XML for Text Markup

An introduction to XML markup.
Where did XML come from?

SGML is the Mother Language

- XML is an offshoot of SGML
- HTML is an offshoot of SGML
The Good About SGML

• **Standard General Markup Language**
• A meta-language for creating descriptive markup languages
• Ability to create unique markup for different projects with the same language
The Good about XML

What XML is....

• Extensible Markup Language
• An offshoot of SGML
• The “syntax” of document structure
• Knowledge representation scheme.
XML vs. HTML

• XML is extensible: it does not contain a fixed tag set
• XML documents must be well-formed according to a defined syntax and may be formally validated
• XML focuses on the meaning of data, not its presentation
What XML is not

• HTML

• A language used for document formatting

• A language where rules don’t apply
Uses of XML

• Storing of metadata
• Communication between machines
• Markup of text for preservation
<?xml version="1.0"?>
    <dc:title>UKOLN</dc:title>
    <dc:description>UKOLN is a national focus of expertise in digital information management. It provides policy, research and awareness services to the UK library, information and cultural heritage communities. UKOLN is based at the University of Bath.</dc:description>
    <dc:publisher>UKOLN, University of Bath</dc:publisher>
    <dc:identifier>http://www.ukoln.ac.uk/</dc:identifier>
</metadata>
<?xml version="1.0" encoding="UTF-8"?>
  <tblZip>
    <Zip>01001</Zip>
    <City>AGAWAM</City>
    <State>MA</State>
  </tblZip>
  <tblZip>
    <Zip>01002</Zip>
    <City>CUSHMAN</City>
    <State>MA</State>
  </tblZip>
  <tblZip>
    <Zip>01005</Zip>
    <City>BARRE</City>
    <State>MA</State>
  </tblZip>
</dataroot>
Document Type Definition

• The list of elements, attributes or entities that a document is allowed to contain.

• A blueprint for what is considered a legitimate markup.

• A device that allows XML documents to have structure that can be interpreted by machines as well as people.
XML Well-Formed ness

- All tags have start and end tags and case matches present.
- There is only one root element in a document tree.
- Empty elements are correctly formatted.
- All elements are properly nested.
- Attribute values are always quoted.
XML Validation

• Well-Formed
• All elements are present in the DTD and have Unique Identifiers
• All attributes and relations between elements are used as described in the DTD
• Parsers check validity of documents based on rules set by the DTD
TEI

- Text Encoding Initiative
- An application of SGML. Specifically a DTD that was designed for encoding text.
- A well accepted, maintained, and supported DTD for text encoding
- Wide coverage, modular, extensible
Basic Structure of TEI

<TEI.2>
  <teiHeader> {Header information} </teiHeader>
  <text>
    <front> {front matter} </front>
    <body> {body of text} </body>
    <back> {back matter} </back>
  </text>
</TEI.2>
Some Rules of XML

All tags must have end tags*

<tag> data </tag>

All tags must be in the same case

<lower> </lower>

All end tags should have a space at the end

Only one root element per document

<TEI.2>

<body> </body>

</TEI.2>
More Rules For XML

All lines must have a space after the last tag.

Special characters ( & $ % ) are important.

& = \&#x0026;
$ = \&#0024;
% = \&#0025;
First tags that we use

All documents start with these tags.

```xml
<?xml version="1.0"?>
<!DOCTYPE TEI.2 SYSTEM "http://www.tei-c.org/Lite/DTD/teixlite.dtd">
```
Most Common Tags

For our projects we will be dealing with some familiar tags.

- `<div> </div>` Division
- `<p></p>` Paragraph
- `<lb />` Line Break
- `<pb n="#"/>` Page break
- `<q> </q>` Quotation
Typographic Tags

<hi rend="italics">italics</hi>
<hi rend="bold">bold</hi>
<hi rend="underscore">underscore</hi>
Paragraph, Quotations...

<p></p> paragraph

<q></q> quotation

<pb N="1" /> page break
Division Tags

When we write division tags `<div1>` we always follow them with the `<head>` tag immediately after, even if there is no heading used.

```html
<div1><head></head>
<p>
</p>
```
Page Break Tags

Page Break Tag

End of Page

Related External Reference

<p><pb n="p042"/><xref to="images/088.jpg">Page Image</xref></p>
XML for Text Markup

Closing Statements

Review