Exam XML & Databases [211096]

Wednesday April 14, 2010; 8:45 – 12:15 h.

Allowed on exam: slides, reader, print outs, notes on paper

Not allowed on exam: electronic devices

About the exam

There are 10 questions. For each question, an indication of the associated paper from the reader is given, if appropriate. Moreover, the number of points for each question is mentioned. The points add up to a total of 90 points. You receive 10 points bonus for showing up at the exam. The final grade is determined by dividing the total score by 10.

Consider the following example XML document (eurocottage.xml) found in the XML database of a website for renting holiday cottages. The example document contains only one cottage whereas in reality there are of course 10,000s. The database receives many more select queries than update queries. The descriptions are geo-tagged, i.e., names of things that have a geographic meaning are surrounded by an element "geotag" giving its type (e.g., river or city) and its id (for lookup in some other document).

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE eurocottage [
 <!ELEMENT eurocottage (cottage*)>
 <!ELEMENT cottage (#PCDATA | geotag)* >
 <!ELEMENT geotag (#PCDATA) >
 <!ATTLIST cottage name CDATA #REQUIRED
                   code CDATA #REQUIRED
                   maxpersons CDATA #IMPLIED>
 <!ATTLIST geotag type CDATA #REQUIRED
                  id CDATA #REQUIRED>
] >

<eurocottage>
   <cottage name="Rosmol, De Lutte" code="447675" maxpersons="8">
     This detached newly built holiday house (210m2) on a villa park lies
      against the <geotag type="country" id="2634">German</geotag> border
      in the village <geotag type="city" id="93745">De Lutte</geotag> close
      to the <geotag type="road" id="921">A1</geotag>. Ideal for peace
      lovers, hikes, cycling and the nature. Discover one the most
      beautiful area of <geotag type="region" id="7243">Twente</geotag>.
      Diverse hiking and cycling routes are possible in this woody
      surrounding. The beautiful nature area the
      <geotag type="area" id="854234">Lutterzand</geotag> where the rivulet
      <geotag type="river" id="77122">Dinkel</geotag> has a free scope. A
      bit over the border you will find the splendid
      <geotag type="city" id="165">Bad Bentheim</geotag> with its beautiful
      castle and swimming pool (health spa).
</cottage>
</eurocottage>
```

Question 1 (16 points)

[XQuery] Give XQuery queries that are as short as possible for the questions below. Note that the queries need to properly answer the questions for any document that is valid according to the embedded DTD of eurocottage.xml, i.e., not only for the given document.

a) *Give the names of all cottages that mention a river in the description.*
b) *Produce an overview that gives per city: the name of the city and the number of cottages that refer to that city in their description.*
c) *The eurocottage.com website obviously displays the information around of one cottage on a web page. Given the code of a cottage bound to a variable $cottagecode, produce the description of the cottage as plain text (i.e., without the geotags). The result should look like "This detached newly built holiday house (210m2) on a villa park lies against the German border in the village De Lutte close to the A1. ..."*
d) *Modify the previous query such that the geo-tagged items become clickable, i.e., they are surrounded by '' and '' where ID is the id in the geotag. The result should look like "This detached newly built holiday house (210m2) on a villa park lies against the German border in the village De Lutte close to the A1. ..."*

***Question 2** (13 points)*

[Paper F/G] Follow the XPath Accelerator approach to answer the following question:

a) *Draw the pre/post plane of eurocottage.xml. To keep it small, restrict yourself to only the bold part of the XML document.*
b) *By how many nodes does the XML database grow if we add one geo-tag to some part in the middle of the description of some cottage? Explain your answer.*
c) *Paper G describes a technique called estimated skipping. Given the pre-order rank, post-order rank and level of the first geotag-element in the document, explain how you calculate the pre-order rank of the first following node.*

***Question 3** (11 points)*

[Paper F] Given the query: `//geotag[text()="Twente"]/@id` follow the XPath Accelerator approach to answer the following question:

a) *Give the corresponding SQL-query to evaluate the above XPath-query. Make sure that it closely adheres to XPath semantics, i.e., the result should be in document order and duplicate-free and for each axis step the appropriate test for node kind is applied (element, attribute or text node).*
b) *XPath semantics specifies that the result of a query is "duplicate-free". Does that mean that the above query will always return one value? Explain your answer.*

***Question 4** (7 points)*

Suppose the people running *eurocottage.com* plan to redesign their application using an XML database solution. Answer the following questions and motivate your answer (this question will be judged based on the arguments presented).

a) *Which of the XML database storage and query methods that were addressed in the course is most appropriate?*
b) *What XML benchmark would be most appropriate for evaluating the system?*

***Question 5** (9 points)*

[Paper W/X] Suppose there is another XML database running on a remote server map.eurocottage.com. It contains a module "mapservices.xq" with among others the function "getCoordinate($id as xs:string) as element(gml:coordinate)". Given the following query:

```
import module namespace ec="eurocottage"
   at  "http://map.eurocottage.com/mapservice.xq";
<gml:coordinates>{
   let $cottage := doc("eurocottage.xml")//cottage[@code="447675"]
  for $g in $cottage//geotag
  return  execute at {"xrpc://map.eurocottage.com"}
                     {ec:getCoordinate($g/@id)}
}</gml:coordinates>
```

a) *How many messages are being exchanged with the remote server map.eurocottage.com if we execute this query? Explain your answer.*
b) *Suppose we delete the geo-tag for "German", i.e., replace ' <geotag type="country" id="2634">German</geotag>' with simply 'German'. In the pre/post/level storage scheme of Papers F and G, which rows need to be updated besides the ones that are deleted? Explain your answer.*

***Question 6** (13 points)*

[Paper I/Y/Z] Assume you are looking for: "All geotags, inside cottages about *cycling* or about *swimming*, that (the cottages) also contain geotags about *Bentheim*." You want the best matches. Answer the following question:

a) *Give an expression of the query using Burkowski's algebra for contiguous text extents.*
b) *Give the PF/Tijah query*
c) *Give the XQuery Full-text query*
d) *What is the relation between the approaches under a), b) and c)? What are the main differences?*

***Question 7** (8 points)*

[Paper S] Given the query below:

```
//cottage[geotag="Bad Bentheim"]/geotag
```

Follow the Timber approach to answer the following questions

a) *Give the TAX query plan for this query including, if needed, the pattern trees, adornment lists, project lists, etc.*
b) *For each pattern tree, give the witness trees.*

***Question 8** (8 points)*

[Paper A] The people at *eurocottage.com* recently acquired *hiking4you.com*, a company which main asset is a huge database of hiking routes. The database has the followin relational schema:

```
route(route_id, name, length)
directions(id, route_id, description, geotag, geotype)
```

In the directions table, `route_id` is a foreign key, `description` contains things like "This route starts at the main square. Walk in the direction of the curch", and `geotag` and `geotype` contain information similar to the *eurocottage.com* database, for instance "Bad Bentheim" and "city", respectively. To integrate the data with the *eurocottage.com* database, the relational database will need to produce data that conforms to the following DTD:

```
<!DOCTYPE route [
  <!ELEMENT route (name?, length?, direction+) >
  <!ATTLIST route route_id ID #REQUIRED >
  <!ELEMENT direction (description, geotag, geotype) >
  <!ELEMENT name (#PCDATA) >
  <!ELEMENT length (#PCDATA) >
  <!ELEMENT description (#PCDATA) >
  <!ELEMENT geotag (#PCDATA) >
  <!ELEMENT geotype (#PCDATA) >
]>
```

Answer the following questions.

a) *Give the XQuery query that produces "all routes that pass the city Bad Bentheim" from the XML standard mapping of the relational database. The answer should conform to the DTD.*
b) *Give the SQL/XML query that produces the same result directly from the relational data.*

***Question 9** (5 points)*

Found on amazon.com:

> "Data on the Web : From Relations to Semistructured Data and XML" by Serge Abiteboul (Author), Dan Suciu (Author), Peter Buneman (Author) The book is aimed at readers already familiar with database concepts, it includes little introductory material. It quickly lays out the concepts of self-describing semi-structured data and how XML fits into this approach to data representation.

Define the following terms and explain how XML fits into this approach to data representation:

a) *self-describing*
b) *semi-structured data*