Exam XML & Databases [211096]

Wednesday April 15, 2009; 09:00 – 12:30 h.

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

Not allowed on exam: electronic devices

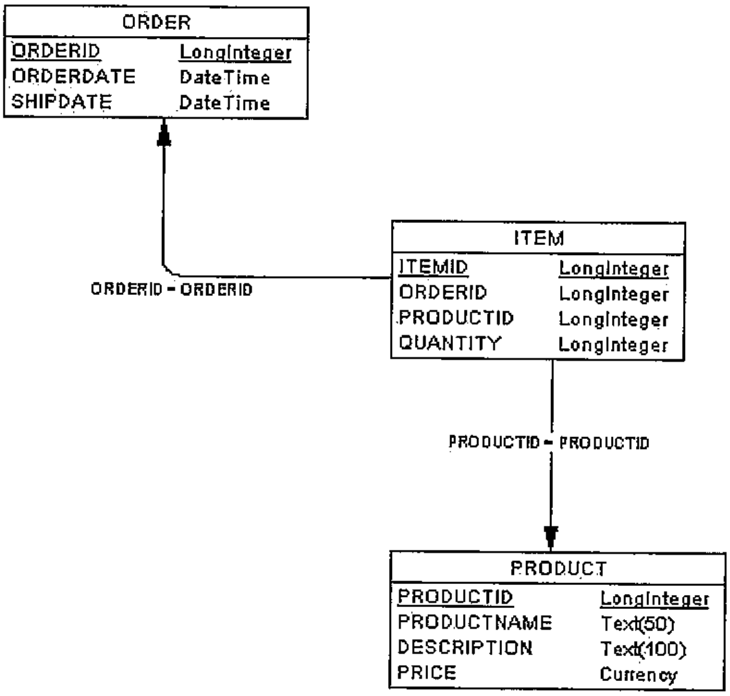
About the exam

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.

1 Relational Data

***Question 1** (13 points)*

[Paper A] Given the relational schema at the right:



a) Give an SQL/XML query that produces Order elements, and within those elements product elements. All relational attributes ending in "ID" have to be XML attributes in the result. All other relational attributes have to be XML elements.

b) Will the resulting XML data under a) contain all data from the relational database? Explain your answer.

c) Give an XQuery query that takes the standard XML mapping, and produces the same results as under question a).

2 XML Data

Consider the following XML document (recipes.xml):

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE recipe [
 <!ELEMENT recipe (title, ingredient+, instructions)>
 <!ELEMENT instructions (step+) >
 <!ELEMENT title (#PCDATA) >
 <!ELEMENT ingredient (#PCDATA) >
 <!ELEMENT step (#PCDATA) >
 <!ATTLIST recipe prep_time CDATA #REQUIRED cook_time CDATA #REQUIRED>
 <!ATTLIST ingredient amount CDATA #REQUIRED unit CDATA #REQUIRED>
] >
<recipe prep_time="5 mins" cook_time="3 hours">
   <title>Basic bread</title>
   <ingredient amount="8" unit="dL">Flour</ingredient>
   <ingredient amount="10" unit="grams">Yeast</ingredient>
   <ingredient amount="4" unit="dL">Warm water</ingredient>
   <ingredient amount="1" unit="teaspoon">Salt</ingredient>
   <instructions>
     <step>Mix all ingredients together.</step>
     <step>Knead thoroughly.</step>
     <step>Cover with a cloth, and leave for one hour in warm room.</step>
     <step>Knead again.</step>
     <step>Place in a bread baking tin.</step>
     <step>Cover with a cloth, and leave for one hour in warm room.</step>
     <step>Bake in the oven at 180(degrees)C for 30 minutes.</step>
   </instructions>
</recipe>
```

***Question 2** (14 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 recipes.xml, i.e., not only for the given document.

a) *All distinct units in which "Salt" is measured.*
b) *An overview of all recipes with prep_time, cook_time, title, number of ingredients and number of steps.*
c) *An overview that per ingredient lists all recipes in which it plays a role.*

***Question 3** (8 points)*

[Paper D] Derive a relational database schema from the embedded DTD of recipes.xml by the following steps:

a) *Give the simplified DTD.*
b) *Give the DTD graph of the simplified DTD.*
c) *Derive the relational schema from the DTD graph using the shared inlining technique.*

***Question 4** (14 points)*

[Paper E/F] Paper E defines three order encoding methods for representing XML's ordered data model: Global Order, Local Order en Dewey Order. Furthermore, the paper considers two cases: when the schema of the input documents is unknown (the schema-less case) and when the schema is known (the schema-aware case). Finally, the paper addresses three query types: Unordered Selection, Ordered Selection, and Reconstruction.

a) *Which of the three methods is used by paper F: "Accelerating XPath location steps"? Explain your answer*

b) *Which of the two cases is considered by paper F? Explain your answer.*

c) *Which of the three query types is addressed by paper F? Explain your answer.*

***Question 5** (19 points)*

Given the following query: `//recipe[title="Basic bread"]/instructions`

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

a) *How many nodes does recipe.xml produce in the document table (Paper F: table "accel"; Paper G: table "doc")? Explain your answer.*

b) *Give the corresponding SQL-query to evaluate the above XPath-query.*

[Paper S] Follow the Timber approach to answer the following questions:

c) *Give the TAX query plan for this query including, if needed, the pattern trees, adornment lists, project lists, etc.*

d) *For each pattern tree, give the witness trees.*

***Question 6** (8 points)*

[paper Y and Z]

a) *Paper Z describes 2 alternative approaches for implementating text search in XQuery by "extending XQuery with full-text functions". Does PF/Tijah follow one of these approaches, and if so, which of the 2 approaches? Explain your answer.*

b) *Suppose we want to retrieve a ranked list of titles from the best matching Wikipedia articles about "high protein flour" with the following PF/Tijah query:*

```
let $d := collection("recipes")
return tijah-query($d, "//recipe[about(., high protein flour)]")/title
```

c) *Given that the syntax of the query is correct, explain why the query does not give the desired result.*

d) *Suppose we want to retrieve the same ranked list as under b) with the following XQuery Full-Text query:*

```
let $d := collection("recipes")
for $a in $d//article[. ftcontains ("high protein flour")]
return $a/title
```

e) *Given that the syntax of the query is correct, explain why this query does not give the desired result.*

***Question** 7 (9 points)*

[Paper O] Suppose new recipes are broad-casted as an XML stream of recipe-elements (e.g., an RSS-feed) according to the given embedded DTD.

a) *Suppose there are 4 users (U1, U2, U3, and U4) who are interested in being notified whenever there is a new recipe. Draw the corresponding Finite State Machine (FSM) according to the YFilter approach.*

b) *Suppose we have a 5th user who is only interested in recipes that have more than 3 ingredients. YFilter as described in the paper only supports child and descendant axes, but no counting of children. Discuss how the technique could be extended to also support counting.*

***Question** 8 (5 points)*

[Paper W] Suppose we have 3 servers running an XML DBMS supporting Bulk XRPC. Each server contains an XQuery module with the functions and variables as given below. The idea is simple: server A calculates for the numbers $i from 1 to 100 the formula $bf * $i + $cf * i. The complication is that the factors $bf and $cf are hidden and only known to servers B and C respectively. There are **two methods of calculation**: (1) A asks B which in turn asks C (functions a1, b1, and c), or (2) A asks B and C directly (functions a2, b2, c).

Server A:

```
declare function a1() as xs:integer
{  for $i in (1 to 100) return execute at {B} {b1($i)} }
declare function a2() as xs:integer {
{    for $i in (1 to 100)
     return execute at {B} {b2($i)} + execute at {C} {c($i)} }
```

Server B:

```
declare variable $bf := 5;
declare function b1($i as xs:integer) as xs:integer
{ $bf * $i + execute at {C} {c($i)} }
declare function b2($i as xs:integer) as xs:integer { $bf * $i }
```

Server C:

```
declare variable $cf := 3;
declare function c($i as xs:integer) as xs:integer { $cf * $i }
```

a) *How many messages are being sent between A↔B, B↔C, and A↔C for calculation method (1)? Explain your answer.*

b) *How many messages are being sent between A↔B, B↔C, and A↔C for calculation method (2)? Explain your answer.*

c) *Both calculation methods hide factors $bf and $cf in their respective servers, i.e., server A does not need to know them. An alternative calculation method would be that server A asks B and C for the value $bf and $cf and do the rest of the calculation itself. What do you expect regarding execution time: a big improvement, a small improvement, a small worsening, or a big worsening? Explain your answer.*