XML BASICS

What is XML?

XML (*Extensible Markup Language*) is an internationally recognized standard for describing document content for reproduction in electronic form. XML is made up of a set of rules - called a *markup language* – that are used together for encoding text. XML can support publishing in many media, including conventional paper based, online multi-media, electronic text, and CD-ROM.

Benefits of XML

- **More meaningful searches** - XML has *Extensibility,* which means unlimited search options and more powerful searches. A search for a book can yield results by author, title, ISBN number or other criteria.
- **Development of flexible Web applications** - Once the data is found, XML can be delivered to other applications, objects, or other servers. XML can combine data that can be edited, changed, or viewed in various ways.
- **Flexibility** - XML has the ability to structure flexible, electronic documents, and to define different parts of a document's contents.

Tags

XML documents are written in plain text with start and end **Tags** that define the content. **Start tags** and **end tags** are both bounded by angle brackets **< >**, and the end tag includes a forward slash mark **/**.

Elements

- Value
- Empty
- Root
- Wrapper

An **Element** is made up of a **start tag**, an **end tag**, and the data in between. The start and end tags describe the data within the tags, which is considered the **value** of the element. For example, the following XML element within the start and end tags is an "editor" with the value "Anna Smith":

<editor>**Anna Smith**</editor>

Each element has a different tag.

<actress>Anna Smith**</actress>**

Empty elements are used as placeholders and action triggers for processing data. They do not need a start and an end tag, only a forward slash **/** just before the end of the empty element's tag:

<graphic/>

Each entire XML document must have a unique starting and ending element, referred to as the **root** element**:**

```
<weather-report >
        <date>March 25, 1998</date>
            <time> 08:00</time>
             <city>Seattle,/city>
 </weather-report>
```

Wrapper elements consist of 1 or more elements:

```
<document>
    <title>Vanity Fair</title>
    <author>Thackeray</author>
 </document>
```

Note: XML tags are case sensitive, so each of the following is a unique element.

<city> <City> <CITY>

Attributes

An element can contain one or more **Attributes.** Attributes are typically used to attach more information to or about an element. An attribute is a name-value pair separated by an equal sign **=,** and bounded by quotation marks.

<CITY ZIP =”20814”>Bethesda</CITY>

Note: Many people use uppercase text for attribute names so they can easily distinguish between the names of elements and the names of attributes.

Hierarchy

Elements are arranged in a parent-child *hierarchy* e.g., “city” is a child of “area”, which is a child of “weather report”.

<weather-report >
<area>
<city>Seattle</city>

Document Type Definition (DTD)

The documentation guidelines or rules are contained in the **Document Type Definition (DTD).** For example, the definition of a report might be that it consists of a title, an author, an abstract, and one or more paragraphs. Any document without a title, according to this definition, could not be considered a report.

- The DTD indicates rules for elements and attributes, and ensures that required elements are entered in the correct order and with the correct data.

Entities

Entities are structural mechanisms that enable XML authors to organize documents. An entity associates an entity name with text. Every time that entity name is used, XML will replace that entity name with a specific piece of text. Entities are noted with an **!ENTITY** declaration at the beginning of a document or DTD.

Format And Structure

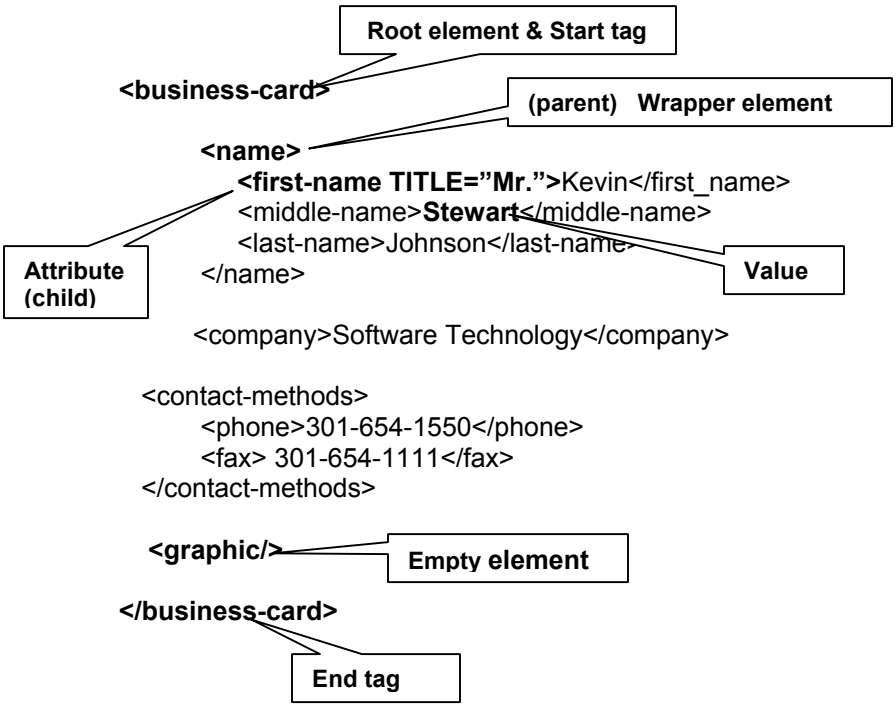
An XML document does not contain any information about how to display itself. A formatting technology is used to display documents in a Web browser or application.

Style sheets are optional, and are attached only if the document needs to be formatted for end users.

- **Format** - How a document looks (font, size)
- **Structure** - Defines the elements that make up the document, and the relationship those elements have to each other.

XML in action

A simple business card, with the components highlighted and named, illustrates the key technical features of XML:



DTD in action

DTD's describe the allowable structure of XML documents. Although you will not have to create DTD's (they are developed for ATLAS by the data team), it is helpful to understand their structure and relationship to XML.

<!ELEMENT business_card (name, title, contact_methods)> is a wrapper element and contains the following elements

<!ELEMENT name (first_name, middle_name?, last_name)>
<!ELEMENT given _name (#PCDATA)>
<!ELEMENT middle_name (#PCDATA)>
<!ELEMENT family_name (#PCDATA)>

<!ELEMENT company (#PCDATA)>

<!ELEMENT contact_methods (phone*, fax)>
<!ELEMENT phone (#PCDATA)>
<!ELEMENT fax (#PCDATA)>

- **#PCDATA** stands for Parsed Character Data, which means text
- **?** means 0 or 1 times. For example: **<!ELEMENT** name (first_name, middle_name?, last_name)> indicates that the person must have a first name, and a last name, but a middle name is not required
- ***** means 0 or more times - So, **<!ELEMENT** contact_methods (phone*, fax) indicates that contact methods may have any number of phone numbers or none.