Universität Freiburg
Institut für Informatik
Prof. Dr. Peter Fischer
Lecuture XML and Databases
Winter Semester 2011/12

04.11.2011

Solution Sheet 1

XML and DTD validation

Exercise 1: One document, several XML representations

1.1.

a) Your document should look like the exercise.

b) Here is one possibility for doc1.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<document>
    <title>Exercise 1: One document, several XML representations</title>
    <section nr="1">Perform the following tasks
        <subsection>Create a sample document in Microsoft Word containing
this exercise with about the same formatting, and save it as
<b>sample.doc</b>.</subsection>
        <subsection>Copy the contents (as raw text) of <b>sample.doc</b> in
oXygen and specify its structure using XML tags. (Title, sections). Also
mark specific styles (such as italic) inside the text. Save it as
<b>doc1.xml</b>.</subsection>
        <subsection>Open <b>sample.doc</b> in Word and save it as XML (Save
as... XML Document (*.xml)) with the name <b>doc2.xml</b>.</subsection>
        <subsection>Open <b>sample.doc</b> in OpenOffice and save it in the
OpenOffice format. Change the extension of the file from .write to .zip.
Extract from the zip file the <b>content.xml</b> file and rename it to
<b>doc3.xml</b>.</subsection>
    </section>
    <section nr="2">
        Open <b>doc1.xml</b>, <b>doc2.xml</b> and <b>doc3.xml</b> in
oXygen. Check that they are well-formed (oXygen tells you). Which format is
better? Microsoft Office, OpenOffice, yours?
    </section>
    <section nr="3">
        Is this data structured, unstructured or semi-structured?
    </section>
</document>
```

c) Do it!
d) Do it!

1.2. No format is better: it depends on the needs of the application.

1.3. This data is semi-structured: there is some structure, but not all of the content is structured as in a flat database.

Exercise 2: Well-Formed XML

2.1. This document has the following problems:
- the quotes in XML must always be simple quotes or double quotes, but not "Word-style" quotes (‘ or ”...)
- the movie start tag does not correspond to the Movie end tag
- The entity &copy; is not defined in XML. Some XML-based languages define it as the caracter © though. You have to define it explicitely
- You cannot have the < sign inside attributes. Use &lt; instead (defined by XML). Also it is advised to use &gt; for the > symbol.
- The first comment element has two attributes named text, this is forbidden.
- Comments <!-- --> cannot include the characters --
- The lang attribute should be quoted.
- XML names beginning with xml are reserved by the W3C. Their usage should be avoided (except if it is as specified as the W3C, e.g. xml:space, xml:lang, xmlns...).

Here is the corrected document:

```
<?xml version="1.0" encoding="utf-16"?>
<!DOCTYPE movies [
<!ENTITY copy "&#169;">
]>
<movies>
    <Movie id="56225">
        <title>Love Story</title>
        <title></title>
        <year>1980</year>
        <_director name='Coppola'></_director>
        <comment text="Five start"/>
        <comment text="Average"/>
        <newcomment text="An &lt;important&gt;
                          text">Oscar</newcomment>
        <comment lang="de">&copy; 1980 Warner Bros.</comment>
        <!-- Famous movie of the 80s -->
    </Movie>
</movies>
```

2.2. This will be shown correctly in most browsers. However, it is not well-formed XML: the br and p tags are not closed. The following would be well-formed XML:

```
<?xml version="1.0" encoding="UTF-8"?>
<html>
    <head>
        <title>Untitled</title>
    </head>
    <body>
        Dear jane <br/>
        <p>You are invited at the weekly meeting</p>
        <p>Yours sincerely, <br/>
            John</p>
    </body>
</html>
```

But XHTML is more than just XML: it also has to have a certain structure (this is called to be "valid"). Among others, the tags have to live in the XHTML namespace (which is a little bit like a family name), and the text in the body has to be embedded in a div tag:

```
<?xml version="1.0" encoding="UTF-8"?>
<html xmlns="http://www.w3.org/1999/xhtml">
    <head>
        <title>Untitled</title>
    </head>
    <body>
        <div>Dear jane
            <p>You are invited at the weekly meeting</p>
            <p>Yours sincerely, <br/>
                John</p>
        </div>
    </body>
</html>
```

Valid XML documents as well as namespaces will be studied in detail later in this course.