Birkbeck
UNIVERSITY OF LONDON

XML

Session 1:
Introduction to XML

Introduction

- **Muawya Eldaw**

  eldaw@dcs.bbk.ac.uk
- **ITApps Administrator:**

  itapp@dcs.bbk.ac.uk
- **Access to My Birkbeck:**

  http://www.bbk.ac.uk/mybirkbeck/
- **Systems Group (DCSIS):**

  sg@dcs.bbk.ac.uk
- **Birkbeck IT Services (ITS):**

  http://www.bbk.ac.uk/its/

Module Timetable

Session 1: Introduction to XML

Session 2: Using and Creating XML Documents

Session 3: Document Type Definition (DTD) – Part 1

Session 4: DTDs, Namespaces and XML Schemas

Session 5: More XML Schemas

Session 6: XSL and XPath

Session 7: E**X**tensible **S**tylesheet **L**anguage **T**emplates

Session 8: More XSLT and examples of use of XML in programming

Module Timetable

- **Tutor Marked Assignment (TMA)**
  - 25% of overall marks
  - To be completed and submitted by 6pm, Monday 16th February 2015
- **Final Module Assignment (FMA)**
  - 75% of overall marks
  - To be completed and submitted by 6pm on Tuesday 14th April 2015

Note: there is a cut-off deadline for tutors to review FMA submissions, which is 2 weeks before the relevant deadline

- **Further details about the ITApps Assessment process**

  See: http://www.dcs.bbk.ac.uk/itapps/assessment.html

XML – Learning Objectives

- Upon completion of the module you will be able to:
  - Learn how an XML document is defined to be well-formed and valid.
  - Create markup pages of data using tags and attributes.
  - Know the difference between well-formed and not well-formed XML documents.
  - Know how to validate documents in order to make them well-formed XML documents.

XML – Learning Objectives

- Create and design DTDs (Document Type Definition) for XML documents.
- Create and design XML Schemas.
- Learn how to create namespaces.
- Learn XPath to query an XML document.
- Have the ability to transform an XML document into another XML document using XSLT.
- Learn how XSLT can be used to create HTML files.

Recommended Reading

- Fawcett, J. et al, Beginning XML 5th edition, Wrox, ISBN 978-1-118-16213-2
- Bradley, N, The XML companion, 3rd edition, Addison Wesley, ISBN 978-0201770599
- Holzner, S. Sams teach yourself XML in 21 days (3rd Edition). Sams, 2003.
- Harold, E. R. & Means, W. S. XML in a Nutshell (3rd Edition). O'Reilly, 2004.
- Eric Ray, Learning XML, O'Reilly & Associates, Inc., Sebastopol, CA, January 2001, ISBN 0-596-00046-4.

Introduction to the Extensible Markup Language (XML)

- SGML, HTML, and XML are the most important markup languages.
  SGML because it is the parent language of both HTML and XML, HTML because it is the language of the web, and XML because it describes the data on the web.

Standard Generalised Markup Language (SGML)

Birkbeck UNIVERSITY OF LONDON

- In the late 1960s, IBM researchers worked on the problem of building a portable system for the interchange and manipulation of legal documents.
- Their prototype language marked up structural elements, with formatting information kept in separate files, called style sheets. The document structure was defined in yet another file, called a Document Type Definition (DTD).
- By 1969, the researchers had developed the General Markup Language (GML).
- After further work worldwide, in 1986, the International Standards Organisation (ISO) adopted a particular version called the Standard Generalised Markup Language (SGML). It quickly became the business standard for data storage and interchange.

– Advantages of SGML
  - Long-term viability as an ISO standard
  - Non-proprietary and platform-independent
  - Supports user-defined tags reflecting the richness of documents
– Disadvantages of SGML
  - Costly to set up, requiring real expertise
  - SGML tools are expensive, compared to those for HTML
  - Creating DTDs with SGML is expensive - especially labour
  - SGML has a steep learning curve
  - Put bluntly, it is too elaborate for the ever-changing web.

HyperText Markup Language (HTML)

- Tim Berners-Lee and Robert Calliau, working independently from the other at CERN, invented the *HyperText Markup Language* (HTML) based on SGML.
- HTML is one particular SGML DTD that is easier to learn and use than SGML.
- HTML is a trimmed-down version of SGML, eliminating SGML features that are rarely needed, but including hyperlinks to link web documents.

Sample of an HTML page

```
<html>
   <head>
      <title>This is the title of the page</title>
   </head>
   <body>
      <p> This is the main details of my page </p>
   </body>
</html>
```

Filename: MyFileSample.html

Cascading Style Sheets (CSS)

- With earlier versions of HTML, web browsers controlled the appearance (rendering) of every web page.
- With the advent of Cascading Style Sheets (CSS), the document author can control the way the browser renders the page, or the entire web site for that matter.
- Style sheets allow document authors to specify the style of their page elements (spacing, margins, etc.) separately from their structure (section headers, body text, etc.), thus allowing greater manageability.

Sample CSS

```
/* This is a CSS example */
p
{
    text-align: center;
    color: black;
    font-family: Arial, sans-serif;
}
```

Filename: MyStyle.css

Extensible Markup Language (XML)

- The Extensible Markup Language (XML) is also a descendant of SGML.
- XML was conceived as a solution to the problem of different formats to define data in a uniform way across many domains, while at the same time the file created is easily readable by both machines and humans.
- It is used to represent low-level data, e.g. configuration files, or to add metadata to documents, e.g. as a way to inform that a particular sentence is of importance.

Sample XML Document

```
<?xml version="1.0" encoding="UTF-8"?>
<student>
   <firstname>John</firstname>
   <surname>Smith</surname>
   <birthday>
      <day>06</day>
      <month>12</month>
      <year>1975</year>
   </birthday>
</student>
```

Filename: MyStudentExample.xml

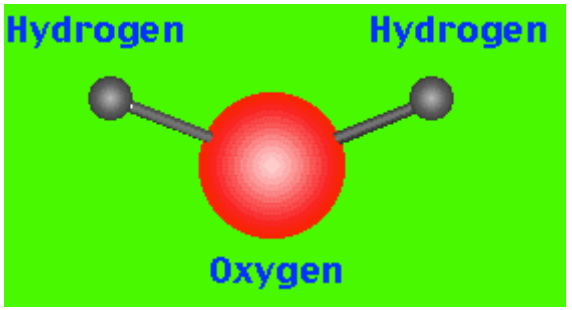
Why make use of XML and the importance of it in the business world

- XML *languages* are being developed for many areas of document processing and e-commerce.
- Example: Chemical Markup Language (CML)

  Peter Murray-Rust's Chemical Markup Language is used for representing molecular and chemical information (www.cellml.org).

Birkbeck
UNIVERSITY OF LONDON

Example of a Water Molecule:



- The following illustrates the CML document for a water molecule (H2O):

```
<?xml version="1.0" encoding="UTF-8"?>
<cml>
        <mol title="Water">
                <atoms>
                        <array builtin="elsym">H O H</array>
                </atoms>
                <bonds>
                        <array builtin="atid1">1 2</array>
                        <array builtin="atid2">2 3</array>
                        <array builtin="order">1 1</array>
                </bonds>
        </mol>
</cml>
```

Filename: WaterMoleculeCMLExample.xml

Mathematical Markup Language (MathML)

- The Mathematical Markup Language [MathML] was developed for describing mathematical notations and expressions using XML.
- It allows mathematical expressions to be processed by different applications for different purposes (www.w3.org/Math).
- MathML Example for the quadratic equation $x^2+4x+4=0$.

Sample MathML


```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE math SYSTEM "http://www.w3.org/TR/MathML2/dtd/mathml2.dtd">
<math xmlns="http://www.w3.org/1998/Math/MathML">
    <mrow>
        <mrow>
            <msup>
                <mi>x</mi>
                <mn>2</mn>
            </msup>
            <mo>+</mo>
            <mrow>
                <mn>4</mn>
                <mo>&InvisibleTimes;</mo>
                <mi>x</mi>
            </mrow>
            <mo>+</mo>
            <mn>4</mn>
        </mrow>
        <mo>=</mo>
        <mn>0</mn>
    </mrow>
</math>
```

MathMLExample.xml

- The **<mi>** element is for identifiers.
- The **<mn>** element is for numbers.
- The **<mo>** element is for operators, etc.
- The entity **⁢** is important – it is invisible when rendered for viewing, spoken when rendered for voice, but indicates multiplication if the equation is being computed!
  - Due to security limitations, the invisible times entity causes a parsing error when the file is viewed on new browsers.

Wireless Markup Language (WML)

- The *Wireless Markup Language* [WML] allows web pages to be displayed on wireless devices such as cellular phones and PDAs.
- WML works with the *Wireless Application Protocol* (WAP version 1) to deliver the content.
- It is still used extensively in the Asian and African markets, due to its design geared towards low bandwidth connectivity.

Sample WML

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE wml PUBLIC "-//WAPFORUM//DTD WML 1.1//EN" "http://
www.wapforum.org/DTD/wml_1.1.xml">
<wml>
   <card title="Welcome to my WML page">
      <p>
         This is my holiday movie page <br/> Click on the link
          below <br/> to play my 3gpp holiday movie.
      </p>
      <p>
         <a href="rtsp://wmlweb.com/holidaymovie.3gp">
            Play my holiday
         </a>
      </p>
   </card>
</wml>
```

Filename: WMLExample.wml

SVG vector images using XML

- Vector images can be represented with XML. SVG is the most popular format.
- The advantage of using XML over a traditional bitmap binary is that the images can be manipulated far more easily. Scaling and other changes become transformations of the XML rather than complex intensive calculations.

Sample SVG

```
<?xml version="1.0" standalone="no"?>
<!DOCTYPE svg PUBLIC "-//W3C//DTD SVG 1.1//EN"
"http://www.w3.org/Graphics/SVG/1.1/DTD/svg11.dtd">

<svg xmlns="http://www.w3.org/2000/svg" version="1.1">
  <circle cx="100" cy="50" r="40" stroke="black"
  stroke-width="2" fill="red" />
</svg>
```

Filename: simpleExample.svg

RDF describing web resources

- RDF (Resource Description Framework) is a framework for describing web resources, such as the title, author, modification date, content and copyright information of a web page.
- It is a W3C standard, designed to be read and **understood** by computers. It is not designed to be displayed to people.
- The RDF language is a part of W3C's Semantic Web activity. The "semantic web vision" is a future where web information has exact meaning; can be understood and processed by computers; computers can integrate information from the web.

Sample RDF

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
syntax-ns#" xmlns:si="http://www.dcs.bbk.ac.uk/rdf/">

<rdf:Description rdf:about="http://www.dcs.bbk.ac.uk/">
  <si:title>DCS homepage</si:title>
  <si:author>Phil Gregg</si:author>
</rdf:Description>
</rdf:RDF>
```

Filename: simpleExample.rdf

RSS web content distribution

- RSS (Rich Site Summary) is a format for delivering regularly changing content.
- Many news-related sites, blogs and other online publishers syndicate their content as an RSS feed to whoever wants it.
- There are 7 different formats all called RSS. Most popular are 1.0 and 2.0, which are incompatible. For more info:

  http://www.xml.com/pub/a/2002/12/18/dive-into-xml.html

RSS sample

A partial example from http://feeds.bbci.co.uk/news/uk/rss.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet title="XSL_formatting" type="text/xsl" href="/shared/bsp/xsl/rss/nolsol.xsl"?>
<rss xmlns:media="http://search.yahoo.com/mrss/" xmlns:atom="http://www.w3.org/2005/Atom" version="2.0">
  <channel>
    <title>BBC News - UK</title>
    <link>http://www.bbc.co.uk/news/uk/#sa-ns_mchannel=rss&amp;ns_source=PublicRSS20-sa</link>
    <description>The latest stories from the UK section of the BBC News web site.</description>
    <language>en-gb</language>
    <lastBuildDate>Wed, 18 Sep 2013 19:06:13 GMT</lastBuildDate>
    <copyright>Copyright: (C) British Broadcasting Corporation, see http://news.bbc.co.uk/2/hi/help/rss/
4498287.stm for terms and conditions of reuse.</copyright>
<!– … -->
<item>
      <title>VIDEO: The Big Debate: Choosing Scotland's Future</title>
      <description>Watch coverage of The Big Debate: Choosing Scotland's Future.</description>
      <link>http://www.bbc.co.uk/news/uk-scotland-18210776#sa-ns_mchannel=rss&amp;ns_source=PublicRSS20-
sa</link>
      <guid isPermaLink="false">http://www.bbc.co.uk/news/uk-scotland-18210776</guid>
      <pubDate>Wed, 18 Sep 2013 12:05:29 GMT</pubDate>
      <media:thumbnail width="66" height="49" url="http://news.bbcimg.co.uk/media/images/69913000/jpg/
_69913113_flags_wind.jpg"/>
      <media:thumbnail width="144" height="81" url="http://news.bbcimg.co.uk/media/images/69913000/jpg/
_69913114_flags_wind.jpg"/>
</item>
<!-- … -->
```

References

Useful links:

- www.w3schools.com/xml/default.asp
- en.wikipedia.org/wiki/XML
- xml.coverpages.org/xml.html
- www-128.ibm.com/developerworks/xml/newto
- https://developer.mozilla.org/en/docs/XML_in_Mozilla
- www.cellml.org/tutorial/xml_guide
- webdesign.about.com/library/nosearch/bl_xmlclass1-1.htm
- www.w3.org/TR/NOTE-sgml-xml-971215
- http://www.w3schools.com/svg/default.asp
- http://www.w3schools.com/rdf/default.asp
- http://www.w3schools.com/rss/default.asp
- http://www.xml.com/pub/a/2002/12/18/dive-into-xml.html