Bases de Données II, Charleroi, 10 janvier 2022

NOM + PRÉNOM :	
Orientation + Année :	

Cet examen contient 7 questions.

Le document XML de la Figure 1 est utilisé pour stocker le palmarès des cyclistes les plus importants de l'histoire du cyclisme. La DTD est incluse au début du document XML. Par exemple, l'élément suivant indique qu'Eddy Merckx est né en 1945 et qu'en 1975, il a gagné (`classement="1"`) la course abrégée comme MSM.

```
<cycliste nom="Eddy Merckx" naissance="1945">
   <participation cid="MSM" annee="1975" classement="1"/>
</cycliste>
```

L'élément suivant indique que MSM est un raccourci pour la course Milan-San Remo :

```
<course cid="MSM">Milan-San Remo</course>
```

Dans chaque course, les coureurs sont classés 1, 2, 3, ... sans ex æquo (i.e., deux coureurs ne peuvent pas avoir le même rang).

Question 1 Écrivez une requête en XPath qui renvoie le nom de chaque cycliste qui n'a jamais participé à Paris-Roubaix. **Il n'est pas permis d'utiliser des fonctions d'agrégation telles que count, min et max. La chaîne de caractères "PR" ne peut pas apparaître dans votre requête ; la recherche doit se faire sur la chaîne "Paris-Roubaix".** Évitez l'usage des axes *parent* et *ancestor*.

Pour le document XML de l'exemple, la réponse est comme suit :

```
nom="Francesco Moser"
nom="Walter Godefroot"
```

.../5

Question 2 Écrivez une requête en XPath qui renvoie le nom de chaque cycliste qui a participé une ou plusieurs fois à une même course. **Il n'est pas permis d'utiliser des fonctions d'agrégation telles que count, min et max.** Évitez l'usage des axes *parent* et *ancestor*.

Pour le document XML de l'exemple, la réponse est comme suit :

```
nom="Roger De Vlaeminck"
nom="Eddy Merckx"
nom="Francesco Moser"
```

En effet, Roger De Vlaeminck et Eddy Merckx ont participé deux fois à Liège-Bastogne-Liège ; Franscesco Moser a participé deux fois à Milan-San Remo. Walter Godefroot n'est pas dans le résultat, car il n'a jamais participé deux fois à une même course.

.../5

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE PALMARES [
<!ELEMENT PALMARES (COURSES, CYCLISTES)>
<!ELEMENT COURSES (course)*>
<!ELEMENT CYCLISTES (cycliste)*>
<!ELEMENT cycliste (participation)*>
<!ELEMENT course (#PCDATA)>
<!ELEMENT participation (#PCDATA)>
<!ATTLIST course cid CDATA #REQUIRED>
<!ATTLIST cycliste nom CDATA #REQUIRED>
<!ATTLIST cycliste naissance CDATA #REQUIRED>
<!ATTLIST participation cid CDATA #REQUIRED>
<!ATTLIST participation annee CDATA #REQUIRED>
<!ATTLIST participation classement CDATA #REQUIRED>
]>

<PALMARES>
<COURSES>
   <course cid="LBL">Liège-Bastogne-Liège</course>
   <course cid="MSM">Milan-San Remo</course>
   <course cid="PR">Paris-Roubaix</course>
</COURSES>
<CYCLISTES>
   <cycliste nom="Roger De Vlaeminck" naissance="1947">
      <participation cid="LBL" annee="1970" classement="1"/>
      <participation cid="PR" annee="1972" classement="1"/>
      <participation cid="LBL" annee="1975" classement="8"/>
   </cycliste>
   <cycliste nom="Eddy Merckx" naissance="1945">
      <participation cid="LBL" annee="1970" classement="3"/>
      <participation cid="PR" annee="1972" classement="7"/>
      <participation cid="LBL" annee="1975" classement="1"/>
      <participation cid="MSM" annee="1975" classement="1"/>
   </cycliste>
   <cycliste nom="Francesco Moser" naissance="1951">
      <participation cid="MSM" annee="1975" classement="2"/>
      <participation cid="MSM" annee="1984" classement="1"/>
   </cycliste>
   <cycliste nom="Walter Godefroot" naissance="1943">
      <participation cid="LBL" annee="1975" classement="3"/>
      <participation cid="MSM" annee="1975" classement="30"/>
   </cycliste>
</CYCLISTES>
</PALMARES>
```

FIGURE 1 – Les palmarès des cyclistes.

Question 3 Écrivez un programme en XSLT qui renvoie les gagneurs (i.e., les cyclistes avec `classement="1"`) des différentes éditions des courses. Les résultats doivent être groupés par course, comme suit :

```
<WINNERS>
   <RACE name="Liège-Bastogne-Liège">
      <EDITION year="1970">Roger De Vlaeminck</EDITION>
      <EDITION year="1975">Eddy Merckx</EDITION>
   </RACE>
   <RACE name="Milan-San Remo">
      <EDITION year="1975">Eddy Merckx</EDITION>
      <EDITION year="1984">Francesco Moser</EDITION>
   </RACE>
   <RACE name="Paris-Roubaix">
      <EDITION year="1972">Roger De Vlaeminck</EDITION>
   </RACE>
</WINNERS>
```

Notez que les balises sont en anglais.

.../10

Question 4 Écrivez une requête en XQuery qui classe les coureurs selon leur nombre de victoires, dans le format suivant :

```
<NOMBRE-DE-VICTOIRES>
  <COUREUR nom="Roger De Vlaeminck" victoires="2"/>
  <COUREUR nom="Eddy Merckx" victoires="2"/>
  <COUREUR nom="Francesco Moser" victoires="1"/>
  <COUREUR nom="Walter Godefroot" victoires="0"/>
</NOMBRE-DE-VICTOIRES>
```

En XQuery, le mot clé `descending` est utilisé pour trier en ordre descendant, par exemple,

`order by $unNombre descending`.

. . . /10

Question 5 Pour un problème de classification à deux classes (disons “yes” et “no”), l’exactitude (*accuracy*) d’un classificateur est défini comme $\frac{\mathsf{TP}+\mathsf{TN}}{\mathsf{P}+\mathsf{N}}$. Si l’exactitude d’un classificateur est faible, ce classificateur n’a pas d’intérêt pratique. L’inverse n’est pas forcément vrai : expliquez pourquoi une exactitude élevée n’indique pas forcément que le classificateur sera utile en pratique. Discutez d’autres mesures pour quantifier la qualité d’un classificateur. Illustrez votre réponse à l’aide d’un **exemple concret**.

.../15

Question 6 Dans la construction d'un arbre de décision pour la table ci-dessous, on souhaite déterminer la valeur de v telle que le gain d'information du test binaire *"Est-ce que $A \leq v$?"* soit maximal. Détaillez une méthode efficace pour déterminer cette valeur de v. Illustrez cette méthode à l'aide de la table ci-dessous.

Identifiant	$\cdots$	Poids	$\cdots$	Classe
1	$\cdots$	20	$\cdots$	yes
2	$\cdots$	41	$\cdots$	yes
3	$\cdots$	21	$\cdots$	yes
4	$\cdots$	40	$\cdots$	yes
5	$\cdots$	25	$\cdots$	no
6	$\cdots$	39	$\cdots$	yes
7	$\cdots$	27	$\cdots$	no
8	$\cdots$	36	$\cdots$	no
9	$\cdots$	27	$\cdots$	no
10	$\cdots$	35	$\cdots$	no
11	$\cdots$	29	$\cdots$	yes
12	$\cdots$	33	$\cdots$	yes
13	$\cdots$	32	$\cdots$	yes

.../15

Question 7 Supposez que dans l'exécution de l'algorithme *A priori*, les ensembles F_3 et F_4 (i.e., les itemsets fréquents de taille 3 et 4) sont comme suit :

$F_3 = \{ABC, ABD, ABE, ABF, ACD, ACE, ADE, ADF, BCD, BCE, BCF, BDE, BDF, CDE, CDF\}$
$F_4 = \{ABCD, ABCE, ABDE, ABDF, BCDE, BCDF\}$

Détaillez une méthode efficace pour déterminer l'ensemble C_5 (i.e., les candidats de taille 5). Illustrez cette méthode à l'aide de l'exemple ci-dessus.

.../10