person></Row>
  <Row> <Car> <Id>3</Id> <Model>VW</Model> <Year>2013</Year> <Owner>2</Owner> </Car></Row>
</PC>
```
Semi-Structured Data Models

... evolve by extending the simple descriptor-oriented model by a union data type:

Union operator $\cup$ introduced as additional type constructor!

In the context of XML, often also named sequence (ordered set)
An object O is described by set C of descriptors D which consist of a name N and either
- another object O or
- a type-value pair A with the type T and the value V

- Flexibility introduced by union type!
- Semi-structured data model
Tree Representation with Edge and Node IDs

Person

Cars

Address

Name

Id

3

Hugo

Zürich

4

Franz

Basel

5

VW

2013

Owner

Model

Year

Id

6

2
Flexibility of Semi-Structured Data Models

- Data must not be compliant to a fixed schema a priori

- Reasons against explicit structuring of data
  - Not meaningful
  - Not uniquely possible
  - Not known (completely)
  - Complete structuring to expensive

- Data not always easily mappable to simple tables
  - Optional characteristics
  - Repetition of elements
  - Element order potentially relevant
  - Structuring however desired, e.g. for fast data access or fine-grained search
Semi-Structured Data Models: Example (1)

Authors do not have always a first and a last name
Avoid null values by allowing different author structures
Semi-Structured Data Models: Example (3)

Books can have arbitrary reviews but must not have one
Conclusions: Motivation for XML via Semantic Data Models

- **Semi-structured data model supports flexible structures**
  - XML documents of a given type can be structured differently

- **Self-description avoids separation of schema and instance**
  - XML documents can describe data as well as data structure
Motivation of XML via Information Retrieval

- **Database systems manage well-structured data**
- **Information-Retrieval-Systems manage less-structured documents**
  - Text documents: arbitrary content length, potentially structured in paragraphs
  - Hyper-Text documents: additionally contain links to related text positions
  - Multi-Media documents: may include images, audios, and videos
  - Documents may significantly differ in structure and size

- **Compared to flat information retrieval models, XML supports**
  - Better structuring possibilities
  - Integrated meta-data description

- **Information-Retrieval may benefit from XML**
  - Structure information is explicitly available in contrast to HTML
  - Access granularity is more fine-granule in contrast to flat document structures
Information Retrieval on Flat Documents

- Document index and query described by characteristics
- Query hits based on characteristics match

**TEXT DOCS**
- D1: The dogs walked home...
- D2: Home on the range.

**INVERTED INDEX**
- dogs → D1, D5, D78
- walked → D1, ...
- home → D1, D2, ...
- on → D2, ...
- range → D2, ...

**TRANSFORMED QUERY**
- dogs home

**SIMILARITY MEASURE**
- $RSV(Q,D1) = .75$
- $RSV(Q,D2) = .34$

**LIST OF RANKED DOCUMENTS**
- D1
- D2

**RESULTS**

Searchers

Lecture "XML and Databases" - Dr. Can Türker
Information Retrieval on Structured Documents

TAGGED DOCS

<DOC>
<DOCNO> D1
</DOCNO>
<TEXT> The dogs walked home. ... </TEXT>
</DOC>

<DOC>
<DOCNO> D2
</DOCNO>
<TEXT> Home on the range. </TEXT>
</DOC>

INVERTED INDEX

dogs --> D1, D5, D78
walked --> D1, ... home --> D1, D2, ... on --> D2, ... range --> D2, ... ... 

TRANSFORMED QUERY

dogs home

RSV(Q,D1) = .75
RSV(Q,D2) = .34

SIMILARITY MEASURE

LIST OF RANKED ELEMENTS

D1: /DOC/TEXT
D2: /DOC/TEXT

Can we use the same ranking function as for flat documents?
Conclusions: Motivation of XML via IR

- **Information-Retrieval-Systems**
  - Support retrieval of relevant text fragments
  - Extract words and word sequences
  - Abstract from document structure

- **Information Retrieval with XML**
  - XML makes structure of certain text elements explicit
  - Finer access and retrieval granularity, e.g., restriction to certain XML elements
  - Increases document search hit potential by exploiting XML document structure
  - Better retrieval results by improved ranking based on integrated structure information

- **Traditional IR ranking approaches such as the vector-space model deliver better quality when document structure is included**

![Diagram of XML structure and retrieval process](image)
Conclusions: Motivation for XML

General advantages

- Platform-independent
- Producer-independent
- Language-independent
- Unicode-based („international“)
- Extensible
- Usable at all application levels
  - Storage
  - Application logic
  - Presentation

Advantages from business perspective

- Support extended search possibilities on distributed, heterogenous data
- Allow new functionality on existing data
- Ease data transformation
- Connect applications
- Reduce application development time
- No (little) license costs since many XML tools are freely available
Motivation for XML in Databases

- **XML as data exchange format**
  - XML documents describing business relevant data
  - Archiving original documents
  - Transformation into other formats
  - Storing XML data in pipeline steps

- **Applications on XML data**
  - XML as internal data structure, e.g., to describe configuration files, application data, or style sheets
  - Requirements for adequate XML storage?

- **Advantages of storing XML in databases**
  - Powerful and efficient search functions
  - Transactional, persistent storage
  - Concurrency control
  - Access control