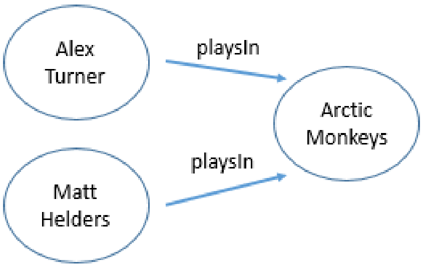
SDM / OD EXAM

8th of June 2018. *The exam will take 2 hours. Answer each question in the provided space.* Answers out of such space will not be considered. You are only allowed to have a pen and the papers provided by the lecturers on the table.

Name: …….

Question 1. [2p]

Given the following conceptual graph:



Two people decide to implement this information, one as a property graph and the other one as a knowledge graph. After doing so, they trigger the following queries:

Q1. Property graph (Cypher query)

```
MATCH (s1:Singer) – [:playsIn] – (b:Band) – [:playsIn] – (s2:Singer)

RETURN s1.name, s2.name
```

Q2. Knowledge graph (SPARQL query)

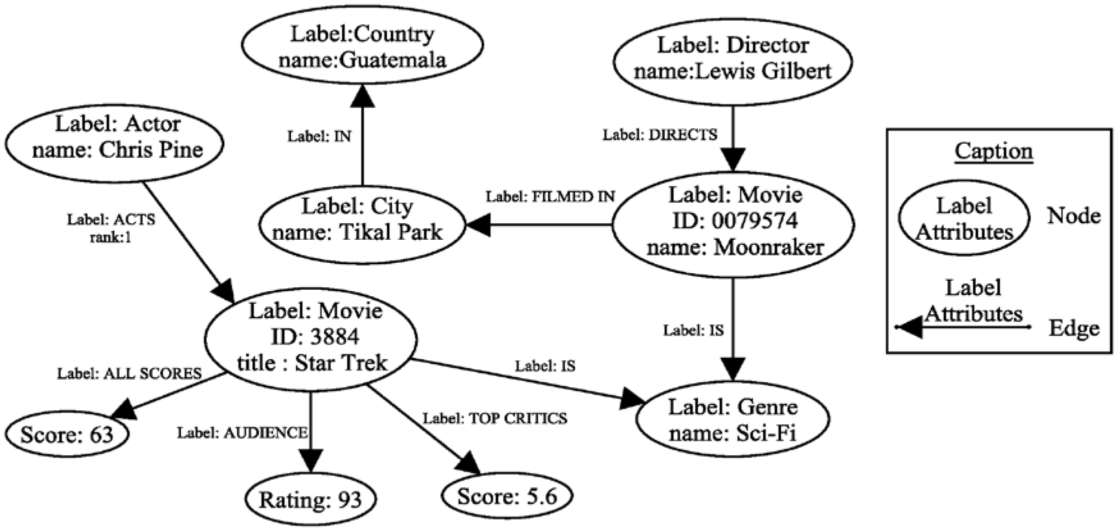
```
SELECT ?a ?b

WHERE { ?a ex:playsIn ?c .

      ?b ex:playsIn ?c . }
```

a) Besides the obvious differences (one query will retrieve strings and the other one IRIs), will the answers to these queries be the same? **Justify your answer**.

b) Rewrite the previous Cypher query into an equivalent navigational graph pattern (i.e., by using regular expressions to describe the path between s1 and s2). Since Cypher does not support most regular expressions on paths, you are allowed to use any of them even if not available in Cypher.

Question 2. [2p]

Given the following representative sample of a property graph showing its most relevant node and edge labels and properties:



What **graph operation(s)** would you need to answer each of the following queries over the whole graph? **<u>Justify your answer</u>**:

- Who is the most successful director (i.e., director of best rated movies) by the audience?

……….

……….

……….

……….

- How many movies have been shot in Barcelona? ………………………………………………………………

……….

……….

……….

- List all movies with two or more directors ………………………………………………………………………

……….

……….

……….

- What is the movie with highest betweenness among actors? …………………………………………….

……….

……….

……….

What is the meaning of this query? ……………………………………………………………………………………

……….

……….

Question 3. [5p]

We want to build a data integration system virtually integrating data from several sources. **Furthermore, we want data to stay in their original placement.** Thus, we first generate a TBOX representing data from each source. From these source TBOXes we will generate an integrated TBOX. One of the sources is an Open Data resource provided by Barcelona’s Town Council about the city bike sharing service (aka as *bicing*). When querying the available API, the information is retrieved in JSON format. A representative JSON example from this source follows:

```
{
   “id”: 1,
   “type”: BIKE,
   “latitude”: 41.397952,
   “longitude”: 2.180042,
   “streetname”: “Gran Via de les Corts Catalanes”,
   “streetnumber”: 760,
   “altitude”: 21,
   “slots”: 19,
   “bikes”: 8,
   “status”:”OPN”,
   “nearbyStations”: [24, 369, 387, 426]
}
```

a) Use RDFS to represent a **TBOX modelling the embedded schema in the above JSON**. Draw it as an RDFS graph. Draw all TBOX concepts within rectangles. Assume the regime entailment is on. Furthermore, your TBOX must capture as much semantics as possible. Clearly identify in the graph the RDFS constructs and define your own namespace prefix for the URIs you need to create. Note that since the regime entailment is activated there is no need to specify metamodel constructs. **[1,5p]**

b) The next source you want to deal with is a similar API but for bus stations. Assume the TBOX for this source has been created and its main concept is bcn:busStation. Now, in the integration layer, you want to state that bcn:Station is a superclass of bcn:busStation and the “*bike station*” concept you created in the previous exercise. This generalisation must be complete and disjoint. Model this information in RDFS. **[1,5p]**

Express in DL those semantics not captured with RDFS.

What would be the equivalent OWL statements to the DL axioms you described above?

c) Finally, assume all the sources have their own TBOX and they have been integrated into a single knowledge graph. Now, it is time to write the mappings to relate the physical instances to the ontology created. Besides the bike stations and bus stations, you want to integrate metro stations (coming from Barcelona's Open Data portal), Tweets (author, location, tweet, hashtags) and Instagram posts (author, picture, description, hashtags) mentioning alterations in the public transport service in Barcelona. In such setting, would you go for LAV or GAV mappings? **Justify your answer**. **[0,5p]**

d) Instead of going for a virtual integration scenario, we could alternatively materialize the source instances as an ABOX for each source TBOX. Consider the TBOX you created for item *a* and assume a correct ABOX has been generated for it. Write the SPARQL query (it must compile) retrieving "*the total number of bikes in Gran Via de les Corts Catalanes*" **[0,75p]**

e) Following with the materialized approach, we want to store each source ABOX and TBOX in Virtuoso as separated named graphs. All TBOX triples (respectively all ABOX triples) are stored in a graph. In order to speed up the previous query, which index would you create on the ABOX? **[0,75p]**

With the index you chose, would Index-Only Query Answering hold? **Justify your answer**.

Question 4. [1p]

Answer the following questions about knowledge graphs:

a) RDFS has powerful constructs that, in certain cases, may generate basic set theory problems such as the Russell's paradox. Justify what design actions you need to undertake when designing an RDFS knowledge graph to guarantee your RDFS knowledge graph is indeed a correct ontology. **[0,5p]**

……

……

……

……

b) Is RDFS a subset of OWL (respectively, of DL)? **Justify your answer**. [**0,5p**]

……

……

……

……