Tutorial – 2
XML and Semantic Web Technologies – SoSe 2012

Instructor: Prof. Dr. Dr. Lars Schmidt-Thieme
Tutor: Umer Khan 27-04-2012

Q-1.

Given is the following DTD:

```
<!ELEMENT movies (Movie+) >
<!ELEMENT Movie
( title, year, _director, (comment | newcomment)+)>
<!ATTLIST Movie id ID #REQUIRED>
<!ELEMENT title (#PCDATA)>
<!ELEMENT year (#PCDATA) >
<!ELEMENT _director (#PCDATA)>
<!ATTLIST _director name CDATA #IMPLIED>
<!ELEMENT comment (#PCDATA)>
<!ELEMENT newcomment (#PCDATA)>
<!ATTLIST comment lang CDATA #IMPLIED>
```

1.1 Correct the following documents w.r.t to above given DTD.

- title and xml are not allowed as children of movies
- the id may not begin with a digit
- comment may not have an attribute

a)

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE movies [
<!ENTITY copy "&#169;">
]>
<movies>
    <movie id="y56225">
        <title>Mamma Mia</title>
         <year>2008</year>
        <_director name=" Phyllida Lloyd"/>
        <comment text="Musical Romantic Comedy"/>
        <newcomment text="An &lt;important&gt;text">Oscar</newcomment>
        <comment lang="en">&copy; 2008 Universal Pictures</comment>
        <!-- Famous for ABBA music -->
    </movie>
</movies>
```

b)

```
<?xml version="1.0" encoding="utf-16"?>
<movies>
<movie id=”y56225”>
<title>Love Story</title>
<!-- <title></title> -->
<year>1980</year>
<_director name='Coppola'></_director>
<comment text=”Five start” text=”Average”/>
<-- <xml>Introduce XML content</xml> -->
<newcomment text="An <important> text">Oscar</newcomment>
<comment lang=de>&copy; 1980 Warner Bros.</comment>
<!-- Famous movie of the 80s -->
</Movie>
</movies>
```

1.2. What is the difference between CDATA and PCDATA ?

PCDATA means "parsed character data". It means that this character data is to be parsed. In particular:
- Entity references (< > ' " and &) will be resolved (to < > ' " and & respectively), as well as any additional entities defined in the DTD.
- It may not contain any unencoded < or & characters, because they would be confused with an opening tag or an entity reference.

In the DTD, PCDATA is used to say that an element may only contain parsed character data, without child elements.

CDATA is more confusing, because it can have several meanings in the XML world:

- In the DTD, it is used to give the most general type for an attribute: an attribute of type CDATA may contain any attribute value. Note, however, that entity references are resolved, and that & and < must be encoded as well. In addition, the single quote ' must be encoded to ' if the attribute value is single-quoted, and the double quote " must be encoded to " if the attribute value of double-quoted.

- Where PCDATA is expected in an element, one can explicitely use a CDATA construct to escape the special XML characters like < or >, which will not be recognized as markup. The only sequence recognized as markup in a CDATA section is]]>, which is interpreted as the end of the CDATA section.

```
<![CDATA[
if (a<2) { // notice the use of < without needing to encode it as &lt;
Writeline("The number is too low");
}
]]>
```

1.3. Provide reasons for using elements instead of attributes.

This is a big debate between programmers.
- Some say that attributes are for metadata whereas elements contain information
- In general, an element is better if there is a lot of data inside
- One has to use an element if one wants to nest children
- Attributes are in a set (so two attributes of an element may not have the same name), whereas two sibling elements may have the same name.

Q-2. Create an XML document to store bank account and customer data with following information:

```
<?xml version="1.0" encoding="UTF-8"?>

<bank>
   <accounts>
    <savings_accounts>
      <savings_account id="a1" interest="0.03">
        <balance>2500</balance>
      </savings_account>

      <savings_account id="a2" interest="0.03">
        <balance>15075</balance>
      </savings_account>
    </savings_accounts>

    <checking_accounts>
      <checking_account id="a3">
        <balance>4025</balance>
      </checking_account>
      <checking_account id="a4">
        <balance>-125</balance>
      </checking_account>
      <checking_account id="a5">
        <balance>325</balance>
      </checking_account>
    </checking_accounts>
  </accounts>

  <customers>
    <customer id="c1">
      <name>Ben Richerdson</name>
      <address>Park Drive 2</address>
    </customer>
    <customer id="c2">
      <name>Marc Wretcher</name>
      <address>Mill Drive 75</address>
    </customer>
    <customer id="c3">
```

```
      <name>Angel Steady</name>
      <address>Lake Sight 15</address>
    </customer>
  </customers>

  <customer_accounts>
    <customer_account c_id="c1" ac_id="a2"/>
    <customer_account c_id="c1" ac_id="a3"/>
    <customer_account c_id="c2" ac_id="a4"/>
    <customer_account c_id="c3" ac_id="a1"/>
    <customer_account c_id="c3" ac_id="a5"/>
  </customer_accounts>

</bank>
```

Q-3. Use an external DTD for the attached courses-ID xml document.

- Use Oxygen xml editor to create and edit the XML document
- Save the file in the same directory as the XML document
- Reference the DTD in the document prolog of the XML document
- Check that the XML document is valid according to the DTD

```
<!ELEMENT Course ( Title, Description? ) >
<!ATTLIST Course Enrollment NMTOKEN #IMPLIED >
<!ATTLIST Course Instructors CDATA #REQUIRED >
<!ATTLIST Course Number ID #REQUIRED >
<!ATTLIST Course Prerequisites CDATA #IMPLIED >
<!ELEMENT Course_Catalog ( Department+ ) >
<!ELEMENT Courseref EMPTY >
<!ATTLIST Courseref Number NMTOKEN #REQUIRED >
<!ELEMENT Department ( Course | Lecturer | Professor | Title )* >
<!ATTLIST Department Chair NMTOKEN #REQUIRED >
<!ATTLIST Department Code NMTOKEN #REQUIRED >
<!ELEMENT Description ( #PCDATA | Courseref )* >
<!ELEMENT First_Name ( #PCDATA ) >
<!ELEMENT Last_Name ( #PCDATA ) >
<!ELEMENT Lecturer ( First_Name | Last_Name | Middle_Initial )* >
<!ATTLIST Lecturer InstrID NMTOKEN #REQUIRED >
<!ELEMENT Middle_Initial ( #PCDATA ) >
<!ELEMENT Professor ( First_Name | Last_Name | Middle_Initial )* >
<!ATTLIST Professor InstrID ID #REQUIRED >
<!ELEMENT Title ( #PCDATA ) >
```