Semantic DATA :
High-level questions to prepare for the oral exam.
June 2021.

Reminder from the modalities of the oral exam.

The exam counts for 50% of the note, the project counting for the other 50%. Only those students having submitted a project are allowed to present the exam.

The exam will be an individual oral discussion*, either in remote or in presential depending on the circumstances. By preference it will be done in presential, if possible. If in remote, the exam will use by preference Collaborate.*

Students will be asked first to present one main question out of a list of high-level questions (this list). For this main question, the students will be allowed to use the relevant slides from the theory of the course and will be asked to explain them. They may also use any notes they may have prepared about the high-level questions.

Students will then be asked oral complementary questions testing their understanding and their capability to make links between different chapters of the course. One of these complementary questions will be related to one of the case studies (cf. also this list).

These complementary questions are not limited to the topic of the main question mentioned above but may range over the whole material, ***and in order to provide answers in timely fashion, students are still required to study the main ideas of the course*** *and to be able to answer them without taking time to research their notes.*

Note : in the same spirit, slide indications mentioned for each high-level questions are indications only, meant to focus the discussion, which does not suppress the need to understand and be able to use what is said in other slides.

Typical examples of complementary questions concern the definition, use, key characteristics, and limitations of the concepts seen in the course and their relationship to other concepts. Students may also be asked to resolve or explain an example covered in the theory section of the course, and when doing so, are also expected to be able to talk about the concepts illustrated by that example. What will not be included in the complementary questions are formal developments outside of the scope of the focus slides indicated for each high-level question.

The ***required material for the exam*** *covers chapters 1 to 11 of the theory and the two case studies. To take into account the Covid situation:*
- the material from the practice sessions is suppressed from the material for the exam (but still useful for the project). However, as said, examples from the theory part of the course are still part of the exam.

- some specific sections or slides have been excluded from the theory chapters. They are explicitly identified at the start of each chapter (starting with chapter 6).

Question 1

Explain the challenges/issues of knowledge representation in a semantic network, including the difficulties of multiple defeasible inheritance and one heuristic to handle it (focusing on chapter 1 slides 39 to 43, 46, 47).

Question 2

Explain the principles of model-theoretic semantics and the decidability characteristics of propositional logic and first order logic (focusing on chapter 2, slides 24, 28, 29, 47, 48, 49).

Notes :

1) That question implies to master the concepts of satisfiability, validity and entailment.
2) The concepts included in that question may be needed in support for answering other questions.

Question 3

Explain the nature and usage of the languages of the semantic web stack (RDF, RDFS, OWL, SPARQL) (chapter 3, slide 5). Then introduce the main characteristics and limitations of RDF and RDFS based on the examples in the course (focusing on chapter 3, slides 23, 24, 37, 43, 60, 61).

Notes :

1) That question also requires to be able to contrast the use of RDF, RDFS and OWL.

Question 4

Explain the general architecture of description logics, and the different contents that can be included in each element of that architecture (focusing on chapter 4, slides 11, 12, 17, 18, 19, 34, 35).

Question 5

Explain why description logics are a family of languages and what varies between the different languages of that family. Present the basic logic *ALC*, the principle of defining its semantics by mapping it into first order logic and discuss the decidability of description logics (focusing on chapter 4, slides 22, 26, 27, 36, 39, 40).

Question 6

Explain the role, foundations and limitations of the language OWL, and position its different variants and profiles, as well as the tradeoff between OWL Full and OWL DL. Explain on the example how OWL can support more inferences than RDFS (focusing on chapter 6 slides 4, 5, 7, 8, 26, 27, 34).

Question 7

Explain the principles of graph querying and the syntax and semantics of basic graph patterns in SPARQL. Discuss the impact of different entailment regimes on the results of SPARQL queries (focusing on chapter 7 slides 9, 11, 19, 21, 23, 24).

Question 8

Explain the main principles of the tableau decision algorithm for description logics, including TBox elimination, input normal form, structure of the algorithm and the rules it uses. Also explain the types of decision problems it can answer (focusing on chapter 8 slides 24, 25, 27, 29, 30, 31, 38).

Question 9

Explain the approach of ontology-based data access based on query rewrite, including its general principles, problems to solve, formalisms being used (focusing on chapter 9, slides 51, 52, 54 plus explanation of the choice of each language, then 59, 60). Taking the input from case study 1 into account, explain the main advantages for the users.

Questions on case studies

One question will always be asked either on case study 1 (Statoil) or 2 (Montefiore).
In addition complementary questions may be asked on the EPIM case study.

For case studies 1 and 2, be able to explain the main points of the case study from the point of view of the analysis grid of the course.

For case study 1 (Statoil), be able in addition to explain the delays and costs involved in modifying relational databases to support new types of data and data analyses.

For case study 2 (Montefiore), be able in addition to explain how semantic graphs address issues of data silos and how ontologies are put in use in biogenetics and medical care to ensure a semantic normalization of data facilitating data interchange and analysis.

For the EPIM case study, be able to explain the advantages of RDF-based data integration relying on triple stores, and the advantages of NoSQL databases in general in complex integration situations when there is a strong need to cope with changes.