Markup data with XML

How to mark up the structural parts of a document with XML

The task

- How television listings might be stored in XML.

Examining the Data

- The first step in developing an XML vocabulary for any domain of interest is to identify the relevant categories.
- In this case the obvious place to look is *TV Guide* or the television listings in the daily newspaper.

Television schedule

July 3	***7:00pm***	***7:30pm***	***8:00pm***	***8:30pm***
CBS 2	Hollywood Squares Repeat, CC	Entertainment Tonight CC American Juniors remaining contestants; Sex and the City Preview.	The Amazing Race 4 CC. Eight twenty-somethings speed through foreign cultures as quickly as possible in attempt to avoid learning anything.	
FOX 5	The Simpsons TVPG, CC	Seinfeld TVPG, CC	The Hurricane (1999) *** (R) TV14, CC Jailed boxer seeks to be exonerated after imprisonment for murders he did not commit.	
…….	……..	…….	……	

Any successful format must provide:

- Station
- Network
- Channel
- Title
- Date
- Start time
- Length or end time
- Description
- Rating
- Whether or not the show is closed captioned
- Year when a movie was made
- Movie type
- Number of stars

XMLizing the Data

- XML is based on a containment model.
- Each XML element can contain text or other

  XML elements, both of which are called the element's *children*.
- Some XML elements may contain both text and child elements.
- However, there's often more than one way to organize the data, depending on your needs.

Q: Which information is a part of which other information?

- Q: The rating and title belong to the show ?
- A: Yes, then, the rating and title elements should be children of the show element.

Q: Which information is a part of which other information?

1. Does a network contain a show or does a show contain a network?
2. Is the network a characteristic of a show, or is a show part of a network?

- Both approaches are plausible.
- There's no one right answer to this question, though some approaches are likely to work better than others.
- I suggest the 2 approach.

ER-Model ?

- Readers familiar with database theory might recognize XML's model as essentially a hierarchical database, and, consequently, recognize that it shares all the disadvantages (and a few advantages) of that data model. There are times when a table-based relational approach makes more sense. This example certainly looks like one of those times.
- However, XML doesn't follow a relational model.
- On the other hand, it is completely possible to store the actual data in multiple tables in a relational database, and then generate the XML on the fly. This enables one set of data to be presented in multiple formats. Transforming the data with style sheets provides still more possible views of the data.

Start to markup

- The root element:

```
<?xml version="1.0"?>
<schedule>
</schedule>
```

Q/A

- Q:There's any information in previous table that applies to the entire table, rather than individual rows or columns ?
- A: Yes, the date the table describes: July 3

```
<?xml version="1.0"?>
<schedule>
  <date>July 3, 2004</date>
</ schedule >
```

What do you know about a station?

- The network affiliation
- The call letters
- The channel number

  Choosing the most obvious names for each of these elements, the first station looks like this:

```
<station>
 <network>CBS</network>
 <call_letters>WCBS</call_letters>
 <channel>2</channel>
</station>
```

- Later we add the shows as children of the station that broadcasts them.

Could be exceptions

- Not all stations have all these pieces, however.
- For example, independent stations aren't affiliated with a network, and cable-only channels don't have call letters.
- You can include those that apply and leave out those that don't.
- For example, here are station elements for WLNY, an independent channel, and HBO, a cable-only network with no local affiliates

Stations

```
<station>
 <call_letters>WLNY</call_letters>
 <channel>55</channel>
</station>
<station>
 <network>HBO</network>
 <channel>501</channel>
</station>
```

Easy to use

- XML makes it very easy to include the information that applies and leave out the information that doesn't.
- There aren't any special null values, or elements used just to fill an expected slot.

The Stations in the Schedule

```
<?xml version=”1.0”?>
<schedule>
  <date>July 3, 2003</date>
  <station>
    <network>CBS</network>
    <call_letters>WCBS</call_letters>
    <channel>2</channel>
  </station>
  <station>
    <call_letters>WLNY</call_letters>
    <channel>55</channel>
  </station>
  <station>
    <network>HBO</network>
    <channel>501</channel>
  </station>
</schedule>
```

Markup shows

- The key component of the information will be the individual shows.

Show: Hollywood Squares

- After examining it, you know the following:
  - The name of the show: *Hollywood Squares*.
  - The start time: 7:00 P.M.
  - The end time: 7:30 P.M.
  - The length of the show: 30 minutes.
  - The channel: 2.
  - The network: CBS.
  - The air date: July 3, 2004.
  - The show is closed captioned.
  - The show is a repeat.

More about shows

- The channel and network will become part of each station element.
- There's no need to duplicate them.
- The remainder of these items can each be made a child element of a show element like:

Hollywood Squares

```
<show>
 <name>Hollywood Squares</name>
 <start_time>7:00 p.m.</start_time>
 <end_time>7:00 p.m.</end_time>
 <length>30 minutes</length>
 <air_date>July 3, 2003</air_date>
<closed_captioned>yes</closed_captioned>
 <repeat>yes</repeat>
</show>
```

Observations

- The start and end times are in the eastern time zone. These are the correct local times for New York.
- Indicate the time zone in which these times are stated, typically by giving the offset from GMT and often using a 24-hour clock.
- For example, New York is five hours behind GMT, so the start time could be written as 19:00-0500.
- One of the three numbers—start time, end time, and length—is redundant. Given two of these it's possible to calculate the other. It might be wiser not to include all three.
- The air date at least seems redundant with date of the entire schedule. However, most television listings prefer to start the day somewhere around 5:00 or 6:00 A.M., rather than at midnight. Thus, it's not uncommon for a show that's broadcast in the early morning one day to appear in the schedule for the previous day.

An improved show

```
<show>
 <name>Hollywood Squares</name>
 <start_time>19:00-0500</start_time>
 <length>30 minutes</length>
 <air_date>July 3, 2003</air_date>
 <closed_captioned>yes</closed_captioned>
 <repeat>yes</repeat>
</show>
```

More information is available

- The cast: Don Rickles, Jerry Springer, Richard Simmons, Vicki Lawrence, John Salley, Joanie Laurer, Martin Mull, Jillian Barberie The Producers: Henry Winkler, Michael Levitt

- The original air date: January 16, 2003

Some improvements

```
<show>
 <name>Hollywood Squares</name>
 <type>Series/Game Shows</type>
 <episode_number>5074</episode_number>
 <start_time>19:00-0500</start_time>
 <length>30 minutes</length>
 <air_date>July 3, 2003</air_date>
 <original_air_date>January 16,2003</original_air_date>
 <closed_captioned>yes</closed_captioned>
 <repeat>yes</repeat>
 <cast> Don Rickles, Jerry Springer, Richard
   Simmons, Vicki Lawrence, John Salley, Joanie
   Laurer,Martin Mull, Jillian Barberie</cast>
 <producer>Henry Winkler</producer>
 <producer>Michael Levitt</producer>
</show>
```

What's wrong with cast element ?

- It has significant substructure that is not yet reflected in the XML markup.

- The cast is composed of individual actors.

A cast element

```
<cast>
 <actor>Don Rickles</actor>
 <actor>Jerry Springer</actor>
 <actor>Richard Simmons</actor>
 <actor>Vicki Lawrence</actor>
 <actor>John Salley</actor>
 <actor>Joanie Laurer</actor>
 <actor>Martin Mull</actor>
 <actor>Jillian Barberie</actor>
</cast>
```

Improving actor and producer elements

```
<cast>
 <actor>
  <given_name>Don</given_name>
  <surname>Rickles</surname>
 </actor>
……
 <producer>
  <given_name>Henry</given_name>
  <surname>Winkler</surname>
 </producer>
……
</cast>
```

Observations:

- The tags `<given_name>` and `<surname>` are preferable to the more obvious `<first_name>` and `<last_name>` or `<first_name>` and `<family_name>`.
- Whether the family name or the given name comes first or last varies from culture to culture. Furthermore, surnames aren't necessarily family names in all cultures.

Look at multiple examples

```
<show>
 <name>Entertainment Tonight</name>
 <type>Series/News</type>
 <episode_number>5689</episode_number>
 <start_time>17:30-0500</start_time>
 <length>30 minutes</length>
 <air_date>July 3, 2004</air_date>
 <original_air_date>July 3, 2004</original_air_date>
 <closed_captioned>yes</closed_captioned>
 <repeat>no</repeat>
 <description>American Juniors remaining contestants;
  Sex and the City preview.
 </description>
</show>
```

Q/A

- Q: Has description element an identifiable substructure ?
- A: Yes, you could mark it up as two separate segments, each of which also contains a series element:

```
<description>
  <segment>
   <series>American Juniors</series> remaining
  contestants
  </segment>
  <segment>
   <series>Sex and the City</series> preview
  </segment>
</description>
```

A movie example

```
<show>
 <name>Final Fantasy: The spirits Within</name>
 <type>movie/animated</type>
 <start_time>18:30-0500</start_time>
 <length>105 minutes</length>
 <air_date>July 3, 2004</air_date>
 <closed_captioned>yes</closed_captioned>
 <repeat>yes</repeat>
 <rating>PG-13</rating>
 <stars>3</stars>
 <description> The last city on earth  defends…
  </description>
 <director>
  <given_name>Hironobu</given_name>
  <surname>Sakaguchi</surname>
 </director>
```

A movie example (cont'd)

```
<writer>
  <given_name>Al</given_name>
  <surname>Reinart</surname>
 </writer>
 <producer>
  <given_name>Hironobu</given_name>
  <surname>Sakaguchi</surname>
 </producer>
   ……
 <cast>
  <actor>
   <given_name>Ming</given_name>
   <surname>Na</surname>
  </actor>
   ……
 </cast>
</show>
```

tvschedule2004-07-03.xml

```
<?xml version=”1.0”?>
<schedule>
 <date>July 3, 2004</date>
 <station>
  <network>CBS</network>
  <call_letters>WCBS</call_letters>
  <channel>2</channel>
  <show>
   <name>Hollywood Squares</name>
   <type>Series/Game Shows</type>
   <episode_number>5074</episode_number>
   <start_time>19:00-0500</start_time>
   <length>30 minutes</length>
   <air_date>July 3, 2004</air_date>
   <original_air_date>January 16, 2003</original_air_date>
   <closed_captioned>yes</closed_captioned>
   <repeat>yes</repeat>
   <cast>
```

tvschedule2004-07-03.xml (cont'd)

```
    <actor>
      <given_name>Don</given_name>
      <surname>Rickles</surname>
    </actor>
    ……
  </cast>
  <producer>
    <given_name>Henry</given_name>
    <surname>Winkler</surname>
  </producer>
      ……
</show>
<show>
  <name>Entertainment Tonight</name>
  <type>Series/News</type>
  <episode_number>5689</episode_number>
```

tvschedule2004-07-03.xml (cont'd)

```
    <start_time>19:30-0500</start_time>
    <length>30 minutes</length>
    <air_date>July 3, 2004</air_date>
    <original_air_date>July 3, 2004</original_air_date>
    <closed_captioned>yes</closed_captioned>
    <repeat>no</repeat>
    <description> American Juniors remaining contestants;
      Sex and the City preview.
    </description>
  </show>
  <show>
    ……
  </show>
 </station>
……
</schedule>
```

Observations

- Even as large as it is, this document is incomplete. It contains only a couple of hour's worth of shows from three networks. Showing more than that would make the example too long…
- If you continued to look at more shows, you would discover numerous other relevant pieces of information that deserve to be marked up, including role played by an actor, broadcast language, pay-per-view prices, and more.
- However, I will stop the XMLization of the data here

Conclusion: **The Advantages of the XML Format**

- The data is self-describing.
- The data can be manipulated with standard tools.
- The data can be viewed with standard tools.
- Different views of the same data are easy to create with style sheets.

TO DO LIST

- Consult web page for exercises.
- Consult CLIX (soon, I think).
- I respect the design principles ?
- Why I abandon the series markup ?
- Please submit exercises.

A very good reading