Exam XML & Databases [192110961]

Tuesday April 8, 2013; 13:45 – 17:15 h. in CU-B103
Maurice van Keulen and Djoerd Hiemstra
Allowed on exam: slides, reader, print outs, notes on paper
Not allowed on exam: electronic devices

There are 9 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.

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE schema [
  <!ELEMENT schema (game*)>
    <!ELEMENT game (group,(notplayed|(result, review?)))>
  <!ELEMENT group (#PCDATA)>
  <!ELEMENT review (#PCDATA)>
  <!ELEMENT notplayed EMPTY>
  <!ELEMENT result EMPTY>
  <!ATTLIST game teama CDATA #REQUIRED teamb CDATA #REQUIRED>
  <!ATTLIST result teama CDATA #IMPLIED teamb CDATA #IMPLIED>
] >
<schema>
  <game teama="GER" teamb="CRC">
     <group>A</group>
     <result teama="3" teamb="1"/>
     <review>A great opening game in which the Costa Ricans gave good
performance, but could not beat the Germans.</review>
  </game>
  <game teama="ECU" teamb="GER">
     <group>A</group>
     <notplayed/>
  </game>
</schema>
```

Question 1 *(6 points)*
Given the XML data from "games.xml" above, answer the following questions:

a) *Is the data **valid** and/or **well-formed**? Explain your answers.*

b) *Is this so-called data-centric XML or document-centric XML? Explain your answer.*

Question 2 *(12 points)*
[XQuery] Give XQuery queries that are as short as possible for the questions below. Note that the queries need to properly produce answers for any document that is valid according to the embedded DTD, i.e., not only for the given document.

a) *Give the review of the game between Germany (GER) and Costa Rica (CRC).*

b) *Give results for all games in group "A" that have been played. The result should look like this:* `<result>GER - CRC: 3-1</result>...`

c) *Somebody wrote the query below. (S)he expected the answer to be* `(6,8,10)`, *but the actual answer is* `(6,7,8)`. *What are the main reasons/causes for this to be the correct answer.*

```
LET $a := 0
FOR $b in (5,6,7)
LET $a := $a + 1
RETURN $a + $b
```

Question 3 *(12 points)*
[Paper F/G]
Given the example document. Follow the XPath Accelerator approach to answer the following questions:

a) *Draw the pre-post plane of the example document (i.e., a graph where every node is drawn at the (x,y) coordinate where x is the pre-order and y is the post-order rank for the node).*

b) *There are two game-elements in the example document. Let us call these two nodes G1 and G2. Indicate in the pre-post plane of question 3a the region of the plane that is necessarily empty, because G2 is a following node of G1.*

c) *Give the corresponding SQL-query to evaluate the XPath-query*

```
/descendant::notplayed[preceding-sibling::group='A']/parent::game
```

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).

Question 4 *(8 points)*
[Paper O] Given the XPath queries Q1=`//game/group`*, Q2=*`//game//*`*, Q3=*`/schema/game/group`*, Q4=*`//game//result`*, and Q5=*`//game/@teama`.

a) *Draw the YFilter for all queries.*

b) *Section 3.3 discusses how to deal with predicates. They mention two approaches, an "intuitive approach" and an approach with a "separate selection operator". Suppose we would have an addition query Q6 =* `//game/group[. = "A"]` *(i.e., the same as Q1 but with a predicate). Extend the NFA according to the "intuitive approach".*

Question 5 *(8 points)*
[Paper I, Z] Given the following query in Burkowski's algebra for contiguous text extents:

```
(<result>) SN (<game> SW ( <review> SW {"batted", "foul"} ))
```

a) *Give the XQuery Full-Text equivalent for this query, assuming that the system orders the results by their relevance to the query*

b) *Suppose we need the results in document order: Is there an equivalent XPath query? Explain your answer.*

Question 6 *(12 points)*

[Paper T] Consider the storage scheme proposed by Meier for the eXist Database System. Suppose we need to store the *games.xml* document.

a) *How many unique identifiers, including the spare identifiers, would be used by the level-order numbering scheme to store the document? Explain your answer.*

b) *How many unique identifiers, again including the spare identifiers, would be used by the alternating level-order numbering scheme. Explain your answer.*

c) *Compare this approach to the XPath Accelerator storage approach. What are the adventages and disadvantages of level-order numbering?*

Question 7 *(12 points)*

[Paper W] A database server at `espn.com` *which is located somewhere in the United States contains a collection of XML documents with students. The server provides the following XQuery module:*

```
module namespace espn="http://espn.com/teams-v1.0";
declare function
  espn:GetTeams($id as xs:string) as node()*
  {
    doc("teams.xml")//team[id = $id]
  };
```

a) *Explain the input and output of the function* `GetStudents()`.

At the University of Twente, for each game that plays at home (the **teama** attribute of element **game**) from the database from Question 1, we want to use GetTeams() to check if teams occur in the database from `espn.com`.

b) *Give the XQuery XRPC query needed at* `utwente.nl` *to execute* `GetTeams()` *for this scenario.*

c) *How many XRPC calls will be done for the query under b)? Explain your answer.*

Question 8 *(12 points)*

[Paper M, Y and Z] Query 14 of the XMark benchmark test collection addresses full text querying: **Q 14.** *Return the names of all items whose description contains the word 'gold'.* Answer the following questions.

a) *Can this query be answered by XQuery or XPath without full text extensions? Explain your answer.*

b) *What functionality of XQuery Full-Text can be used for XMark Query 14 that is not supported in ordinary XQuery? Give the XQuery Full-Text query that shows this.*

c) *Is the XMark benchmark well-suited for evaluating full text querying? Explain your answer.*

```
CREATE TABLE movies (
        movie_id INTEGER,
        title VARCHAR NOT NULL,
        year INTEGER NOT NULL,
        plot_outline VARCHAR,
        rating INTEGER,
        PRIMARY KEY (movie_id),
);
CREATE TABLE persons (
        person_id INTEGER,
        name VARCHAR NOT NULL,
        PRIMARY KEY (person_id)
);
CREATE TABLE actors (
        movie_id INTEGER,
        person_id INTEGER,
        role VARCHAR,
        FOREIGN KEY (movie_id) REFERENCES movies(movie_id),
        FOREIGN KEY (person_id) REFERENCES movies(person_id)
);
```

Question 9 *(8 points)*

[Paper A] Given the SQL schema above, answer the following questions:

a) *give the most strict DTD for which the SQL/XML "standard mapping" is valid, assuming "absent" behaviour of NULL values. Elements of type (#PCDATA) do not have to be listed.*

b) *Give the SQL/XML query that produces for each movie the names of all persons that acted in that movie. Use the element names "movie" and "name".*