CARDIFF UNIVERSITY
EXAMINATION PAPER

Academic Year:	2012-2013
Examination Period:	Spring
Examination Paper Number:	CMT207
Examination Paper Title:	Information Modelling and Database Systems
Duration:	2 hours

Do not turn this page over until instructed to do so by the Senior Invigilator.

Structure of Examination Paper:
There are 5 pages.
There are 4 questions in total.
There are no appendices.
The maximum mark for the examination paper is 60 and the mark obtainable for each part of a question is shown in brackets alongside the question.

Students to be provided with:
The following items of stationery are to be provided:
1 answer book.

Instructions to Students:
Answer 3 questions.
The use of a translation dictionary between English or Welsh and another language, provided that it bears an appropriate school stamp, is permitted in this examination.

Question 1

(a) Consider the following XML document:

```
<?xml version="1.0" encoding="UTF-8"?>
<cv>
 <name>John Smith</name>
 <pic>http://images.com/js.jpg</pic>
 <skills>
  <skill>XML</skill>
  <skill>DTD</skill>
  <skill>XSD</skill>
 </skills>
 <url>http://john-smith.com</url>
</cv>
```

A Web server technology is made available which can be used to inject values collected from this XML through XPath expressions into an HTML page. To use this technology, the delimiters <% and %> must surround the XPath expression. For example *<% /my/xpath/element %>*. Using this technology write a simple HTML page extracting data from the XML document above. The HTML page should have:

- as title "Curriculum Vitae (CV)"
- A heading of type 1 displaying the name, in blue (use inline CSS).
- A paragraph with the image with a specified width and height of 200 pixels.
- A heading of type 2 displaying the text "Skills:" in italic (use inline CSS).
- A paragraph containing a bullet list displaying the skills.
- In the same paragraph, one line below the list, add a hyperlink to the URL specified in the XML.

[7]

(b) Explain the common purpose of XML Schema and DTD, as well as *three* differences between their expressive powers. [2]

(c) Write the DTD corresponding to the XML document in part (a), assuming that a pic is optional, that there can be an unlimited number of skills (possibly 0), and that the *<url>* element can be replaced by a *<message>* one containing a short text. Either *<url>* or *<message>* should appear. [4]

(d) Write the XSD corresponding to the XML document in part (a), following the same assumptions as in question (c). [7]

Total 20 marks

Question 2

(a) Briefly explain the similarity and give two differences between a database and a data warehouse. [3]

(b) Define the three operations involved in ETL, giving two example tasks for each. [3]

(c) Explain the nature of the Entity Resolution problem in terms of its goal and provide two examples of domain applications where ER would be useful. [2]

(d) Apply the Levenshtein distance on the following string pairs, using 1 as weight for every operation, and justifying the numeric answer by listing the operations involved.

place - pace
knitting – writing [4]

(e) Use the first part of the Apriori algorithm, showing the computation steps, to find itemsets containing more than two items and with support larger than 50% in the following transactions.

Bread, Milk
Beer, Bread, Diaper, Eggs
Beer, Coke, Diaper, Milk
Beer, Bread, Diaper, Milk
Bread, Coke, Diaper, Milk [8]

Total 20 marks

Question 3

(a) Describe the problems in the current Web and the vision of the Semantic Web. [2]

(b) What is Linked Data? [2]

(c) Explain the differences between RDF, RDFS and OWL. [3]

(d) Consider the following XML (partial):

```
<ancestors>
  <person name="Kelly">
    <father>
      <person name="Paul">
        ....
      </person>
    </father>
    <mother>
      <person name="Samantha" />
    </mother>
  </person>
  ...
</ancestors>
```

and the following facts:

A man is a person
A woman is a person
A mother is a woman
A father is a man

Give an RDF representation for the facts above. They can be stated as a set of triples, there is no need to use an RDF notation. Assume the XML document provides ancestry data about several families and that the RDF references entities of the XML document. Give one sample query that demonstrates the benefit of using RDF and SPARQL compared to pure XML with XPath. It is OK to state the query in natural language; you do not need to use the SPARQL syntax. You must explain the benefits and why you choose the query. [6]

(e) A consumer retail brand is looking to identify the most frequently purchased products from a selection of customers as part of a market research initiative. Customer data is distributed in local centres C_1 to C_l, each user $(u_1, \ldots, u_m)$ is linked to products $(p, \ldots, p_n)$. Describe the steps, inputs and outputs of a MapReduce algorithm which would determine the most frequent products bought. [7]

Total 20 marks

Q4 (a) The purchasing department of a manufacturing company is responsible for ordering components to be used in the manufacture of their range of products. The following table shows information taken from two order forms, order# 1234 and order# 1235. An order is placed with a supplier for quantities of one or more different parts. A part is identified uniquely by the part#, and different suppliers of the same part use the same part#.

Draw one or more functional dependency diagrams for the data presented on the order form and hence produce a set of 3rd normal form relations. [10]

Order#	Supplier#	SupplierName	Part#	Description	Quantity
1234	501	Bloggs & Co	7123	Flange	100
1234	501	Bloggs & Co	4832	Sprocket	5
1235	413	Widgets-R-Us	2975	50cm Cam	20
1235	413	Widgets-R-Us	1863	2.5cm Bracket	100
1235	413	Widgets-R-Us	4115	Standard Beam	2

(b) In a database system, two transactions A and B progress in the following manner. At time t1 transaction A fetches a record R. At time t2 transaction B fetches the same record R. At time t3 transaction A updates record R. At time t4 transaction B updates record R.

Analyse this situation showing clearly the problems associated with it. How do database management systems typically use locking to overcome the problems associated with this situation? Are there any further issues with this solution? [10]

Total 20 marks