Statistical Methods for Database Integration
Examination
DATABASES

The exam consists in two parts:

1) Part A: **The exam is closed-book, closed-notes**;

2) Part B: **You are allowed to use lecture and labs notes**.

Each questions is assigned points expressed in cents.

PART A

Ex. 1

(a) **(10 points)** "Currently the Web offers a variety of data, even if it is embodied into HTML code". Marketing needs to compare flight offers of different companies. Consider the following data offers and describe it by means of the XML language. The second flight is active, and the economy options are displayed. **Describe carefully the visible data of this option.**

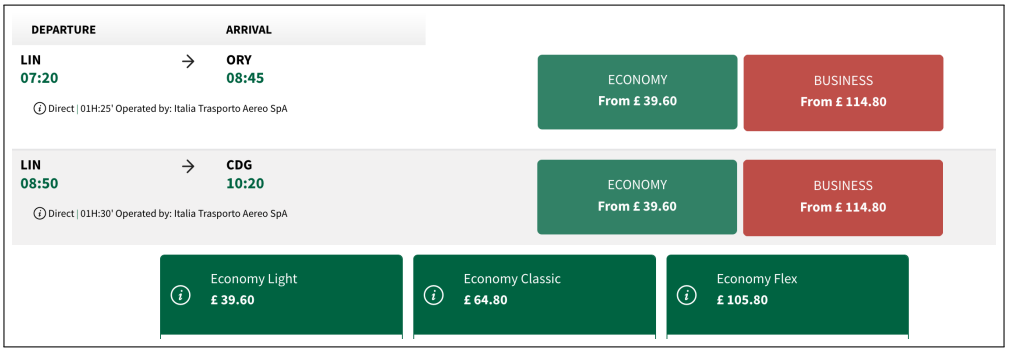


Figure 1: source: **www.itaspa.com**:

[*Sol.*:]

```
▼<flight-offers>
  ▼<flight>
    ▼<departure>
       <airport> LIN </airport>
       <time format="hh:mm"> 8:50 </time>
     </departure>
    ▼<arrival>
       <airport> CGD </airport>
       <time format="hh:mm"> 10:20 </time>
     </arrival>
    ▼<fare>
       <type> Economy </type>
      ▼<offer>
         <name> Economy Light </name>
         <price unit="£"> 39.60 </price>
       </offer>
      ▼<offer>
         <name> Economy Classic </name>
         <price unit="£"> 64.80 </price>
       </offer>
      ▼<offer>
         <name> Economy Flex </name>
         <price unit="£"> 105.80 </price>
       </offer>
     </fare>
   </flight>
 </flight-offers>
```

Figure 2: **XML file**

(b) **(10 points)** The election of the president of the cultural association held last week and you registered the results of each session (consider that 5 sessions has been held!) arranging data in the following table (relations). Reflecting a little bit you observed that **anomalies can occur**. Which ones? Maybe a normal form should be a solution. Describe how you normalize the relation. [**Explanation,** type: for presidential elections, nr.: for sessions, ...].

type	date	nr.	surname	name	birth-date	votes

[*Sol.*:] The relation should be split in three relations. Artificial keys are added.

Candidate(cID, surname, name, birth-date)

Session(code, type, date, nr.)

Session_Candidate(code, cID, votes)

(c) **(Optional: 5 points)** Data integration really represents a challenge in data management. ODBC technology can be easily used to integrate some applications and relational databases. What does ODBC technology enable? Describe the experience of data integration we have set up in the lab.

[*Sol.*: See teaching material]

Es. 2 - Data Modeling

(1) **(35 points)** "Socials, whatsapp, SMS are frequently used to collect opinions, song votes in a competition, customer satisfactions about a service, or generally appreciations, feelings,...". The goal is to promote historical buildings, gardens, artwork, enabling people to give a vote for 'appreciation, like, felling, ...' by means of socials, whatsapp, and SMS.

Draw the E/R diagram that capture the requirements stated below. **Use "ID" as key only if strictly necessary**.

(a) We generally submit people **items** to vote, which are uniquely identified by a code, they have a title, and the period (from, to) people is allowed to vote it.

(b) Each item could be identified by *one or more* **hashtag** [**group**], which is registered, adding the meaning we have assigned to it and a status, that is if this hashtag is again in use or obsolete.

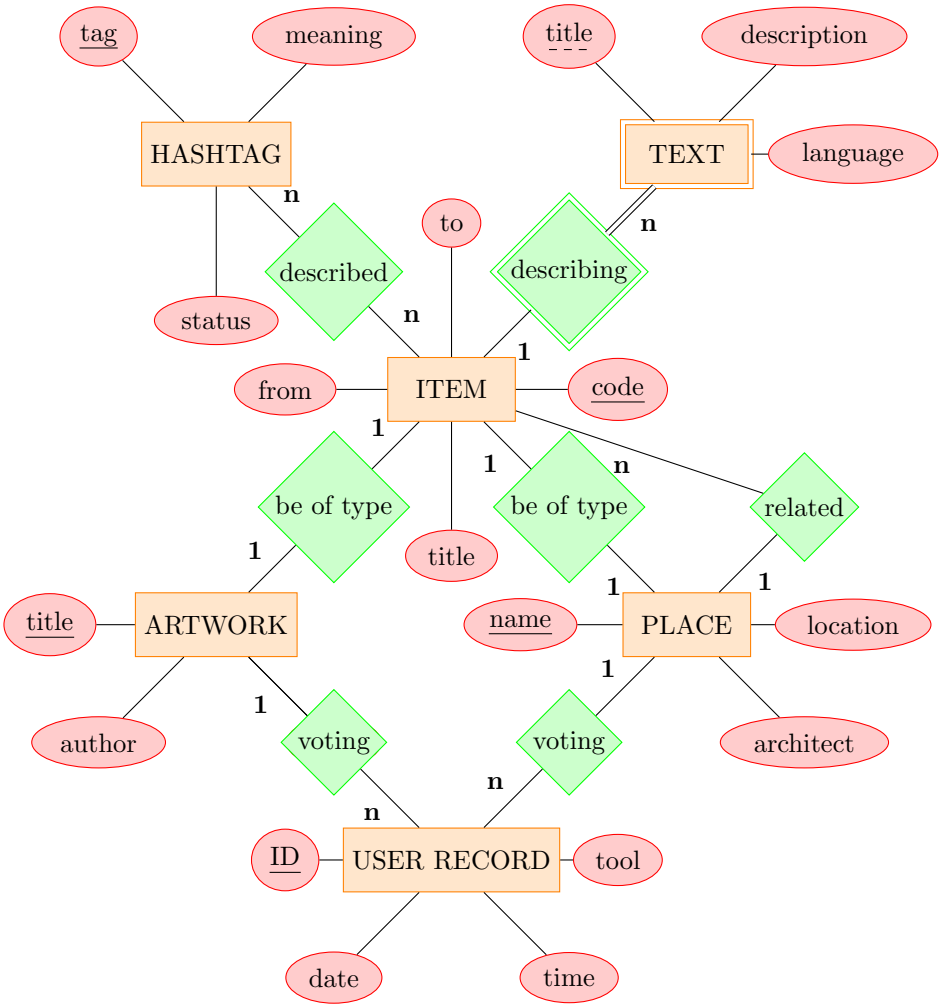
(c) Items could be a **place**, like an historical building (castle, palace, church, ...) a garden, a natural park, described in the database by means of a common name, the location, the architect - if known.

(d) Alternative items could be **artwork**, described in the database by means of a title, author and a year of realization.

(e) It could happen an item, that is a building, be related to another item that is a garden or a painting or other ...

(f) For items both, is available a detailed **text** describing it. Specifically multi-language texts are allowed and available.

(g) Everyone can vote by means of whatsapp, SMS, Twitter and Facebook. Any vote received through the listed tools is converted into a **user record**: an automatic ID is assigned, the tool used, and when (date + time). The relationships between a user record and a specific item (place or artwork) represents **one vote**.



(2) **(Optional: 5 points)**. Write the SQL statement to `CREATE` the "relation" that describes all **texts**.

```
CREATE TABLE Text(
   title VARCHAR(100),
   description VARCHAR(5000),
   language VARCHAR(30),
   itemCode CHAR(5),
   FOREIGN KEY (itemCode) REFERENCES item(code),
   PRIMARY KEY (title, itemCode)
);
```

PARTE B

Es. 3 - SQL (45 points) Let assume the database "online-market".

(1) Region(name, description)

(2) Sheet(ID, description, Region.name)

(3) Producer(name, description)

(4) Produced(Producer.name, Sheet.ID)

(5) Ingredient(name, description)

(6) Made(Ingredient.name, Sheet.ID)

(7) Menu(name, description, main)

(8) Food(name, unit, weight, label, price, startDate, endDate, Menu.main_name, Sheet.ID)

(9) GiftBasket(name, description)

(10) BasketCombines(GiftBasket.name, Food.name, Food.unit, Food.weight)

(11) User(ID, date, time, network_info)

(12) Consulted(User.ID, Food.name, Food.unit, Food.weight, time)

(13) Selected(User.ID, Food.name, Food.unit, Food.weight, time, quantity)

Questions

1) In order to update the web site to be more informative, marketing needs to know for which producer we have the largest number of food products, specifically their descriptive sheets (only the ID). [**Tip:** Firstly identify the producer who produced the largest number of food products for which sheets exist]. The best strategy uses sub-queries.

   [*Sol.*]

```
SELECT sheet_ID
  FROM produced
 WHERE producer_name = (SELECT producer_name
                          FROM produced
                      GROUP BY producer_name
                      HAVING COUNT(*) >= ALL (SELECT COUNT(*)
                                                FROM produced
                                            GROUP BY producer_name));
```

2) For similar purposes of the previous question, marketing needs to know for which food products there exist descriptive sheets and they are produced in **Piemonte**, **Sicilia**. Return the food product identifier and further compute for how many years this product is in the virtual shop [**Tip**: to compute years use a suitable expression.]

[*Sol.*]

```
SELECT name, unit, weight, year(now()) - year(startDate) AS years
FROM food
WHERE NOT EXISTS (SELECT *
                    FROM sheet
                   WHERE ID = food.sheet_ID
                     AND region_name IN ('Piemonte', 'Sicilia'));
```

3) Report how many products have been selected (**not the quantity of items!!!**) for a purchase, distinct for year and month. Show the year, the month, and the number of selections, and the most expensive food price. **Use the explicit JOIN if it is necessary to join tables.** [**Tip**: the date is available in `USER` relation.]

[*Sol.*]

```
SELECT year(user.date), month(user.date), COUNT(food_name) AS nr, MAX(price) AS ex
  FROM food JOIN selected ON (food.name = selected.food_name
                              AND food.unit = selected.food_unit
                              AND food.weight = selected.food_weight)
            JOIN user ON (selected.ID = user.ID)
GROUP BY year(user.date), month(user.date);
```