UNIVERSITY OF LONDON

CM3010

BSc EXAMINATION

COMPUTER SCIENCE

Databases and Advanced Data Techniques

Release date: Thursday 19 September 2024 at 12:00 midday British Summer Time

Close date: Friday 20 September 2024 by 12:00 midday British Summer Time

Time allowed: 4 hours to submit

INSTRUCTIONS TO CANDIDATES:
Part A of this assessment consists of a set of **TEN** Multiple Choice Questions (MCQs). You should attempt to answer **ALL** the questions in **Part A.** The maximum mark for Part A is **40**.

Candidates must answer **TWO** out of the **THREE** questions in **Part B**. The maximum mark for Part B is **60**.

Part A and Part B will be completed online together on the Inspera exam platform. You may choose to access either part first upon entering the test area but must complete both parts within **4 hours** of doing so.

Calculators are **NOT** permitted in this examination.

You may use **ONE** A4 page of previously prepared notes in this examination. Please hold up your notes to the camera at the start of the examination.

File upload is **NOT** permitted.

Do not write your name anywhere in your answers.

PART A

Question 1

Candidates should answer the **TEN** Multiple Choice Questions (MCQs) in Part A.

PART B

Candidates should answer any **TWO** questions from Part B.

Question 2

An enthusiast website for people interested in historical music for lute, guitar and similar instruments has a list of tens of thousands of pieces as they occur in thousands of manuscripts and prints (called **Sources**, since they are the source of the music).
The data is stored in csv files and compiled into HTML web pages by PHP scripts. The data files and scripts are stored on GitHub and edited directly by two users. The first line of each file lists the field names for the csv. Below is some sample data (edited for this exam).

First, there is a file representing **Sources** (extract given below). It lists library reference, names of the source, and date and instrument information:

```
Ref_Short;Ref_Long;Library;Name_German;Name_English;Date;Instruments
NL-At;Ms. 205.B.32;Amsterdam, Toonkunst-Bibliotheek;;;1600-1680;Baroque Lute
D-DI_M297;Mscr Dresd. M. 297;Staats- und Universitätsbibliothek Dresden;Liederbuch eines Jenenser Studenten (Lautenbuch);Songbook of a student from Jena (lutebook); 1603; Renaissance Lute
```

Then, a file representing musical works that may occur in different versions in different sources (**Concordances**). It gives the work 'concordance number' as identifier, a composer, and lists places where that work occurs:

```
Conc_no;Composer;Concordances
Conc_51;V. Gaultier;NL-At/2v – F-PnVmb7/188 – D-B40068/59r
Conc_15;V. Gaultier or D. Gaultier;NL-At/24v – GB-Balcarres/86
```

Finally, a file for each **Source**, listing the individual pieces it contains (in order). The extract below comes from a file called `NL-At.csv`, which matches the rows in the **Concordances** extract above. It lists piece and page number, musical key, and composer, along with a concordance number. In the example below, 'Page_no' numbers each piece of paper ('folio'), with 'r' referring to the first face of the paper you see, and 'v' referring to the back of the sheet (that you see when you turn the page):

```
Piece_no;Key;Page_no;Title;Composer;Conc_no
4;c minor;2v;sans titre;V. Gaultier;Conc_51
17;d minor;24v;Caprice;D. Gaultier;Conc_15
```

(a) What might be the advantages and disadvantages of a database approach compared to the file-based approach used here? Where relevant, refer to specific examples in the data above.

[6 marks]

(b) What database model would you recommend as the easiest to use, given the current state of the data? Why?

[2 marks]

(c) It has been decided to use a relational database in a future version of the site. Propose a model, listing the tables, fields, and keys of the new database. List any concerns you have about the data and your model.

[12 marks]

(d) One of the most common pieces is a song by John Dowland called ‘Lachrimae’ or ‘Flow my tears’, but we would like to check that all the relevant versions have been grouped. Write a query that finds all pieces with ‘lachrimae’ or ‘flow’ in their names that are not included in a Concordance associated with a composer called ‘John Dowland’.

[6 marks]

(e) The developer has added a web form that allows a user to provide new data, adding sources and their contents. Write an appropriate GRANT command for the account that the web application will use.

[4 marks]

Question 3

The text below is an extract of a file collecting entries for a series of poetry contests. Entrants fill in a web form with their details, a submission, and the category and competition date they are applying for. The submissions are collated into this file and then judged. Awards are given based on the judgement.

```
...
<competition theme="limericks" date="2024-01-03">
    <entry>
        <authors>
            <author viaf="23156">Edward Lear</author>
        </authors>
        <poem>
            <tei:lg type="stanza">
                <tei:l>There was an old man of Dumbree</tei:l>
                <tei:l>Who taught little owls to drink tea</tei:l>
                <tei:l>For he said, "To eat mice
                        is not proper or nice"</tei:l>
                <tei:l>That amiable man of Dumbree</tei:l>
            </tei:lg>
        </poem>
    </entry>
    ...
 </competition>
 ...
```

(a) What is the format of this file?

[1 mark]

(b) The competition website says that they save data as Text Encoding Initiative files. Are they correct? Give a more specific (and accurate) statement.

[3 marks]

(c) Write a simple XPATH expression to retrieve the first line from all entries to competitions with theme 'limericks'. Note: the precise syntax is less important in this question than the logic of your answer.

[3 marks]

(d) Design a relational model for the files, adding the ability for judges to give numerical assessments of each entry (usually, three judges score each). Give the CREATE commands needed to build a MySQL. Explain your choices and show what Normal Forms your solution is in.

[12 marks]

(e) Give a query for your database that retrieves the winning (highest scoring) entry for the Limerick challenge of the 3 Jan 2024.

[5 marks]

(f) How do the XML document-based approach and the relational model compare for this use case? What works best in each? Would there be any benefit to a hybrid approach in this example?

[6 marks]

Question 4

A researcher has received a grant from the Belgian government for a project focusing on Belgian artists from before 1600. To get started, the researcher types the following into the Wikidata query engine:

```
SHOW person ?personLabel ?placeLabel ?dob
{{
  BIND (wd:Q31 as ?country)      # Q31 is Belgium
  BIND (wd:Q483501 as ?job)      # Q483501 is Artist
  person wdt:P19 ?place ,        # P19 is place of birth
         wdt:P569 ?dob ,         # P569 is date of birth
         wdt:P106 ?job ;         # P106 is occupation
  place wdt:P17 ?country .       # P17 is country

  FILTER( YEAR(?dob) < 1600 )

  SERVICE wikibase:label {
    bd:serviceParam wikibase:language "en" .
    # allows English language labels to be returned
    # for Wikibase items
  }
}}
```

(a) What language does this query use?

[1 mark]

(b) Some of the syntax of this query is wrong. Correct it.

[2 marks]

(c) What might cause the **retrieval** of this query to be less than 100%?

[4 marks]

(d) Given the purpose of the query, what results might be returned that are not wanted?

[3 marks]

(e) The researcher originally considered querying by **country of citizenship** (P27) rather than **place of birth** (P19). Since Belgium is a young country (it was declared as an independent state in 1830), this was unsuccessful. For example, Jan Brueghel the Elder is a painter considered Belgian for this study, but his **country of citizenship** in Wikidata is **Habsburg Netherlands**. How does the query above avoid this problem?

[4 marks]

(f) Each artist is to be connected to a database of artworks held in various galleries. For each work, the database records the artist, data of creation, the medium and materials (e.g. watercolour on canvas), the height and width of the object, and a link to a digital image. The researcher has decided to create a MongoDB database of the combined data. Give a query for this database that might return all artworks made between 1520 and 1530 by artists born in Antwerp.

[6 marks]

(g) Do you think the researcher was right to use an object database for this? Evaluate a graph, object, and relational model in this context. What is special about this case that makes these work better or worse than they would in other circumstances?

[10 marks]

END OF PAPER