Prof. Dr. Stefan Brass
Institut für Informatik
MLU Halle-Wittenberg

October 19, 2017

XML and Databases
— Exercise Sheet 1 —

Part a) to e) will be solved in class, you only have to submit Part f). But please think about the questions in Part a) before the meeting! Send your solutions for the homework f) to the instructor via EMail: `brass@informatik.uni-halle.de` (with "`xml17`" in the subject line). The official deadline is October 26, 10:00 (before the lecture time).

In this course, no group submissions are permitted.

Not all submitted homeworks will be corrected, but all homework exercises will be discussed in class. If you should have questions about your homework, please ask! A precondition for getting credit for this course is that you submit solutions to two thirds of the homework sheets. Obviously wrong or very incomplete submissions do not count.

In-Class Exercises

a) Answer the following questions (similar questions might be asked in an oral exam):

- Describe the tree structure for the internal representation of XML documents (as far as it was introduced in the course so far).
- What is the relationship between tags and elements?
- Name the three different kinds of tags in XML and describe their syntax.
- Which characters are forbidden inside the `"..."` specification of an attribute value? What can you do if you need one of these characters in the attribute value?
- Name some differences between XML and HTML.
- Explain the syntax of an element declaration in XML and give an example. What constructs can be used in a content specification?
- Explain the restriction that content models must be deterministic. Give an example for a forbidden content model.
- What does "PC" mean in the keyword "`#PCDATA`"? Why is the "`#`" necessary here, but not in "`EMPTY`" (where it is actually forbidden)?
- XML element declarations and syntax rules of a context-free grammar (e.g., BNF), look very similar. What is the main difference of XML with DTDs and context-free grammars with respect to the definable languages?

- How are attributes declared in DTDs? Give an example and explain the most important options, e.g. commonly used data types for attributes.
- How does a comment look like in XML? Why is the syntax so long/complicated?
- How do ISO Latin 1 (ISO 8859-1) and UTF-8 differ in the encoding of German national characters, e.g. "ä" (Unicode code point: 228).
- Define the syntax of character references. Wie sehen Zeichenreferenzen aus? Is the character encoding defined for the XML file (in the XML declaration) relevant for the number in the character reference?

b) The following XML document contains at least 8 syntax errors. Please find and correct as many as possible. The text in the document is in German, but that has no consequences for the syntax errors. [This exercise was developed by Dr. Henning Thielemann.]

```
<Film 6jahre=freigegeben>
    <titel>Herr Fuchs und Frau Elster</titel>
    <regie>Harald Serowski & Erich Hammer<regie/>
    < altersbeschränkung typ = "für sechsjährige" >
    <beschreibung>
    Wenn es eine Personifizierung von <quote>Neugier</quote> gibt,
    dann ist es <b><i>Frau Elster</b></i>.
    In diesem heiteren Streifen für die ganze Familie
    ist sie wieder voll in ihrem Element.
    ...
    </beschreibung>
</film>
```

You can access the document at the following web address:

[http://www.informatik.uni-halle.de/˜brass/xml17/examples/err_film.xml]

After you corrected some mistakes, you can check the document using one of the following methods:

- The checking for well-formedness can be done with any modern browser, e.g. Firefox. Simply open the XML file in the browser. If the browser displays the data of the XML document (typically the tree structure with an expand/collapse-interface for subtrees), the document is well-formed. It is normal that the browser complains that there is no style sheet associated with the XML document.
- [http://www.xmlvalidation.com/]
  Since there is no DTD, this service shows the error that the "`Film`" element is not defined. If it got that far, the document is well-formed.
- [http://www.w3schools.com/xml/xml_validator.asp]

- You could also install the AltovaXML Community Edition on your computer:

  [http://www.softpedia.com/get/Internet/Other-Internet-Related/AltovaXML.shtml]

  The program has only a simple command-line interface, but it contains a validator (for DTDs and XML Schema definitions), and XSLT processor, and an XQuery implementation. For a graphical user interface and other additional features, one has to buy Altova XMLSpy.

- For later exercises with XQuery, I recommend to install the XML-database BaseX on your computer. Of course, when you create a new database from an XML file, this also checks for syntax errors.

  [http://basex.org/]

c) Please look at an XML file in your web browser, e.g.

   [https://www.ibiblio.org/xml/examples/shakespeare/much_ado.xml]

d) Develop an XML file for a simplified course list of the computer science department. The following information is needed for each course:

   - the course title (mandatory),
   - a web address (optional),
   - a course type (one of: lecture, seminar, project),
   - instructors of the course (zero or more) (only the name must be stored),
   - class meeting times (one or more). For each such time, the day of week, start time, end time and room must be stored.

   Please write an XML file with at least two courses, of which one has two class meeting times. Please check your XML file for well-formedness.

e) Develop a DTD for your data file. If you use a web service for validation, it might be necessary to embed the DTD into the data file as follows:

```
<?xml version="1.0"?>
<!DOCTYPE OUTERMOST_ELEMENT [
    <!-- ELEMENT and ATTLIST declarations -->
    ...
]>
<OUTERMOST_ELEMENT>
    <!-- Data -->
    ...
</OUTERMOST_ELEMENT>
```


Homework Exercises

f) Please develop an XML file that contains information about nutrition supplements (multivitamin tablets). It should be possible to represent the data from the following relational database, but you can also use German names, or other data (of approximately the same size). If you need additional IDs, you may introduce them. You do not have to specify a DTD (that will probably be the next homework).

   - `Supplement(Name, Producer)`
     The underlining indicates primary key attributes.

Supplement	
Name	Producer
Centrum	Lederle
One A Day	Bayer

   - `Vitamin(Substance, Daily_Value, Unit)`
     The second column contains the recommended daily value, and the third column the unit used for amounts of that vitamin.

Vitamin		
Substance	Daily_Value	Unit
Vitamin A	5000	IU
Vitamin C	60	mg
Biotin	300	mcg
Iron	18	mg

   - `Contents(Name → Supplement, Substance → Vitamin, Quantity)`
     This table corresponds to a many-to-many-relationship between “Supplement” and “Vitamin” (the arrows indicate foreign keys). “Quantity” is the contents in one tablet, with the same units as in table “Vitamin”.

Contents		
Name	Substance	Quantity
Centrum	Vitamin A	5000
One A Day	Vitamin A	5000
Centrum	Vitamin C	60
One A Day	Vitamin C	60
Centrum	Biotin	30
Centrum	Iron	18
One A Day	Iron	27

Of course, the concrete data listed here are incomplete and probably outdated. I used this example in a database exam in Pittsburgh in 1999.