LINKÖPING UNIVERSITY
Institutionen för datavetenskap
Patrick Lambrix / Valentina Ivanova

Tentamen

DF22300 Advanced Data Models and Databases
2014

Grades:

For a pass grade all questions need to be answered and almost all correctly answered.

For the 4.5 credits course: answer questions 1-5.

For the 6 credits course (in case you did not take the ontology engineering course): answer all questions.

Instructions:

- This is an *individual* exam. If you have questions, ask the course leader.
- Send your answers to: Patrick Lambrix, Institutionen för datavetenskap, Linköpings universitet, 581 83 Linköping.
- Deadline: February 1, 2015.
- Write a contact phone number/e-mail address on your hand-in. I will phone you in case there are questions regarding your answers.
- If there are references to papers in the questions, then links to the papers will be on the course home page.

Question 1. Semi-structured data

(i) Below, a (simplified) description of the biological data source PIR-PSD, is given. This data source contains information about protein sequences. Each entry in the data source has information about the entry (identification numbers, dates of creation and update), names of the protein sequence and the parts that it contains, the organism from which the sequence was taken, references to the literature describing the sequence, references to other related databases and the protein sequence. The simplified PIR-PSD schema is given as a DTD. Draw a strong data guide for PIR-PSD. Use the OEM model.

```
<!-- *****************************************************************
PIR-International Protein Sequence Database (PSD)
***************************************************************** -->

<!-- ProteinEntry: the root element. -->
<!ELEMENT ProteinEntry (header,protein,organism,reference*,sequence)>

<!-- header: database information. -->
<!ELEMENT header (uid,accession*,created_date,seq-update_date)>

<!ELEMENT accession (#PCDATA)> <!-- accession number -->
<!ELEMENT created_date (#PCDATA)> <!-- date (DD-MMM-YYYY) -->
<!ELEMENT seq-update_date (#PCDATA)> <!-- date (DD-MMM-YYYY) -->

<!-- protein: the protein-names. -->
<!ELEMENT protein(name,alt-name*,contains*)>

<!ELEMENT name (#PCDATA)> <!-- protein name -->
<!ELEMENT alt-name (#PCDATA)> <!-- alternate protein name -->
<!ELEMENT contains (#PCDATA)> <!-- activity name -->

<!-- organism: identification of the biological source. -->
<!ELEMENT organism (source,common-name?,note*)>

<!ELEMENT source (#PCDATA)> <!-- source name -->
<!ELEMENT common-name(#PCDATA)> <!-- common name -->
```

```
<!-- reference -->
<!ELEMENT reference (refinfo,note*)>

<!-- refinfo: identification of the literature source. -->
<!ELEMENT refinfo (authors,citation,title?,xrefs?)>

<!ELEMENT authors (author+)> <!-- list of authors -->
<!ELEMENT author (#PCDATA)> <!-- author name -->
<!ELEMENT citation (#PCDATA)> <!-- citation name -->
<!ELEMENT title (#PCDATA)> <!-- title text -->

<!ELEMENT xrefs (xref+)> <!-- cross-references -->
<!ELEMENT xref (db,uid)> <!-- a cross-reference -->
<!ELEMENT db (#PCDATA)> <!-- database tag -->

<!-- sequence: the amino acid sequence. -->
<!ELEMENT sequence (#PCDATA)> <!-- amino acid symbols and
 punctuation -->

<!-- General elements. Elements that can be contained in several
 other elements. -->
<!ELEMENT note (#PCDATA)> <!-- note text -->
<!ELEMENT uid (#PCDATA)> <!-- entry identifier -->
```

(ii) In figures 1 and 2 (simplified and slightly modified) data from a SwissProt data source is given. (The dashed nodes in the figures refer to nodes in the other figure. Both figures together make up the data source.) Draw a strong data guide for this data. Use the OEM model.

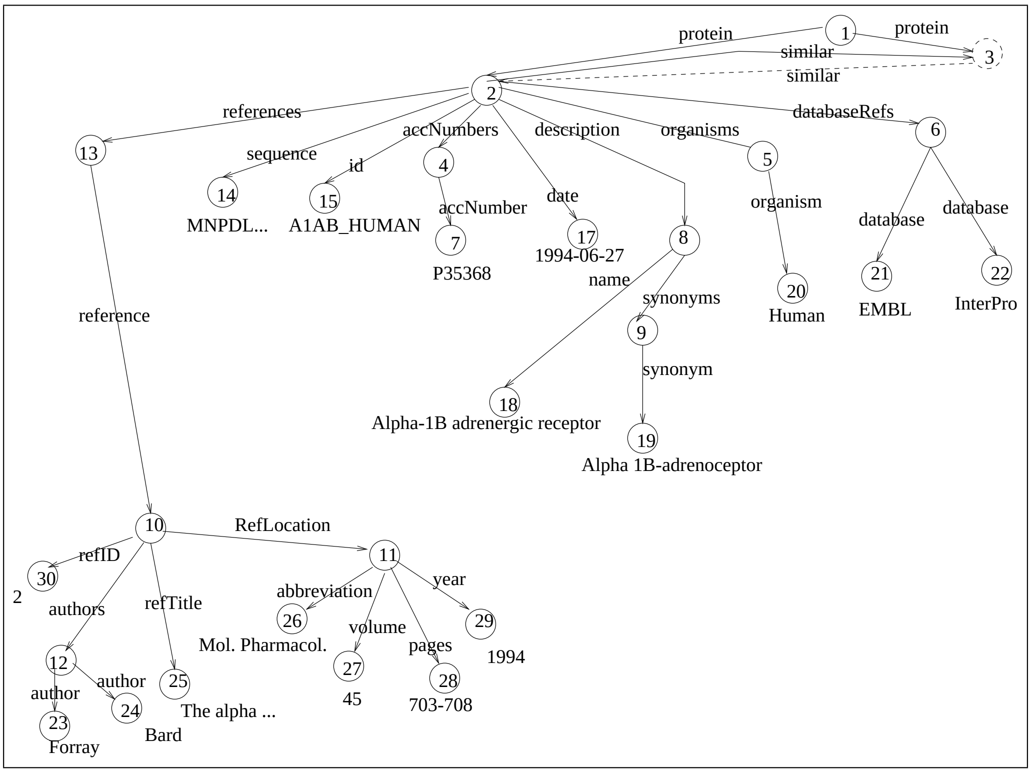

Figure 1: SwissProt db -1

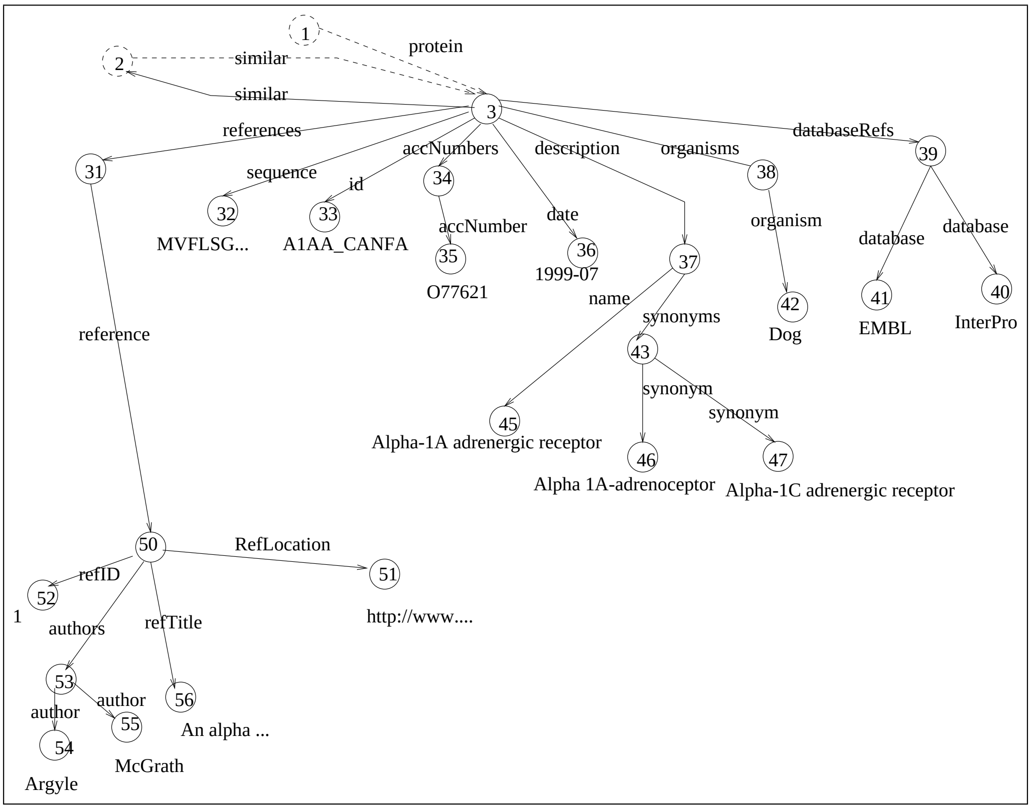

Figure 2: SwissProt db -2

SwissProt concept	PIR-PSD concept
id	uid
accNumbers	accession
date	created_date
description	protein
synonym	alt_name
refTitle	title
refLocation	citation
databaseRefs	xrefs
database	db

Figure 3: Matching concepts.

Question 2. XML and databases

Given the schema in question 1 (i).

(i) Use XPath to find all protein entries taken from the organism with common name 'human'.
(ii) Use XQuery to find the accession number, name and creation date for all protein entries that are referenced by articles that also reference the protein with accession number NM_000684.

Question 3. Integration of data sources

(a) *Global model*

Draw a global domain model based on your data guides from question 1. Use the information in figure 3 about which concepts in SwissProt match to concepts in PIR-PSD. Assume that concepts with the same name in the two data sources match. When labeling, give priority to SwissProt terminology.

(b) *Global as view and local as view*

Discuss the global as view approach versus the local as view approach regarding the mappings between the global model and the content of the local sources as well as regarding query processing. *Exemplify* using examples based on the scenario in question 1 and question 3(a). (If needed, you may extend the scenario or the data source descriptions.)

Question 4: NoSQL databases

(i) Discuss the C, A and P in the CAP Theorem and its connections to the ACID and BASE properties according to the following sources:
http://www.cs.berkeley.edu/∼brewer/PODC2000.pdf
http://www.cs.berkeley.edu/∼brewer/cs262b/TACC.pdf

(ii) Summarize the following article ”CAP Twelve Years Later: How the ”Rules” Have Changed” (http://www.infoq.com/articles/cap-twelve-years-later-how-the-rules-have-changed).

Question 5: Information Retrieval

(i) Assume the boolean model for information retrieval. Assume we are interested in the words ’gene’, ’enzyme’, ’protein’ and ’signal’.

a. Show how to represent documents in the boolean model.

b. Represent the query for all documents containing gene or enzyme, but not protein. Show the completed DNF (disjunctive normal form) of the query.

(ii) Explain tf and idf in the vector model.

FROM HERE ON: ONLY FOR THE 6 CREDITS COURSE:

Question 6: Description Logics

Define the following concepts using description logics:

C1: animals that have at least 4 legs and all foods they eat are wheat plants
C2: animals that have at least 2 legs and all foods they eat are plants

Does C2 subsume C1, i.e. C1 isa C2? Prove your answer using a tableau algorithm. (Use the Baader, Nutt chapter in the reading list.)

Question 7: DB vs KB

Describe the difference between open-world assumption and closed-world assumption. Give examples.

Question 8: Ontology Engineering

(i) Give 4 different kinds of matchers for ontology alignment. For each kind of matcher give an example and explain briefly what it does.

(ii) Given the following axioms in an ontology: $A \sqsubseteq B$; $A \sqsubseteq E$; $B \sqsubseteq C$; $B \sqsubseteq D$; $C \sqsubseteq F$; $D \sqsubseteq F$; $F \sqsubseteq G$; $E \sqsubseteq G$; $A \sqsubseteq \neg G$.

Debug the ontology. Compute MIPS and MUPS.