

# XML background

Debora Donato

Yahoo! Research – Barcelona, Spain

Seminars of Computer Networks  
Thanks to Eyal Oren and Peter Mikka

- 1 Introduction
- 2 XML properties
- 3 XML Data Model

- 1 Introduction
- 2 XML properties
- 3 XML Data Model

# Move toward Next Generation Networks

## Web 2.0 vs Web 1.0

- Web 1.0 was about reading, Web 2.0 is about writing
- Web 1.0 was about companies, Web 2.0 is about communities
- Web 1.0 was about client-server, Web 2.0 is about peer to peer
- Web 1.0 was about HTML, Web 2.0 is about XML
- Web 1.0 was about home pages, Web 2.0 is about blogs
- Web 1.0 was about portals, Web 2.0 is about RSS
- Web 1.0 was about taxonomy, Web 2.0 is about tags
- Web 1.0 was about wires, Web 2.0 is about wireless
- Web 1.0 was about owning, Web 2.0 is about sharing
- Web 1.0 was about IPOs, Web 2.0 is about trade sales
- Web 1.0 was about Netscape, Web 2.0 is about Google
- Web 1.0 was about web forms, Web 2.0 is about web applications
- Web 1.0 was about screen scraping, Web 2.0 is about APIs
- Web 1.0 was about dialup, Web 2.0 is about broadband
- Web 1.0 was about hardware costs, Web 2.0 is about bandwidth costs

This entry was posted on Monday, May 29th, 2006 at 12:34 am and is filed under [AJAX](#), [Business](#), [Ideas](#), [RSS](#), [Software](#), [Web 2.0](#), [bebo](#), [broadband](#), [del.icio.us](#), [flickr](#). You can follow any responses to this entry through the [RSS 2.0 feed](#). You can [leave a response](#), or [trackback from your own site](#).

# Outline

## This week

- XML background
- Index Structures for XML and Text
- Ranking Query Results for XML IR

## Next Week week

- Taxonomies, Folksonomies and Tagging - How social bookmarking can improve Web Search
- Context Search
- Quality in User Generated Content

# XML

## eXtensible Markup Language

The extensible markup language (XML) is a markup language that has been defined to **structure**, **store** and **send** information.

# XML

## eXtensible Markup Language

The extensible markup language (XML) is a markup language that has been defined to **structure**, **store** and **send** information.

It is characterized by:

- Extensible set of tags

<http://www.w3.org/TR/REC-xml>

# XML

## eXtensible Markup Language

The extensible markup language (XML) is a markup language that has been defined to **structure**, **store** and **send** information.

It is characterized by:

- Extensible set of tags
- Describe and exchange arbitrary information (not just Web documents)

<http://www.w3.org/TR/REC-xml>

# XML

## eXtensible Markup Language

The extensible markup language (XML) is a markup language that has been defined to **structure**, **store** and **send** information.

It is characterized by:

- Extensible set of tags
- Describe and exchange arbitrary information (not just Web documents)
- Markup follows syntactic rules: check for well-formedness

<http://www.w3.org/TR/REC-xml>

# XML

## eXtensible Markup Language

The extensible markup language (XML) is a markup language that has been defined to **structure**, **store** and **send** information.

It is characterized by:

- Extensible set of tags
- Describe and exchange arbitrary information (not just Web documents)
- Markup follows syntactic rules: check for well-formedness
- Document schema may be defined in DTD or XSD: check for validity

<http://www.w3.org/TR/REC-xml>

# In a nutshell

## eXtensible Markup Language

The extensible markup language (XML) is essentially a textual representation of the hierarchical (tree-like) data where a meaningful piece of data is bounded by matching starting and ending tags such as

```
<text>This is an XML document</text>
```

# Data on the Web

HTML: Web documents

# Data on the Web

HTML: Web documents

CSS: reusable layouts

# Data on the Web

HTML: Web documents

CSS: reusable layouts

XML: arbitrary Web data

# Data on the Web

HTML: Web documents

CSS: reusable layouts

XML: arbitrary Web data

XSD/DTD: XML schemas

# Data on the Web

HTML: Web documents

CSS: reusable layouts

XML: arbitrary Web data

XSD/DTD: XML schemas

XQuery/Xpath: accessing XML data

# Data on the Web

HTML: Web documents

CSS: reusable layouts

XML: arbitrary Web data

XSD/DTD: XML schemas

XQuery/Xpath: accessing XML data

XSLT: transforming XML data

# HTML

## Hypertext Markup Language

The hypertext markup language that has been defined to **visualize** data on the Web

# HTML

## Hypertext Markup Language

The hypertext markup language that has been defined to **visualize** data on the Web

Defines set of tags and their mandatory/optional interpretation

- Example tags: `<html>`, `<head>`, `<body>`, `<h1>`, `<p>`, `<b>`

<http://www.w3.org/TR/REC-html40/>

# HTML

## Hypertext Markup Language

The hypertext markup language that has been defined to **visualize** data on the Web

Defines set of tags and their mandatory/optional interpretation

- Example tags: `<html>`, `<head>`, `<body>`, `<h1>`, `<p>`, `<b>`
- Head: metadata

<http://www.w3.org/TR/REC-html40/>

# HTML

## Hypertext Markup Language

The hypertext markup language that has been defined to **visualize** data on the Web

Defines set of tags and their mandatory/optional interpretation

- Example tags: `<html>`, `<head>`, `<body>`, `<h1>`, `<p>`, `<b>`
- Head: metadata
- Body: structure and content

<http://www.w3.org/TR/REC-html40/>

- 1 Introduction
- 2 XML properties
- 3 XML Data Model

# XML Syntax (I)

- Hierarchy of elements (XML element tree)
- Elements have names (tags), values (content) and attributes
- Elements can be nested

```
<?xml version="1.0" encoding="UTF?8"?>
  <country name= "The Netherlands" >
    <geography>
      <capital name= "Amsterdam" >
        <remark> The Hague is the seat of the government
        </remark>
      </capital>
      <neighboring_country> Germany </neighboring_country>
      <neighboring_country> Belgium </neighboring_country>
    </geography>
  </country>
```

## XML Syntax (2)

- Mandatory header

```
<?xml version="1.0" encoding="UTF?8"?>
```

- Elements: basic components
  - start-tag, end-tag and content
  - the root must be unique
  - content can be text, other elements (nested) or nothing
- Attributes: name-value pair inside tag

```
<person firstname="John" lastname="Smith"/>
```

## XML Syntax (3)

```
<?xml version="1.0" encoding="UTF?8"?>
<?xml:stylesheet type="text/css2" href="style.css"?>

<country name= "The Netherlands" >
  <geography>
    <capital name= "Amsterdam" >
      <remark> the seat of the government is The Hague
    </remark>
    </capital>
    <neighboring_country> Germany </neighboring_country>
    <neighboring_country> Belgium </neighboring_country>
  </geography>

  <!-- Should be extended with other data ??>
</country>
```

## XML Syntax (4)

- comments: ignored by parser

```
<!-- comment -->
```

- processing instructions: passed to application

```
?xml:stylesheet type="text/css2" href="style.css"?)
```

- XML is well-formed if:
  - nesting is well-balanced
  - attribute names unique within element

## XML Syntax (5)

```
<name>
    <firstName>Vincent
    <lastName>van Gogh
    </firstName>
    </lastName>
</name>
```

```
<name>
  <firstName>Vincent</firstName>
  <lastName>van Gogh</lastName>
```

# XML Namespaces

- Combining documents can lead to naming collisions: book title and recipe title
- Namespace provide naming context URI for elements and attributes
- Namespace prefixes provide shorthand notation

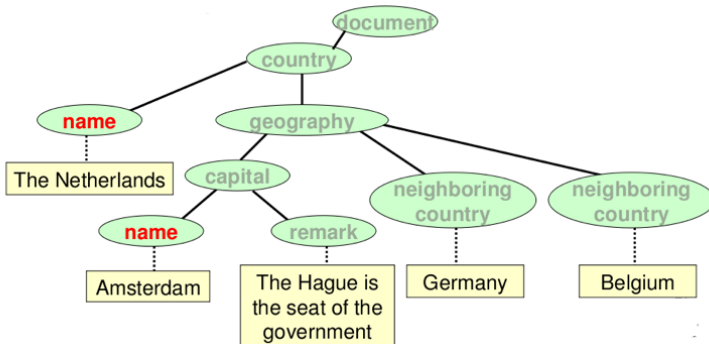
```
<collection
  xmlns:books="http://www.oclc.org/books/1.0/"
  xmlns:webpage="http://www.w3c.org/html/1.0/">

  <book>
    <books:title>Gulliver's travels</books:title>
  </book>
  <web>
    <webpage:title>My first homepage</webpage:title>
  </web>
</collection>
```

- 1 Introduction
- 2 XML properties
- 3 XML Data Model**

# XML Data Model (I)

- Well-formed: if document syntactically correct
- Syntax: alphabet and grammar: '<', '/', '&', etc.
- Data model: how to interpret syntax



## XML Data Model (2)

- Ordered labelled tree
- Exactly one root, no cycles
- Every non-root node has one parent
- Every node can have a label
- Order is important for elements

## XML Data Model (3)

- XML data model: limited meaning
  - elements with names
  - parent-child relations
  - element values
- Not:
  - concepts/classes
  - concept properties
  - class hierarchy
- XML defines document structure

## Beyond XML syntax

- XML represents structured information
- XML allows for arbitrary structures
- Structure can be agreed upon and described using DTDs or XSDs
- Validity of instance documents can be verified against these schemas
- XML document is valid if: well-formed and conforms to XSD/DTD

## Structuring using DTDs

- DTD: document type definition
- Associated using 'DOCTYPE' statement
- Element nesting, order, multiplicity, attributes
- Only few datatypes

```
<!ELEMENT country (geography, people, economy)>
<!ATTLIST country
name CDATA #REQUIRED>
<!ELEMENT geography (capital, neighboring_country*)>
<!ELEMENT capital (remark*)>
<!ATTLIST capital
name CDATA #REQUIRED>
  <!ELEMENT remark (#PCDATA)>
<!ELEMENT neighboring_country (#PCDATA)>
```

## Structuring using XSDs

- W3C standard for XML Schema Definition (2001)
- Schema language like DTD, but:
  - Expressed in XML
  - Several datatypes
  - Richer grammar

```
<complexType name="capital">  
<element name="name" type="string"/>  
<element ref="remark" maxOccurs="unbounded"/>  
</complexType>
```

## XSD grammar (1)

- Cardinality: minOccurs, maxOccurs
- Content models: choice, sequence, all
- Attribute values: default, fixed

```
<complexType name="WindowsType">
  <element name="version" type="string" minOccurs="0"
    maxOccurs="1" default="W98"/>
  <element name="includedBrowser" type="string"
    minOccurs="0" maxOccurs="1" fixed="Internet Explorer"/>
</complexType>
```

## XSD grammar (2)

```
<schema
  xmlns="http://www.w3.org/2001/XMLSchema"
  xmlns:po="http://www.example.com/purchaseOrder">

  <element name="purchaseOrder" type="po:type"/>
  <element name="comment"       type=":string"/>
  <element name="anotherComment" type="xsd:string"/>
  <!-- etc. -->
</schema>
```



## Summary HTML and XML

- HTML: fixed set of tags, represent document structure and layout, presentation model defined in CSS
- XML: arbitrary set of tags, schema may be specified in DTD or XSD
  - well-formed: correct syntax, nested tags
  - valid: conforms to schema definition
- XHTML: HTML4.0 in XML
- How to represent XML data in HTML (web page)?
- How to transform XML document?
- How to query XML data?

Thanks to Eyal Oren and Peter Mika