UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFORMATICS 1 - DATA AND ANALYSIS

Monday 11 May 2009

14:30 to 16:30

Convener: M O'Boyle
External Examiner: R Irving

INSTRUCTIONS TO CANDIDATES

1. **ANSWER ALL QUESTIONS.**
2. **DIFFERENT QUESTIONS MAY HAVE DIFFERENT NUMBERS OF TOTAL MARKS. Take note of this in allocating time to questions.**

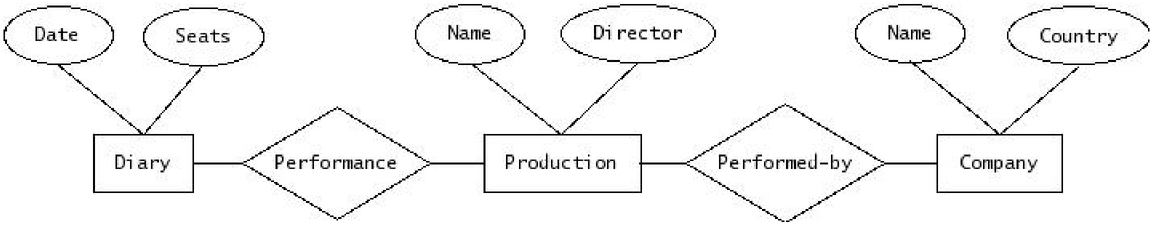

Figure 1: A preliminary ER diagram

1. A theatre wants to set up a database to record its schedule of performances. The database will contain information about each performance, the show being performed and the performing company. They intend to record the following information.

   - For each date, the production that takes place on that day, and the number of seats still available for sale. (The theatre has at most one performance each day.)
   - For each production, the name of the show (for example, *"The Magic Flute"*), the director, and the company performing it (for example, *"Scottish Opera"*). Each production has a single director and is performed by a single company. However, there may be different productions of the same show (for example *"The Magic Flute"*) by different directors and companies.
   - For each company, the name of the company and the country it comes from.

   Figure 1 presents a preliminary (incomplete) ER diagram incorporating some of the above information.

   (a) Define the notion of a *key* for an entity set in an ER diagram. Assign keys to the entity sets in Figure 1, explaining, in each case, any assumptions you need to make in order to do so. [*8 marks*]

   (b) Complete the ER diagram of Figure 1, by redrawing it and including: appropriate notation for the keys, and any missing key and participation constraints. [*6 marks*]

   (c) Use the SQL Data Definition Language to present relational schemata that implement the above ER diagram, using three tables:

   - `Diary` — with 4 fields: date, seats, production-name, director.
   - `Production` — with 3 fields: name, director and company.
   - `Company` — with 2 fields: name and country.

   You may choose your own field names, but make sure that it is clear which of the above fields each name refers to. Also, ensure that you capture all the constraints on the data. [*8 marks*]

(d) Using the relational schemata defined in part (c) above, formulate the following query three times; once each in relational algebra, tuple-relational calculus and SQL.

- Find the dates of every performance by *"Scottish Opera"*. [*12 marks*]

(e) Formulate the following query in SQL.

- Count the number of different companies from Scotland that have given sold-out performances. (A performance is sold out if there are 0 seats available for it.) [*6 marks*]

```
<books>
 <book quantity="15" department="computing">
  <title>
   Database Management Systems
  </title>
  <author>
   Raghu Ramakrishnan
  </author>
  <author>
   Johannes Gehrke
  </author>
  <publisher>
   McGraw-Hill
  </publisher>
 </book>
 <book quantity="5" department="fiction">
  <title>
   A Tale of Two Cities
  </title>
  <author>
   Charles Dickens
  </author>
  <publisher>
   Penguin
  </publisher>
 </book>
</books>
```

Figure 2: An XML document

2. The XML document in Figure 2 illustrates a possible representation of a bookshop's database containing information about its stock.

   (a) Draw the XML tree (following the XPath Data Model) corresponding to the document in Figure 2. [*8 marks*]

   (b) Write a DTD to specify the XML format for bookshop databases illustrated by the example in Figure 2. [*8 marks*]

   (c) Write XPath expressions to return the following lists of text strings from any XML document that is valid with respect to the DTD for Figure 2.

      i. The names of all authors.

      ii. The titles of all books by Charles Dickens.

iii. The titles of all books that are out of stock. (You may assume that these are identified by having a value of `"0"` for the `quantity` attribute.) [*9 marks*]

(d) Mark-up the quotation below with the kind of metadata that would be used to annotate it if it were included in a corpus (such as the British National Corpus), and write out the result in XML format.

> *We reject as false the choice between our safety and our ideals.*

Describe the meaning of the different annotations you use.

(Your answer to this question will be evaluated on the basis of the choice of appropriate forms of annotation, and not on the linguistic correctness of your annotations.) [*10 marks*]

3. (a) What is the *information retrieval task*? What are the underlying assumptions on which it is based? Illustrate these assumptions using a suitably chosen example of an information retrieval task. [*8 marks*]

(b) The performance of an information retrieval system can be evaluated in terms of its *precision*, P, and *recall*, R. Informally, R may be defined as the proportion of those documents matching the objectives of the search that are correctly returned by the system. Give a similar informal definition of P. [*2 marks*]

(c) The mathematical formula defining recall is:

$$R = \frac{TP}{TP + FN}$$

Name and define the terms TP and FN appearing in this formula. Give a formula defining the precision P, and also explain any new terms that appear in this formula. [*7 marks*]

(d) Suppose a document collection contains 100 documents, 20 of which are relevant to a particular query, which is submitted to an information retrieval system. The system returns 25 documents, 15 of which are relevant to the query. Calculate the precision and recall for this system, showing the details of your calculations (you should not need a calculator for this). [*4 marks*]

(e) What is the *precision-recall tradeoff*? Compare the relative importance of precision and recall in the example information retrieval task you discussed in your answer to question 3(a) above. [*4 marks*]