Web of Data – exam*

Paper documents and notes are allowed - duration: 2:30

Computer Science Master 2 – Data Science – Paris Saclay University

16^{th} of December, 2024

Part 1: Data Linking *[12pts]*

A- General course questions (4pts.)

- *Question 1 (2 pts).* Give three main families of data linking approaches and, for each, give its main characteristics.
- *Question 2. (2 pts)* What are the main aspects that may be considered for evaluating data linking approaches.

B- Data Linking application exercises (8pts.)

Let consider two datasets $D1$ and $D2$ given in tables 1 and 2 respectively that conform to the ontology of Figure 1.

They describe some aquariums and zoos in France. They are described by six properties {`name, city, region, opening-date, #animal, lang`}. We note that the property *lang* is multi-valued. We consider the following axioms as fulfilled in the ontology $O1$, for the class `AnimalConservatory`:

*The mark scale is given as an indication.

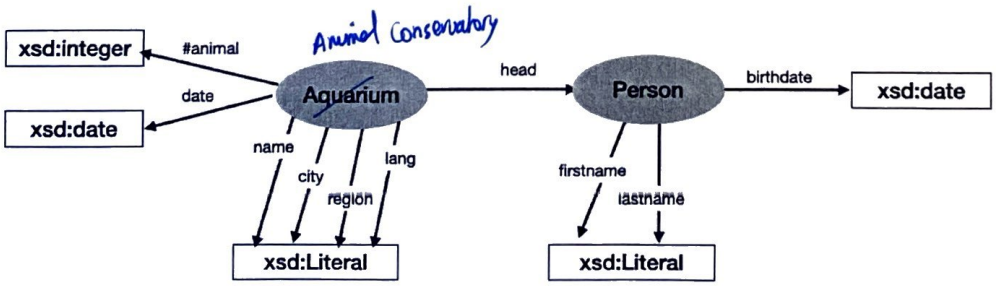

Figure 1: O1: Aquariums and zoos Ontology

- `PFI(name,date)`, `PFI(name, city, region)`
- `PF(head)`, `PF(city)`, `PF(date)`, `PF(#animal)`

	name	city	region	date	#animal	lang	head
a_1	Aquarium du Périgord noir	Le Bugue	Nouvelle-Aquitaine	1933	6000	fr,en	:d1
a_2	Biotropica	Val-de-Reuil	Normandie	2012	1300		
a_3	Réserve ornithologique du Teich	Le Teich	Nouvelle-Aquitaine	1972	600	fr	:d3
a_4	Parc'Ours	Borce	Nouvelle-Aquitaine	1971	150		
a_5	Montagne des singes	Kintzheim	Grand Est	1969	240	fr, de, en	:d5
a_6	Muséum-aquarium de Nancy	Nancy	Grand Est	1933		fr,en	:d6
a_7	Parc zoologique de Paris	Paris	Île-de-France	1933	2800	fr,en,es	:d7
a_8	Ménagerie du Jardin des plantes	Paris	Île-de-France	1794	600	fr,en	

	firstname	latsname	birthdate
:d1	Patrick	Bourgeois	1964
:d3	Nicolas	Charnois	1976
:d5	Marie	Funch	1982
:d6	Claire	Reynaud	1973
:d7	Pierre	Ferré	1993

Table 1: Extract of aquariums and zoos descriptions (D1)

	name	city	region	date	#animal	lang	head
a_{21}	Aquarium du Périgord noir	Le bugue	Nouvelle-Aquitaine	1933	6000	fr,en	:d21
a_{22}	Biotropica	Val-de-Reuil	Normandie	2012	1300		:d22
a_{23}	Réserve ornithologique du Teich	Le Teich	Nouvelle-Aquitaine	1972	600	fr	
a_{24}	Parc'Ours	Borce	Nouvelle-Aquitaine	1971	150		:d24
a_{25}	Montagne des singes	Kintzheim	Grand Est	1969	240	fr,de,en	:d25
a_{26}	Parc zoologique de Paris	Paris	IDF	1933	2800	fr,en,es	

	firstname	latsname	birthdate
:d21	Patrick	Bourgeois	
:d22	Anthony	Denise	1997
:d24	Clémentine	July	1977
:d25	Martin	Sebag	1975

Table 2: Extract of aquariums and zoos descriptions (D2)

We consider that for each pair of values of the common properties that are equal from D1 and D2, we have a fact of the predicate $synVals$:

```
synVals(''Aquarium du Périgord noir'', ''Aquarium du Périgord noir'')
synVals(''Le Bugue'', ''Le bugue'')
synVals(''1969'', ''1969'')
synVals(''Île-de-France'', ''IDF'')
 ...
```

Question 3 (6 pts). Considering these synVals facts, the data descriptions from D1 and D2 and the ontology axioms of the class `AnimalConservatory` give what are the `owl:sameAs` and `synVals` facts that can be obtained if one applies the L2R method (logical method for reference reconciliation)? You are asked to jutifiy the deduced facts.

Question 4 (2 pts). If you consider these axioms of the class `Person`:

- `PF(firstname)`, `PF(lastname)`, `PF(birthdate)`

what are the facts that can be inferred using L2R method?

Part 2: Link invalidation (5pt)

- *Question 5.* According to the research studies conducted on the Web of data, what is the range of the percentage of the number of erroneous identity links in the Web of Data?

- *Question 6.* The linking invalidation approach proposed by [Raad et al 18] relies on the computation of an error degree for each identity link in the network. This computation is based on the density of the communities to which it belongs.

  By using the formulas presented in course 3 and given in equations (1) and (2), considering $w = 2$ and the two communities of C_1, C_2 of Figure 2, give (while detailing the computation) the error degrees for the links : e_{C_1}, e_{C_2} and $e_{C_{12}}$

$$err(e_C) = \frac{1}{w(e_C)} \times (1 - \frac{W_C}{|C| \times (|C| - 1)}),\ where\ W_C = \sum_{e_C \in E_C} w(e_C) \qquad (1)$$

$$err(e_{C_{ij}}) = \frac{1}{w(e_{C_{ij}})} \times (1 - \frac{W_{C_{ij}}}{2 \times |C_i| \times |C_j|}),\ where\ W_{C_{ij}} = \sum_{e_{C_{ij}} \in E_{C_{ij}}} w(e_{C_{ij}}) \qquad (2)$$

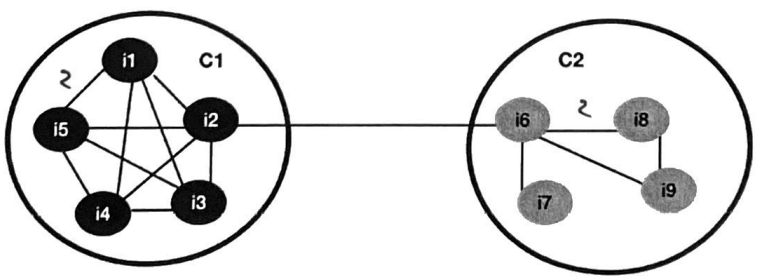

Figure 2: Identity communities C_1 and C_2

Part 3: Ontology Alignment (3pt)

- *Question 7.* Given the ontology alignment problem shown in Figure 3, explain what the inputs `parameters` and `resources` represent for an ontology alignment tool and give an example of each of them.

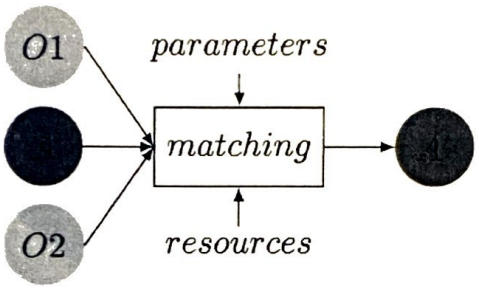

Figure 3: Ontology Alignment problem

- *Question 8.* Give three kinds of heterogeneities in ontologies that can be faced when dealing with ontology alignment.