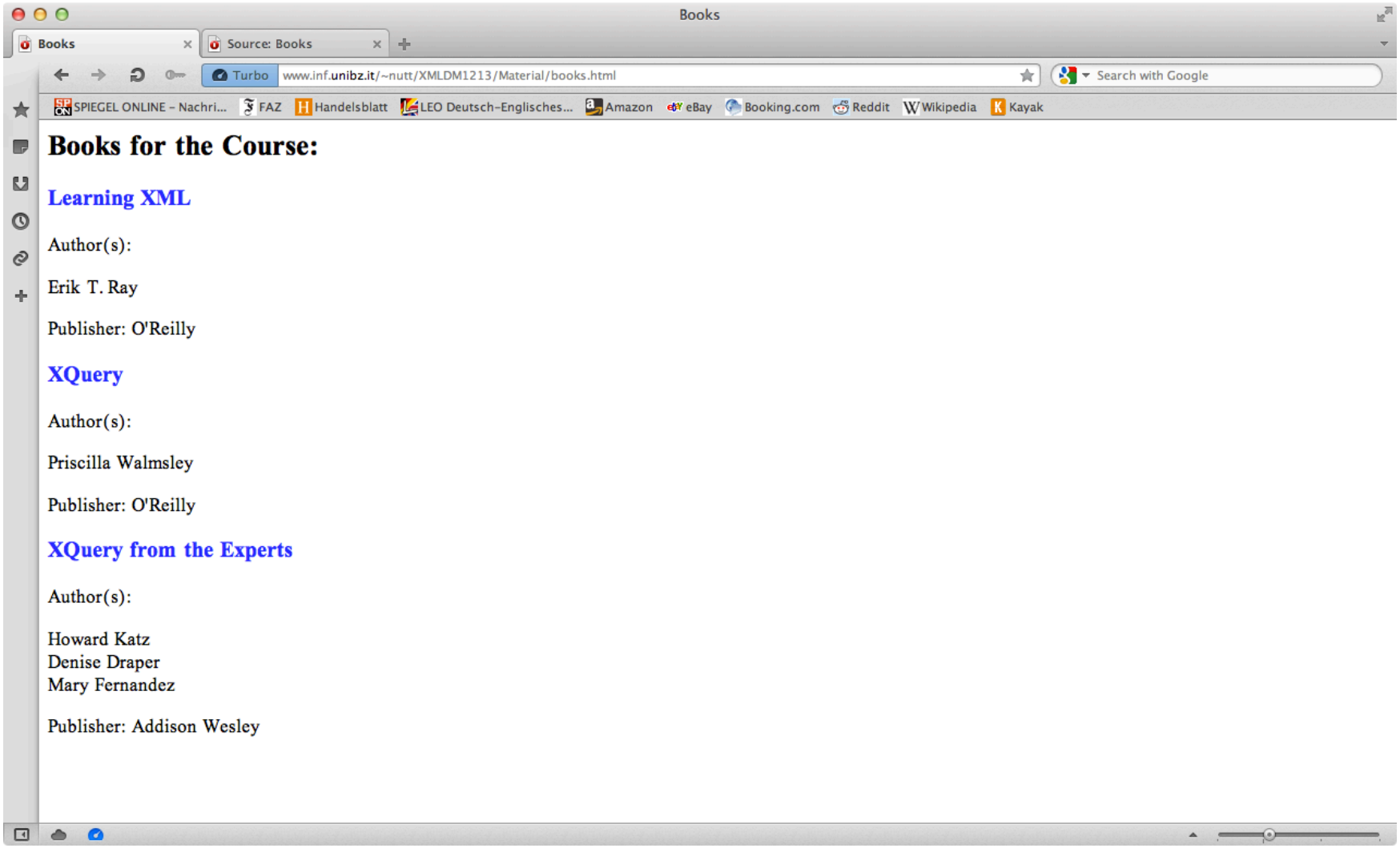
XML Data Management

Motivation:

What is XML? Where does it occur?
How is it used?

Werner Nutt

Web Pages are in HTML



books.html

```
<html>
<head>
<meta http-equiv="Content-Type"
      content="text/html; charset=UTF-8">
<title>Books</title>
</head>
<body>
<h2>Books for the Course:</h2>
<h3><font color="3333ff">Learning XML</font></h3>
<p>Author(s):  <p>Erik T. Ray<br></p></p>
<p>Publisher: O'Reilly</p>
<h3><font color="3333ff">XQuery</font></h3>
<p>Author(s):  <p>Priscilla Walmsley<br></p></p>
<p>Publisher: O'Reilly</p>
<h3><font color="3333ff">XQuery from the Experts</font></h3>
<p>Author(s):  <p>Howard Katz<br>Denise Draper<br>Mary
   Fernandez<br></p></p>
<p>Publisher: Addison Wesley</p>
</body>
</html>
```

tags

attributes

data

Web Pages are in HTML

- HTML is a markup language
- An HTML page consists of tags
  with attributes and data
- HTML describes the style of the page
  (e.g., color, font type, etc.)

HTML: Pros and Cons

+ Easy to read/view the displayed HMTL for humans
+ Standardization makes content publishers independent from specific browsers
+ Many possibilities to display text, images, forms, …

− Fixed vocabulary of tags and attributes
− Content and presentation are mixed
− Humans can grasp the meaning, machines can't:
  what are the titles? where are the authors?
− No easy way to transfer this info and
  combine it with similar other info

Data on the Web are in XML

books.xml

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<?xml-stylesheet type="text/xsl" href="books.xsl"?>
<books>
   <book title="Learning XML">
      <authors>
      <author>Erik Ray</author>
      </authors>
      <publisher>O'Reilly</publisher>
   </book>
   <book title="XQuery">
      <authors>
               <author>Priscilla Walmsley</author>
      </authors>
      <publisher>O'Reilly</publisher>
  </book>
```

tags

declarations
(optional)

attributes

data

books.xml (cntd.)

```
   <book title="XQuery from the Experts">
      <authors>
       <author>Howard Katz</author>
       <author>Denise Draper</author>
       <author>Mary Fernandez</author>
      </authors>
      <publisher>Addison Wesley</publisher>
   </book>
</books>
```

What's the Difference?

-
-
-
-
-

What's the Difference?

- Display info is missing
- Tags express logical structure
- Attributes, too, contain data with information content
- Tree hierarchy reflects logical hierarchy in information
- Tags are not prescribed, but can be chosen freely
- HTML is "essentially" a special case of XML ($\Rightarrow$ XHTML)

Both Files Looked the Same in the Browser …

One line in **`books.xml`** gives us a hint:

```
<?xml-stylesheet type="text/xsl" href="books.xsl"?>
```

This is a reference to the file

`books.xsl` ,

which is a stylesheet telling the browser

how to display the info in

`books.xml`

books.xsl

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<xsl:stylesheet version="1.0"
          xmlns:xsl="http://www.w3.org/1999/XSL/Transform”>

  <xsl:template match="/">
    <html>
      <head>
        <title>Books</title>
      </head>
      <xsl:apply-templates select="books"/>
    </html>
  </xsl:template>
```

books.xsl (contd.)

```
<xsl:template match="books">
  <body>
     <h2>Books for the Course:</h2>
     <xsl:apply-templates select="book"/>
  </body>
</xsl:template>

<xsl:template match="book">
  <h3><font color="3333ff">
       <xsl:value-of select="@title"></xsl:value-of>
      </font>
  </h3>
  <p>Author(s):
      <xsl:apply-templates select="authors"/></p>
  <p>Publisher:
     <xsl:value-of select="publisher"></xsl:value-of></p>
</xsl:template>
```

books.xsl (contd.)

```
 <xsl:template match="authors">
  <p><xsl:for-each select="author">
          <xsl:value-of select="text()">
     </xsl:value-of><br/>
  </xsl:for-each></p>
 </xsl:template>

</xsl:stylesheet>
```

Observations about the Stylesheet

-
-
-
-
-

Observations about the Stylesheet

- Similar syntax as XML
- HTML tags and attributes occur in the code
- Specific tags: **`xsl:template`**, **`xsl:value-of`**
- Specific attributes: **`match`**, **`select`**
- Document consists of templates with
  - match pattern
  - new stuff to be generated (HTML)
  - calls to apply template

Stylesheets

The stylesheet is written in XSLT,
which is part of XSL (= eXensible Stylesheet Language)

XSL consists of three parts

- XPath *non-XML notation*
  - addresses parts of a document
- XSLT *(= XSL Transformations)* *XML notation*
  - transforms XML to XML, HTML, text
- XSL-FO *(= XLS Formatting Objects)* *XML notation*
  - talks about pages, regions, blocks, lines
    can be processed to PDF, RTF etc.

XML, XLS, and (X)HTML

How do they work together?

There is an XSLT interpreter at

http://www.w3schools.com/xsl/tryxslt.asp?xmlfile=cdcatalog&xsltfile=cdcatalog

XML

- XML: Extensible Markup Language
- Defined by the WWW Consortium (W3C)
- Originally intended as
  - a document markup language
  - not a data model

What is the W3C?

World Wide Web Consortium http://www.w3.org/

- International Standards Organization for the WWW
- Standards are called “recommendations”
- Head and founder Tim Berners-Lee

The recommendation for XML is at

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

XML as a Markup Language

- Documents have tags with info about document parts

  e.g. **<title> XML </title>**

  <figure caption="Tree Structure“> …
  </figure>

- XML is derived from SGML (Standard Generalized Markup Language), but simpler

- Extensible, unlike HTML
  - Users can add new tags, and *separately* specify how the tag should be handled for display

- Goal was to replace HTML as the language for publishing documents on the Web

History: SGML, HTML, XML

SGML: Standard Generalized Markup Language

Charles Goldfarb (IBM), ISO 8879, 1986

- DTD (Document Type Definition)
  powerful formalism for structuring information, but
  - full implementation of SGML difficult
  - tools for working with SGML documents expensive
- Two sub-languages of SGML made their way
  - HTML: HyperText Markup Language (Berners-Lee, 1991).
    Describing presentation
  - XML: eXtensible Markup Language, W3C,1998.
    Describing content.

XML Around us …

- Conference Proceedings VLDB 2011 on my laptop:

  Mix of HTML and XML (table of contents, author list),

  XML displayed with XSLT

- "Citizens' Network South Tyrol" (province website)

  http://www.provinz.bz.it/

  The "HTML" there is XHTML
  - HTML is an SGML-language, XHTML is an XML-language
    ⇒ follows stricter rules, can be more easily parsed
  - several differences, e.g. case-insensitive vs. case-sensitive

XML Around us … (cntd)

- Open Office document "HelloWorld.odt"
  - follows the OpenDocument (ODT) standard by OASIS
  - .odt, .ods, .odp file types for text, spreadsheets, presentations etc.
  - a document is zipped archive ⇒ **`unzip HelloWorld.odt`**
    the main file is **`content.xml`**

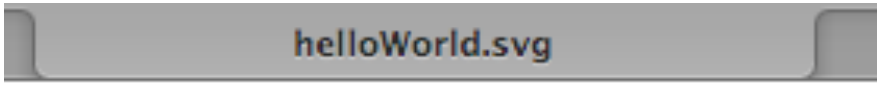
- MS Word 2007 document "HelloWorld.docx"
  - follows the Office Open XML standard by Microsoft
  - .docs, .xslx, .pptx file types
  - Again a document is zipped archive
    the main file is **`document.xml`**

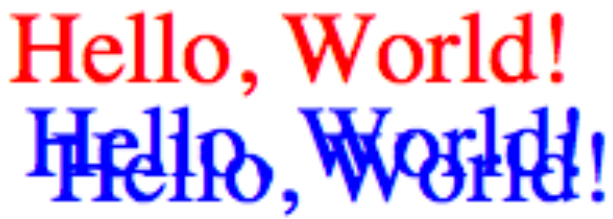
Both standards are based on XML.

ODT is supported by IBM, Google, Adobe, LibreOffice, …

XML Around us … (cntd)

- Scalable Vector Graphics file "HelloWorld.svg"
  - follows SVG 1.1 Recommendation of W3C





Why SVG?

- Scalability: Can be shrunk/enlarged w/o loss of detail
- Accessibility: graphics can be parsed
- Interactivity: SVG elements can trigger events
- Scripting: the elements of an SVG figure can be accessed and manipulated by a scripting language (like JavaScript)

XML Around us … (cntd)

XML dumps are available of

- Wikipedia http://en.wikipedia.org/wiki/Wikipedia:Database_download
- Gene Ontology http://www.geneontology.org/GO.downloads.ontology.shtml
- DBLP (Computer Science Bibliography) http://dblp.uni-trier.de/xml/

. . .

XML updates are available from

- Weather forecast South Tyrol http://www.provinz.bz.it/wetter/services.asp
- Medline/Pubmed http://www.nlm.nih.gov/bsd/licensee/data_elements_doc.html

. . .

GPX — GPS Exchange Format

Go to http://www.trekking.suedtirol.info/, define a hike,

and download the coordinates to your GPS device

`trekking.gpx` contains XML

Google Geocoding API

Google offers a REST (Representational State Transfer) interface for retrieving geocoding:

- Input: address
- Output: geographic coordinates

With coordinates, applications can put symbols on maps

Example:

http://maps.googleapis.com/maps/api/geocode/xml?address=piazza+universita+bolzano+italy&sensor=false

Results come in XML (and JSON)

Other Usages of XML

- MathML (Mathematical Markup Language)

  http://en.wikipedia.org/wiki/MathML

- CML (Chemical Markup Language)

- SPL (Structured Product Labeling)
  human prescription drug labeling

  http://www.fda.gov/ForIndustry/DataStandards/StructuredProductLabeling/default.htm

- OWL (Web Ontology Language)
- RDF XML Serialization
  (RDF = Resource Description Format)
- WSDL (= Web Service Description Language)
- SOAP (= Simple Object Access Protocol) messages
  web services

XML = Data Exchange Format

- Data are exchanged
  - across platforms
  - across enterprises

- Data are integrated
  - from heterogeneous data sources
  - from data sources distributed across different locations

- Different application areas need different standards

  $\Rightarrow$ XML is the basis for data interchange formats