XML – a data sharing standard

Data-sharing challenge

- Companies need to “talk” across different locations/systems and software
- Most database systems (and other software applications, e.g., Excel) can read text files formatted as XML.

Blind Men and Elephants

The ol’ standby: tab (or comma) delimited data...

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? were data collected?
How: were data collected? processed? verified?

… can be pretty useless!

Metadata

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

Early Example of Metadata
Is HTML a good way to share data?

XML: Tags provide meaning (metadata)

XML vs. HTML

Another example: iTunes Library is XML

Acct./Fin XML example

HTML vs. XML

- XBRL (eXtensible Business Reporting Language) is a language for the electronic communication of business and financial data.
- For example, company net profit has its own unique tag.
- Edgar Online (SEC company information) uses XBRL.
- The solution will allow users to request XBRL formatted data for all US equities from within Microsoft Excel, custom templates, and the web.

- HTML started with very few tags, but...
  - HTML now has many tags, and more keep being added over time
  - Messy, yet not customizable

- XML has very few standard tags
  - You add custom tags that are particular to your data needs
XML helps solve the “Is this an elephant?” problem

The following code is legal in HTML:

```html
<p>This is a paragraph</p>
<p>This is another paragraph</p>
```

In XML all elements must have a closing tag like this:

```xml
<p>This is a paragraph</p>
<p>This is another paragraph</p>
```

Opening and closing tags must have the same case:

```xml
<message>This is correct</message>
<Message>This is incorrect</Message>
```

More strict tagging rules than HTML

Visualizing XML data: a “Tree”

In HTML, improperly nested tags like the following are frowned upon, but will not cause an error:

```html
<b><i>This text is bold and italic</i></b>
```

In XML all elements **must** be properly nested within each other like this:

```xml
<b><i>This text is bold and italic</i></b>
```

Database data vs. XML Data

XML data can usually fit a DBMS table...

- Consider the case of: **missing attributes**

  ```xml
  <person id="055">  
    <name>Mary</name>  
    <address>  
      <street>Maple</street>  
      <no>345</no>  
      <city>Seattle</city>  
    </address>  
  </person>
```

- An acceptable fit with table

  Even though blanks are deemed a bit undesirable in database tables

<table>
<thead>
<tr>
<th>name</th>
<th>phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>1234</td>
</tr>
<tr>
<td>Joe</td>
<td>-</td>
</tr>
</tbody>
</table>
...but the fit is not always good

- Problematic case: **Repeated attributes**

  ```xml
  <person>
    <name>Mary</name>
    <phone>2345</phone>
    <phone>3456</phone>
  </person>
  ```

  Two phones!

- A poor fit with table, because database-design rule is that you should not have two columns with the same name:

<table>
<thead>
<tr>
<th>name</th>
<th>phone</th>
<th>phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td>2345</td>
<td>3456</td>
</tr>
</tbody>
</table>

  If XML data had distinct "home" and "cell" columns, the issue goes away.

---

**Formatting (visualizing) XML data**

- Formatting (visualizing in a web page) XML data requires a separate file.

  - What is that separate file? It’s called a “style sheet”

- There are a couple different standards for style sheets:
  - XML-specific standard: XSL
  - Flexible standard: CSS
    - CSS can be used with both .xml and .html files
    - We will look at CSS after XML