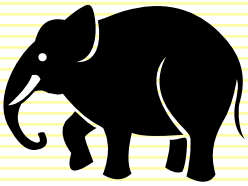
LBSC 690: Session 5

Metadata and XML

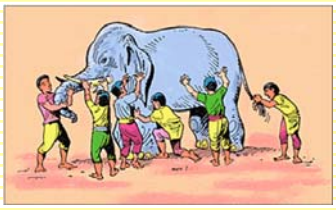
Jimmy Lin
College of Information Studies
University of Maryland

Monday, October 8, 2007

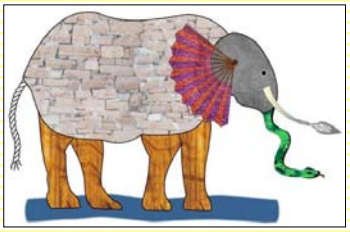
Blind Men and Elephants



Blind Men and Elephants



Is this an elephant?



Metadata

- Literally "data about data"
  - "a set of data that describes and gives information about other data" — Oxford English Dictionary

Data without Metadata...

... can be pretty useless!

Who:
authored it?
to contact about data?

What:
are contents of database?

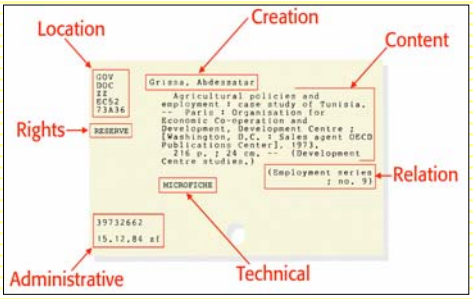
When:
was it collected?
processed? finalized?

Where:
was the study done?

Why:
was the data collected?

How:
were data collected?
processed? Verified?

Early Example of Metadata



Encoding Metadata

- Language for ***expressing*** metadata should be:
  - **Universal** - so all can understand
  - **Flexible** - to incorporate different types
  - **Extensible** - flexible to custom types
  - **Simple** - to encourage adoption
  - **Modular** - so that schemes can be mixed, extended

From: Ian Graham, An Introduction to RDF. http://www.utoronto.ca/ian/talks/

Metadata

- How do we encode metadata?
- How do we encode metadata to support interoperability?

Simple example: **January 31, 2001**
31 janvier 2001
2001-01-31
01-31-2000
31012000

What is the Dublin Core?

- A metadata standard for describing digital resources
- An initiative to create a digital "library card catalog" for the Web
- Dublin Core fields: (all optional)

Title	Creator	Subject
Description	Publisher	Contributor
Date	Type	Format
Identifier	Source	Language
Relation	Coverage	Rights

What is XML?

- XML = eXtensible Markup Language
- XML is a standard for exchanging structured data
  - Provides standardization at the syntactic level
  - Does **not** provide "meaning" for the tags
- XML is a standard recommended by the W3C

Goals of XML

- Easy to use
- Easy to extend and adapt
- Easy to write programs that use XML
- Support a wide variety of applications
- Should be human legible
- Formal and concise

The Basic Rules

- XML is case sensitive
- All start tags must have end tags
- Elements must be properly nested
- XML declaration is the first statement
  - <?xml version="1.0"?>
- Every document must contain a root element
- Attribute values must have quotation marks
  - <item id="33905">
- Certain characters are reserved for parsing
  - < = '<'

Questions about XML

- How is XML like HTML?
- How is HTML like XML?
- What's the relationship between XML and structured documents?
- How are the rules governing a structured document encoded?

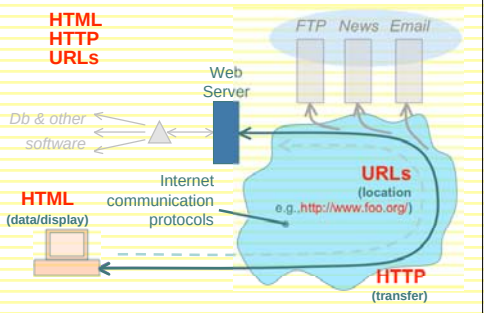
XML: Historic Perspective

- HTML and the birth of the Web
- HTML is not enough
- Development of XML

This section contains slides adapted from presentations by Ian Graham: http://www.utoronto.ca/ian/talks/

In the beginning...

The foundations of the Web:



Three Core Technologies

- **HTTP** - HyperText Transfer Protocol
  - A protocol for transferring data between machines on the Internet
- **URL** - Uniform Resource Locator
  - A scheme for referencing the specific location of a resource
- **HTML** - HyperText Markup Language
  - A markup language for encoding information to be read by humans

HTTP and URLs have pretty-well stood the test of time. But by 1996, HTML was already showing signs of age ....

HTML

- Started with very few tags …
- Language evolved as more tags were added:
  - Forms
  - Tables
  - Fonts
  - Frames
  - …

Problems with HTML

- Desire for personalized tags
  - HTML can't be extended
- Desire to incorporate other types of data
  - Mathematics, database entries, literary text, poems, purchase orders …
  - HTML can't accommodate other types of data
- Desire for automatic processing by software
  - HTML is too messy and inconsistent

Back to the Basics

- HTML was defined using SGML
  - **S**tandard **G**eneralized **M**arkup **L**anguage
  - A meta-language for defining languages
- Complex, sophisticated, powerful
  - … too difficult to use
- Idea: create a simpler version of SGML
  - The birth of XML!

Evolution of XML

- XML can be used to define other languages
- Many XML languages, optimized for different roles
  - MathML: for mathematics
  - SMIL: for synchronized multimedia
  - RSS: for news feeds
  - XHTML: HTML by XML rules
  - RDF: for the Semantic Web
  - …

MathML

- An XML language for defining mathematic formulas

$x^2 + 4x + 4 = 0$

```
<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>
```

See http://www.mozilla.org/projects/mathml/demo/tester.html

MathML

- What advantages does it offer?

SMIL

- Synchronized Multimedia Integration Language
- Integration of multimedia with text, audio, video
- Support in RealPlayer

SMIL Example

```
<smil>
 <head>
   <meta name="title" content="Online Teaching Services promo" />
   <meta name="author" content="Jay Moonah, CAT" />
   <layout type="text/smil-basic-layout">
     <root-layout width="280" height="316" background-color="white"/>
     <region id="AnimChannel1" title="AnimChannel1"
          left="0" top="0" height="265"  width="280" fit="hidden"/>
   </layout>
 </head>
 <body>
   <par title="Online Teaching Services promo" author="Jay Moonah, CAT" >
     <audio src="final.rm" id="Soundtrack"   title="Soundtrack"/>
     <animation src="otscompfin.swf" id="Animation"
          region="AnimChannel1" title="Animation"   fill="freeze"/>
     <text src="cc.rt" id="caption" region="cc" title="cc" fill="freeze"/>
   </par>
 </body></smil>
```


RSS

- RSS = Really Simple Syndication or Rich Site Summary
- An XML format for distributing news headlines on the Web

See example at http://www.nytimes.com/services/xml/rss/

XHTML: Beyond HTML

```
<?xml version="1.0" encoding="iso-8859-1"?>
<html xmlns="http://www.w3.org/TR/xhtml1" >
<head>
  <title> Title of text XHTML Document </title>
</head>
<body>
<div class="myDiv">
   <h1> Heading of Page </h1>
    <p> here is a paragraph of text. I will include  inside this paragraph
        a bunch of wonky text so that it looks fancy. </p>
    <p>Here is another paragraph with  <em>inline emphasized</em>
        text, and <b> absolutely no</b> sense of humor. </p>
    <p>And another paragraph, this one  with an <img src="image.gif"
        alt="waste of time" />  image, and a <br /> line break. </p>
</div>
</body></html>
```


XHTML

- Just like HTML, but based on XML rules
- Will support integration of different data into a single document

XHTML and other Data

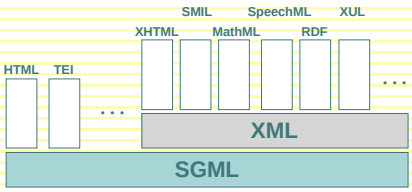
```
<?xml version="1.0" encoding="iso-8859-1"?>
<html xmlns="http://www.w3.org/TR/xhtml1" >
<head>
  <title> Title of XHTML Document </title>
</head><body>
<div class="myDiv">
   <h1> Heading of Page </h1>
    <math xmlns="http://www.w3.org/1998/Math/MathML">
        ... MathML markup ...
    </math>
    <p> more html stuff goes here </p>
   <smil xmlns="http://www.w3.org/TR/smil1">
        ... SMIL markup ...
    </smil>
</div>
</body></html>
```

Demo at:
http://www.umiacs.umd.edu/~jimmylin/LBSC690-2007-Fall/XML-demo/math-demo.xml

And Others…

- **CML** – chemical Markup Lang
- **CellML** – biological models
- **BSML** – bioinformatic sequences
- **MAGE-ML** – Microarray Gene Expression
- **XSTAR** – for archaeological research
- **MARCXML** – MARC in XML
- **AML** – astronomy markup language
- **SportsML** – for sharing sports data

The XML Family Tree



Mixing XML Dialects

- XML is designed to support the integration of multiple standards
- Allows users to mix elements from different standards
  - Snapping together XML dialects like Lego pieces
  - Based on the notion of "namespaces"

Example

```
<?xml version="1.0"?>
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rss="http://purl.org/rss/1.0/"
  xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rss:channel rdf:about="http://www.xml.com/xml/news.rss">
    <rss:title>XML.com</rss:title>
    <rss:link>http://xml.com/pub</rss:link>
    <dc:description>
      XML.com features a rich mix of
      information and services for the XML community.
    </dc:description>
    <dc:subject>XML, RDF, metadata, information
      syndication services</dc:subject>
    <dc:identifier>http://www.xml.com</dc:identifier>
    <dc:publisher>O'Reilly & Associates, Inc.</dc:publisher>
    <dc:rights>Copyright 2000, O'Reilly &
      Associates, Inc.</dc:rights>
  </rss:channel>
</rdf:RDF>
```

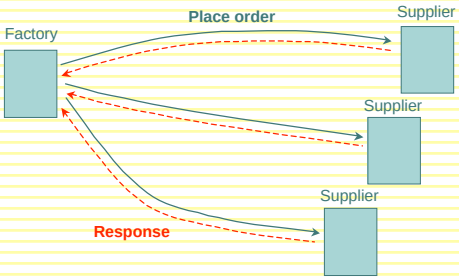
Example from http://www.xml.com/pub/a/2000/10/25/dublincore/

Interoperability

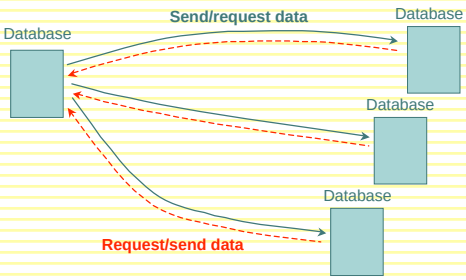
- What does it mean and what's the role of XML?
- XML as a universal format for data interchange
  - Software exchange data as XML-format messages
- Advantages?
  - Eliminates proprietary data formats
  - Promotes interoperability
  - Encourages cooperation
  - Leverages lots of existing XML processing software

Interoperability slides adapted from presentations by Ian Graham: http://www.utoronto.ca/ian/talks/

XML Messaging



XML Messaging



Example Message

```
<partorders xmlns="http://myco.org/Spec/partorders.desc">
  <order ref="x23-2112-2342" date="25aug1999-12:34:23h">
    <desc> Gold sprockel grommets, with matching hamster</desc>
    <part number="23-23221-a12" />
    <quantity units="gross"> 12 </quantity>
    <delivery-date date="27aug1999-12:00h">
  </order>
  <order ref="x23-2112-2342" date="25aug1999-12:34:23h">
     …. Order something else …..
  </order>
</partorders>
```

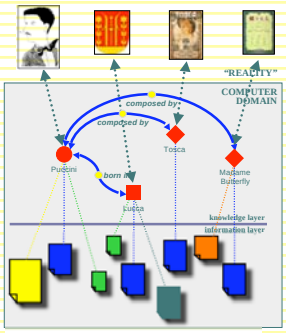

The next best thing since…

- What's the big deal about XML?
- What does XML not do?
- How do XML tags acquire meaning?
- How do standards arise?

What's wrong with the Web?

- It was meant for humans, not machines
- The current Web contains only data, not knowledge
  - From Web of data to Web of knowledge
- Difficult to
  - Aggregate/compare data across sites
  - Delegate complex tasks to "agents"
  - Formulate complex queries involving multiple constraints
  - …

The Semantic Web



Slide from http://www.ontopia.net/

Web 2.0

- Tagging ("folksonomy")
- Blogging
- The "Long Tail"
- Web services
- Wikipedia

Back to the elephant…

- Concepts covered:
  - Metadata
  - XML
  - Semantic Web
- Questions?
- Confused?