CS1001: XML and Applications

Joshua Gordon, Oct 26th 2011
Markup languages

- Markup languages
  - Annotate a document
  - Embed **meta-information** about the structure and content of the document **within** the document
- **HTML**: HyperText Markup Language
  - Tells a web browser how to **display** a document
- **XML**: eXtensible Markup Language
  - Describes the document’s **content**
- How do these relate to programming languages?
Motivation

- You’re an art history student. As part of an undergraduate thesis, you’ve collected thousands of images of historical NYC.
- Goals:
  - Create a website (HTML)
  - Share your photographs (XML)
Pennsylvania Station, 1911 and Today.
106th St and Columbus in 1940 and Today
Old city hall station, 1904-1945
Putting data on the web

- We have
  - Images
  - Descriptions
  - Metadata
- Option 1
  - Create a website using just HTML
Beautiful, but reusable?

```
<html><head>
<title>Columbia University in the City of New York</title>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
<meta name="keywords" content="college, university, new york, graduate, business, law, medical, medicine, mba, public affairs, public health, social work, engineering, science">
<meta name="robots" content="noarchive">
<meta http-equiv="expires" content="0">
<link rel="stylesheet" type="text/css" href="css/default.css">
<style type="text/css" media="all">@import 'css/styles.css';</style>
</head>
<body bgcolor="#ffffff" marginwidth="0" marginheight="0" leftmargin="0" topmargin="0" text="black">
<a name="top"></a>
<center>
table cellpaddings="0" cellspacing="0" border="0" width="760">
<tr bgcolor="#d2a622">
<td bgcolor="#cccccc" width="1"><img src="http://www.columbia.edu/images/spacer.gif" td width="157"></td>
<td class="navlink__base_target" top align="right" bgcolor="#d2a622" width="596">
  class="linkReader__base_target" top href="http://www.columbia.edu/help/index.html" class="linkReader__base_target" top><a href="http://www.columbia.edu/help/index.html" class="linkReader__base_target" top>Help</a><br/>
</td>
</tr>
</table>
</center>
```

The New York Times

N.Y. / Region

1940s-1950s

1946: A view north from 103rd Street, with the supports of the old Ninth Avenue Elevated line remaining at the path to go underground continued.

Getting more from your data

- You’ve invested time and energy
  - Researching our topic; visiting historical societies, scanning in images

- You want an easy way to share your work!
  - Friend: “I’ve have a site with historical maps of New York. Can I link my maps to your images?”
  - You: “Sure.”
  - Friend: “How?”
Uh oh…
Separate content from display

- **We have**
  - Images
  - Descriptions
  - Metadata
- **Option 1**
  - Create a website using HTML
- **Option 2**
  - Instead of directly embedding our images on a webpage, first describe them using XML
<table>
<thead>
<tr>
<th><strong>XML</strong></th>
<th><strong>HTML</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

```xml
<image>
  <address>Penn Station</address>
  <date>1911</date>
  <description>An image of ...</description>
  <file>penn_station_1911.jpg</file>
</image>
```

```html
<html>
  <head>
    <title>Historical NY</title>
  </head>
  <body>
    Check out these images of old NY! The first is of Penn station in 1911
    <img src="penn_station_1911.jpg"/>
    The second is of 106 and 9th ave in 1940
    <img src="106_street_1940.jpg"/>
  </body>
</html>
```
A consistent way of describing data

<person>
  <name>Lisa</name>
  <occupation>student</occupation>
  <favorite food>broccoli</favorite food>
</person>

<person>
  <name>Homer</name>
  <occupation>nuclear safety technician</occupation>
  <favorite food>donuts</favorite food>
</person>
XML

- XML
  - Describes data so it’s readable by people and machines
  - Self-descriptive and (usually) easy to interpret
  - Used to share data on the web
  - Extensible

- XML is a meta-language (a markup language for markup languages)
  - XML uses tags to markup text
  - XML is used to define tags to markup text
XML is not a replacement for HTML

• HTML Tags are standardized
  • <b>Bold tags</b> are <i>understood</i> by <u>web browsers</u>
  • The set of legal tags is fixed
  • You can invent your own, but a web-browser won’t understand them

• XML has a standardized syntax, but no predefined tags!
  • You invent them when you write XML
  • The meaning of the tags is up to the author
This photograph was taken in 1937. City Hall Station opened along with the rest of the Interborough's first subway line on October 27, 1904. It was immediately clear that expansion of the subway system would be necessary and additional lines were built. But ever-increasing ridership eventually required the Interborough's five-car local stations to be lengthened to accommodate longer trains.
Describing data

- What types of information do we want to encode?

<image>

</image>
Describing data

- What types of information do we want to encode?
  
  <image>
  
  <title>City Hall Station</title>
  
  <date>1937</date>
  
  <gps coordinates>40.80, -73.95</gps coordinates>
  
  <description>City hall station prior to the expansion of the subway.</description>
  
  </image>
XML Syntax in a nutshell

- Tags annotate the text between them
  - In HTML tags are used to tell a web browser how to display data
  - In XML, tags indicate to a **person or program** that the content between the tags represents a **specific kind of data**
- Like HTML, tags can be opened, closed, and nested
- Which of these is syntactically correct?
  - `<b>some text</b>`
  - `<b>some text</b>`
  - `<b>>some text</b>`
  - `<b some text</b>`
Sharing information

- Should we all define our own tags?
  - What if we used different names for fields?
- Schemas
  - Describes the structure of an XML document
  - Legal building blocks like the names of tags and their data types
- Namespaces
  - An XML document may contain element or attribute names from more than one vocabulary
  - How can we uniquely specify ours?
Creating Schemas

```xml
<schema xmlns="http://www.cs.columbia.edu/note.xsd">
  <element name="note">
    <complexType>
      <sequence>
        <element name="to" type="string"/>
        <element name="from" type="string"/>
        <element name="priority" type="integer"/>
        <element name="body" type="string"/>
      </sequence>
    </complexType>
  </element>
</schema>
```
Using Schemas

<note schemaLocation="http://www.cs.columbia.edu/note.xsd">
  <to>Bart</to>
  <from>Homer</from>
  <priority>1</priority>
  <body>Find donuts</body>
</note>
Structured vs. Unstructured Data

- Most data on the web is unstructured
  - What language is it in?
  - Can programs work with unstructured data?
- Goal: describe data in the world so that programs can operate on it in useful ways
  - XML defines a syntax for structuring data on the web
  - How we interpret the tags defines the semantics
Natural language processing

NLP research at Columbia

Natural Language Processing research at Columbia University is conducted in the Computer Science Department, the Center for Computational Learning Systems and the Biomedical Informatics Department. Due to the broad expertise and wide ranging interests of our NLP researchers, NLP@CU has a distinctive combination of depth and breadth. Our research combines linguistic insights into the phenomena of interest with rigorous, cutting edge methods in machine learning and other computational approaches.

The NLP@CU group consists of seven senior researchers who serve as principal investigators on multiple projects, and who advise graduate students at all levels. In any one year, the number of Ph.D. students ranges from approximately fifteen to twenty-five, and the number of masters students is much larger. Other members include post doctoral researchers, visitors, undergraduate research assistants, and consultants.
Transmitting data

- In addition to sharing data, XML is also used to transmit data
  - If you know the structure of information, that information is easy to interpret programmatically
<channel>
  <title>Yahoo! Weather - Sunnyvale, CA</title>
  <description>Yahoo! Weather for Sunnyvale, CA</description>
  <language>en-us</language>
  <yweather:location city="Sunnyvale" region="CA" country="United States" />
  <yweather:units temperature="F" distance="mi" pressure="in" speed="mph" />
  <yweather:wind chill="50" direction="0" speed="0" />
  <yweather:atmosphere humidity="94" visibility="3" pressure="30.27" rising="1" />
  <yweather:astronomy sunrise="7:17 am" sunset="4:52 pm" />
  <image>
    <title>Yahoo! Weather</title>
    <width>142</width>
    <height>18</height>
    <link>http://weather.yahoo.com</link>
    <url>http://l.yimg.com/a/i/us/nws/th/main_142b.gif</url>
  </image>
  <item>
    <title>Conditions for Sunnyvale, CA at 9:38 am PST</title>
    <geo:lat>37.37</geo:lat>
    <geo:long>-122.04</geo:long>
    <pubDate>Fri, 18 Dec 2009 9:38 am PST</pubDate>
    <yweather:condition text="Mostly Cloudy" code="28" temp="50" date="Fri, 18 Dec 2009 9:38 am PST" />
    <yweather:forecast day="Fri" date="18 Dec 2009" low="49" high="62"
text="Partly Cloudy" code="30" />
    <yweather:forecast day="Sat" date="19 Dec 2009" low="49" high="65"
text="Partly Cloudy" code="30" />
  </item>
</channel>
Google and Flicker link millions of images automatically using XML
Digital Humanities Projects

• Cultural heritage sites have huge amounts of information
• 3D objects are placed into the scene on the spot in order to overlay the real scene. Like missing paintings, statues or architecture models.
• Transporting a place's original ambiance by virtually placing spatial audio clips in the surroundings.

http://itacitus.org/
Takeaways

• XML helps us separate data from display
  • Embedding information directly in a webpage is brittle
• XML simplifies data sharing and transport
  • Structured data is reusable
  • XML helps you create formats to publish structured information
• XML is platform independent
  • Data is encoded in plaintext using open standards
  • There are others way to structure data (spreadsheets, database) – but many databases use proprietary formats
Designing schemas is hard

- Data representation is an open problem in CS
  - Particularly in medical informatics, Natural Language Processing, A.I.
- Let’s get a flavor for the problem
  - Lavash Pizza With Tomatoes, Mozzarella and Goat Cheese
More complex examples

**Ingredients**
- 1 9-by-12-inch piece of lavash
- 1/2 cup, tightly packed (2 ounces), grated or shredded fresh mozzarella
- 2 plum tomatoes (8 to 10 ounces), sliced
- 1/4 red onion, sliced (optional)
- 1/2 cup (2 ounces) crumbled goat cheese
- 1 to 2 teaspoons fresh thyme
- Salt and freshly ground pepper
- 1 tablespoon extra virgin olive oil

**Instructions**
- Heat the oven to 375 degrees, preferably with a pizza stone in it.
- Place the lavash on a baking sheet.
- Sprinkle the mozzarella over the lavash, and top with the tomato slices and onion.
- Arrange the goat cheese over and between the tomato slices.
- Sprinkle with thyme, salt and pepper.
- Drizzle on the olive oil, and place in the oven.
- Bake 15 to 20 minutes, until the lavash is crisp and dark brown on the edges.