Klausur "Semistructured Data and XML"
Summer Term 2021
Prof. Dr. Wolfgang May
17. August 2021, 11:00–13:00
Working Time: 120 Minutes
(carried out as a computer-based ILIAS exam)

Vorname:

Nachname:

Matrikelnummer:

Setting: The usage of saxon (with the aliases saxonValid, saxonXQ, and saxonXSL defined as in the course) and xmllint (validation error messages are better with xmllint) is recommended. Web access (e.g. for XPath/XQuery and XSLT documentation) is allowed. It was also recommended to have the slides, and a condensed self-prepared "cheat sheet" (preparation of a cheat sheet is a very effective way to work through the materials).
Answers might be given in English or German (most answers are program code anyway). In the text, the german translation is sometimes given in parentheses.
Give *all* answers via the ILIAS system.
Like in a "paper exam", also solutions that do maybe not work (or do not work completely) can be delivered and will be graded with appropriately partial points.
For **passing** the exam, **50** points are sufficient.

	Max. Punkte	Erreichte Punkte
Aufgabe 1 (XML)	20	
Aufgabe 2 (DTD)	15	
Aufgabe 3 (XPath (a))	2	
Aufgabe 4 (XPath (b))	4	
Aufgabe 5 (XPath/XQuery (c))	6	
Aufgabe 6 (XPath/XQuery (d))	6	
Aufgabe 7 (XPath/XQuery (e))	6	
Aufgabe 8 (XPath/XQuery (f))	9	
Aufgabe 9 (XPath/XQuery (g))	9	
Aufgabe 10 (XSLT)	15	
Aufgabe 11 (Miscellaneous (a))	4	
Aufgabe 12 (Miscellaneous (b))	4	
Summe	100	

Note:

Name: MatNr.:

Project: All Years of Tour de France Database

All exercises are based on a common "project": a database about all "Tour de France" instances. The "Tour de France" is a cycling sports (and touristic) event that takes place every year usually during three weeks in the summer (except in 2020, when it was postponed to September) around France and sometimes also other (more or less) neighboring countries.

1. There is a *tour* instance in every year. As examples, we consider the 2020 and mainly the 2021 tour instances.

2. Every year, several *teams* participate; each with several *riders*.

3. So, for every year, it is stored which riders started for which teams. Riders may change teams from one year to the other.

4. For every rider that ever participated, the name, the birthdate, and the country of birth is stored (so we don't have to consider people changing nationalities).

Sample: In 2021, among others the *Jumbo Visma* team participated with riders *Primoz Roglic* (from *Slovenia*, born October 10th, 1989), *Wout van Aert* (from *Belgium*, born September 15th, 1994), *Sepp Kuss* (from the *USA*, born September 13th, 1994), *Jonas Vingegaard* (from *Denmark*, born December 10th, 1996) and others. Also, the *UAE Team Emirates* participated, with rider *Tadej Pogacar* (from *Slovenia*, born September 21st, 1998) and others. Team *"Trek"* participated with riders *Bauke Mollema* (from the *Netherlands*, born November 26th, 1986), *Kenny Elissonde* (from *France*, born July 22nd, 1991) and others.

In 2020, *Jumbo Visma* participated also with *Roglic*, *van Aert*, *Kuss* and some others. UAE also with *Pogacar* and *Vingegaard* (the latter was not true in reality, but serves here as an example for a rider who changed the team).

5. In every year, the tour consists of a sequence of *stages*; for each stage, the date is stored. Assume that there is at most one stage per day. Every stage is assigned a type: flat, hilly, mountains.

6. Every stage leads from one place to another. Places can be towns/cities, or mountains/passes. A subsequent stage does not necessarily start where the previous stage ended (but usually in the same region).

7. Places have a name and an elevation, and are located in a country (usually, but not always France).

8. For every stage, the starting place and the destination place, and optionally, a sequence of intermediate mountains/passes and places is stored.

Sample: In 2021, the first stage on June 26th led from *Brest* to *Landerneau*, it was a *hilly* stage of 198km. The 11th stage in 2021 on July 7th over 199km started in *Sorgues*, to crossed the *Mont Ventoux*, down to *Malaucene*, then to *Bedoin*, crossed the *Mont*

Ventoux again, and finished in *Malaucene*. The 15th stage on July 11th lead over 191km through mountains from *Ceret* to the town *Andorra* (which is in the country *Andorra*); it crossed the *Port d'Envalira* and the *Col de Beixalis* (both also in the country *Andorra*). The (last) 21st stage on July 18th lead flatly over 108km from *Chatou* to *Paris*.
In 2020, the first stage was a 156km hilly round trip from *Nizza* to *Nizza*, and the last, 21st stage was 122km flat from *Mantes* to *Paris*.

Brest, *Landerneau*, and *Nizza* are at an elevation of 10m, *Sorgues* is at 20m, *Chatou*, *Mantes*, and *Paris* are at 30m, *Ceret* is at 120m, *Malaucene* and *Bedoin* are at 300m, *Andorra* has 1011m, the *Mont Ventoux* has 1909m, the *Port d'Envalira* has 2407m, and the *Col de Beixalis* has 1795m.

9. For every stage, the result is stored: the order in which the riders *arrived* at the finish line together with the duration each of them needed for the stage.
Note that it is possible that riders abandon the tour at some day, then they will not be listed in remaining stages.

Sample: The above 11th (*Mont-Ventoux-*)stage 2021 was won by *Wout van Aert* in 5:17:43h, then followed *Elissonde* and *Mollema* (both 5:18:57), *Pogacar* and *Vingegaard* (5:19:21) (and later the rest).

The above 15th stage (to Andorra) was won by *Sepp Kuss* in 5:12:06h, and the others arrived later. For this example, it may be useful also to store that *Elissonde* arrived at 5:16:17h.

The 21st stage was also won by *Wout van Aert*, in 2:09:37h; Pogacar, Vingegaard and the others crossed the line with the same time directly after him.

Exercise 1 (XML [20 Points])
Design an XML structure (use the frame given in file `exam.xml`) and fill it with some sample data (e.g. with some of the example data given in the text).
(Information about handling dates and times in XML can be found in the course's slides around Slide 300.)
Copy-and-paste the XML from the file `exam.xml` afterwards (at the end of the exam, because it will be extended in later exercises) here:

Exercise 2 (DTD [15 Points])
Develop the DTD for your document developed in Exercise 1, use the file `exam.dtd`.
Use one of the calls

```
xmllint -loaddtd -valid --noblanks -noout exam.xml
saxonValid.bat -s:exam.xml
```

for validating it (note that xmllint provides better error messages).
Copy-and-paste the DTD from the file `exam.dtd` afterwards here:

Exercise 3 (XPath (a) [2 Points])
Use your `exam.xml` XML file as a basis for solving this and the following exercises.
None of the results should contain duplicates.

Give an XPath query or an XQuery query that returns the *names* of those riders who participated in the 2021 tour.
Write the query string in the file `query1.xq` and call it with

```
saxonXQ.bat -s:exam.xml query1.xq
```

Copy-and-paste the query from query1.xq afterwards here:

Exercise 4 (XPath (b) [4 Points])
Give an XPath query or an XQuery query that returns the *years* of those tours that started/passed or finished on or over the *Mount Ventoux*.
Write the XPath query string in the file `query2.xq` and call it with

```
saxonXQ.bat -s:exam.xml query2.xq
```

Copy-and-paste the query from query2.xq afterwards here:

Exercise 5 (XPath/XQuery (c) [6 Points])
Give an XPath query or an XQuery query which, for every rider, yields the total number of stage victories he ever had in the Tour de France. The results should be ordered by the number of victories descending in the form

```
<rider name="..." number="..."/>
```

Copy-and-paste the query from query3.xq afterwards here:

Exercise 6 (XPath/XQuery (d) [6 Points])
Give an XPath query or an XQuery query that yields the names of all riders that *at least once finished* a stage *passing* the Mont Ventoux, but never finished the *last* stage of a tour (this stage usually leads to Paris, but also other intermediate stages might have their destination in Paris).
Copy-and-paste the query from query4.xq afterwards here:

Exercise 7 (XPath/XQuery (e) [6 Points])
Give an XQuery query or an XPath query that returns those countries/country codes from which in *every year since 1990* at least one rider participated in the tour.
Copy-and-paste the query from query5.xq afterwards here:

Exercise 8 (XPath/XQuery (f) [9 Points])
Give an XQuery query that returns for the 2021 tour the total result, i.e., for every rider who finally completed the last stage, the total time he needed. The results should be of the format

```
<rider name="..." totalduration="..."/>,
```

ordered by the total time ascending. The query must allow that for evaluating the result for other years, only the year has to be changed.
Copy-and-paste the query from query6.xq afterwards here:

Exercise 9 (XPath/XQuery (g) [9 Points])
Give an XQuery query or an XSLT stylesheet that, for each rider R, returns a list (without duplicates) of all mountains or places P higher than 1000m that he reached during his Tour de France participations. For each such rider R, the result should be of the form

```
<result rider="name-of-R">
   name-of-place-P1 ... name-of-place-Pn
</result>
```

(no duplicate P's, in any order, with arbitrary whitespaces)
Copy-and-paste the query from query7.xq afterwards here:

Exercise 10 (XSLT [15 Points])
Extend the given XSL stylesheet frame `exam.xsl` to an XSLT stylesheet that returns for the 2021 tour a simple HTML page that contains for every rider a table which lists his name, and then his positions in each of the stages that he finished.
Use the following call to execute it:

```
saxonXSL.bat -s:exam.xml exam.xsl
```

Copy-and-paste the XSLT stylesheet from exam.xsl afterwards here:

Exercise 11 (Miscellaneous (a) [4 Points])
Consider that the same task/database should be realized by a relational database using SQL.
What would be a central difficulty in contrast to XML?
Which query/queries from above would be much more complicated?

Exercise 12 (Miscellaneous (b) [4 Points])
Consider that the database should be extended as follows: for every town/mountain, an HTML page/tree with some touristic information should be stored.
Describe (shortly) how to extend the XML and the DTD.

The following frames can be used:

- Frame for XML file `exam.xml`:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE put-name-here SYSTEM "exam.dtd">
  to be extended here
```

- Frame for XML stylesheet `exam.xsl`:

```
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
  version="2.0">
  to be extended here
</xsl:stylesheet>
```