Arquitectura de bases de dades

Pràctica 2: Extensió XML

Pregunta 1 (3 punts)

Enunciat

Respon cadascuna de les preguntes següents justificant breument la teva resposta.

a) Ens demanen que proporcionem un fitxer XML amb les següents característiques i estructura:

- Que tingui l'estructura següent:

```
root
    COVID
        summary_data (attributes: idregion=43 i descriptionregion=tarragona)
            patients_hospitalised
                men: 17
                women: 4
                average_age: 56
            patients_ICU
                men: 6
                women: 2
                average_age: 64
            acumulated_cases: 3947
        summary_data (attributes: idregion=17 i descriptionregion=girona
            patients_hospitalised
                men: 21
                women: 6
                average_age:58
            patients_ICU
                men: 7
                women: 1
                average_age:67
            acumulated_cases: 7998
    gencat:COVID
        summary (attirbutes: day=12/08/2020): Summary of the latest regional data
```

- Que indiqui que l'encoding utilitzat és el ISO-8859-1

- Que indiqui com a namespace per defecte http://www.uoc.edu/subjects/adb/ns i un d'específic amb dades de la Generalitat al que farà referència amb l'dentificador "gencat" que serà http://www.gencat.cat/dadesobertes/ns

- Que indiqui que el contingut del que es mostra dins el node gencat:COVID no ha de ser interpretat com a XML (donat que prové de fonts externes i no en podem garantir la validesa).

b) Si tenim l'XML següent:

```
<?xml version="1.0" encoding="UTF-8"?>
<COVID id="1234">
   <date>
        <year>2020</year>
        <month>08</month>
        <day>12</day>
   </date>
   <region>
        <idregion>43</idregion>
        <description>Tarragona</description>
   </region>
   <patients>
        <hospitalised>17</hospitalised>
        <ICU>6</ICU>
        <gender>male</gender>
   </patients>
   <patients>
        <hospitalised>6</hospitalised>
        <ICU>2</ICU>
        <gender>female</gender>
   </patients>
</COVID>
```

Defineix el seu l'esquema XSD tenint en compte el següent:
- L'atribut "id" és obligatori i tindrà un valor entre 1 – 9999.
- Del node COVID penjaran els elements *date*, *region* i fins a 10 *patients* (donat que les dades ens les poden proporcionar dividides).
- El node *date* estarà format pels elements *year*, *month* i *day*, i només cal validar que el valor sigui un enter.
- El node *region* estarà format pels elements *idregion* que serà un enter i l'element description que serà un string.
- El node *patients* es pot repetir fins a un màxim de 10 vegades i contindrà els elements *hospitalised* que serà un enter, ICU que serà un enter, i *gender* que podrà prendre els valors *male*, *female* o *other*.

Criteris d'avaluació

Els criteris d'avaluació que s'aplicaran en la correcció d'aquesta pregunta són els següents:

- Totes les preguntes tenen el mateix pes.
- Les preguntes no contestades no penalitzen.
- Les preguntes sense argumentació no seran avaluades.
- Es valorarà la qualitat de la resposta.

Pregunta 2 (3 punts)

Enunciat

Utilitzeu el fitxer XML adjunt a la pràctica (dadesLleida.xml), que conté informació diversa relativa al COVID (des de febrer fins a agost) a la regió de Lleida, una de les més afectades pels rebrots durant l'estiu, per a resoldre cadascun dels apartats següents:

a) Proposeu una sentència XPath que retorni la data d'inici (`start_date`) del període en que hi havia més casos confirmats (`confirmed_cases`) a la regió.

b) Utilitzant la mateixa sentència anterior, obteniu el node que es troba dues posicions abans de l'obtingut amb la sentència, i mostreu un text indicant la diferència de casos confirmats entre les dues setmanes (Ex. "Hi ha hagut un increment de xx casos")

c) L'indicador `r0_confirmat_m` indica la taxa de reproducció d'un virus i el nombre de casos addicionals que pot generar un cas particular. És molt important per a reduir l'expansió del virus que aquest indicador estigui per sota de 1. Obté en quins períodes (`start date - end date`) aquest indicador ha sigut > 1 (agafa únicament les dades que no tenen en compte les dades de residències geriàtriques, és a dir, `residence`=No). La sortida d'aquesta sentència ha de mostrar un llistat on cada línia serà com l'exemple que es mostra a continuació: `3.85714-2020-03-09T00:00:00-2020-03-15T00:00:00`

El contingut dels fitxers XML és el següent:

- Region: conté la comaca a la que fan referència les dades. En aquest exercici avaluem la comarca del Segrià (a la pregunta 3 s'agafaran les dades del Tarragonès, Barcelonès i Gironès).
- IdRegion: identificador numèric de la comarca.
- Start date: data inici del període de mesurament.
- End date: data fi del període de mesurament.
- Residence: camp binari de desagregació de dades: inclou opcions Si (població ingressada en residència geriàtrica) No (població no ingressada en residència geriàtrica)
- R0_confirmat_m: camp numèric amb dades de R0 (El valor R0 (Rho) indica la taxa de reproducció d'un virus i el nombre de casos addicionals que pot generar un cas particular)
- Confirmed_cases_rate: camp numèric amb dades de casos confirmats de COVID-19 (Valor de taxa de casos per 100.000 habitants de cada territori)
- Confirmed_cases: camp numèric amb dades de casos confirmats de COVID-19 (La PCR+ ens indica que és un cas actiu i pel seguiment epidemiològic és important la data en què s'inicia el cas. No usem els casos COVID-19 confirmats per testos ràpids serològics o per ELISA)
- PCR_rate: camp numèric amb dades de Proves PCR realitzades (Aquest camp mostra el valor de la taxa de proves PCR per 100.000 habitants realitzades a cada territori)
- PCR: camp numèric amb dades de Proves PCR realitzades
- Total_hospitalization: camp numèric amb dades d'ingressos registrats (Nombre de nous ingressos hospitalaris per COVID-19 que hi ha hagut durant el període descrit).
- Critical_hospitalization: camp numèric amb dades d'ingressos a UCI registrats (Nombre de nous ingressos a la UCI de pacients amb COVID-19 durant el període descrit).

(Font dades: https://analisi.transparenciacatalunya.cat/Salut/Dades-setmanals-de-COVID-19-per-comarca/jvut-jxu8)

Les sentències XPath de cada apartat han d'anar acompanyades d'una explicació i justificació de la sentència.

Criteris d'avaluació

Els criteris d'avaluació que s'aplicaran en la correcció d'aquesta pregunta són els següents:

- Cal lliurar tots els apartats per tal que l'exercici sigui avaluat.
- No es valorarà l'apartat si no hi ha argumentació o no és possible provar la sentència XPath.
- Es valorarà la qualitat de la resposta, així com l'ús dels conceptes estudiats en el mòdul 2 corresponents a l'extensió XML.
- Les respostes incorrectes no descompten.

Pregunta 3 (4 punts)

Enunciat

a) Quines clàusules podem utilitzar en una consulta XQuery? En què consisteix cadascuna d'elles? (1 punt)

b) Per a resoldre aquest apartat cal que proporcioneu una sentència XQuery que utilitzi tots els fitxers XML adjunts a la pràctica (dadesLleida.xml, dadesTarragona.xml, dadesBarcelona.xml, dadesGirona.xml) per a generar un nou XML. L'objectiu d'aquest XML és per un costat, mostrar unes dades agrupades prenent com a referència les dades del fitxer dadesLleida, i per l'altra, obtenir les dades màximes de cada regió.
A la part d'`AgregatedData`, retornarem la situació a les diferents comarques capitals de província en el moment en que a Lleida es produeix el moment amb màxims contagis (tal i com hem obtingut en l'exercici anterior). El format de sortida ha de ser el següent, de manera que es retorni:

- IsInResidence: contindrà el valor que hi ha al fitxer dadesLleida al node que contingui el número més gran de contagis del fitxer dadesLleida.
- MeasuramentData: contindrà el valor d'`start-date` del node que contingui el número més gran de contagis del fitxer dadesLleida.
- Els dos camps anteriors ens serviran per filtrar els nodes de la resta de fitxers i obtenir el valor de `confirmed_cases` i `confirmed_case_rate`.
- Per a cada fitxer que tractem, crearem el node que correspongui a la capital de comarca LleidaData, TGNData, BCNData i GironaData.

Per altra banda a `IsolatedData`, mostrarem per a cada regió els 5 dies amb més casos confirmats sense tenir en compte les dades de les altres regions. Caldrà obtenir únicament els mostrejos que tinguin Residence="No". Mostrarem els valors `confirmed_cases - start_date`.

```
<COVIDData>
<AgrupatedData IsInResidences="xx" MeasuramentDate="xxxxxx">
     <LleidaData>
       <ConfirmedCases>xxx</ConfirmedCases>
       <ConfirmedCasesRate>xxx</ConfirmedCasesRate>
     </LleidaData>
     <TGNData>
       <ConfirmedCases>xxx</ConfirmedCases>
       <ConfirmedCasesRate>xxx</ConfirmedCasesRate>
     </TGNData>
     <BCNData>
       <ConfirmedCases>xxx</ConfirmedCases>
       <ConfirmedCasesRate>xxx</ConfirmedCasesRate>
     </BCNData>
     <GironaData>
       <ConfirmedCases>xxx</ConfirmedCases>
       <ConfirmedCasesRate>xxx</ConfirmedCasesRate>
     </GironaData>
</AgrupatedData>
```

```
  <IsolatedData>
    <LleidaData>
      <ConfirmedCases>xxx - 2020-xx-xxT00:00:00</ConfirmedCases>
      ...
    </LleidaData>
    <TGNData>
      <ConfirmedCases>xxx - 2020-xx-xxT00:00:00</ConfirmedCases>
      ...
    </TGNData>
    <BCNData>
      <ConfirmedCases>xxx - 2020-xx-xxT00:00:00</ConfirmedCases>
      ...
    </BCNData>
    <GironaData>
      <ConfirmedCases>xxx - 2020-xx-xxT00:00:00</ConfirmedCases>
      ...
    </GironaData>
  </IsolatedData>
</COVIDData>
```

Doneu una explicació i justificació de la sentència XQuery realitzada. (3 punts)

Criteris d'avaluació

Els criteris d'avaluació que s'aplicaran en la correcció d'aquesta pregunta són els següents:

- No es valorarà l'apartat a) si no es detalla com funciona cada clàusula, no serveix únicament enumerar-les.
- No es valorarà l'apartat b) si no hi ha argumentació, no és possible provar la consulta XQuery o el format de la resposta no s'ajusta al demanat.
- Es valorarà la qualitat de la resposta, així com l'ús dels conceptes estudiats en el mòdul 2 corresponents a l'extensió XML.

Recursos

Per tal de resoldre aquesta pràctica caldrà utilitzar els recursos que s'enumeren a continuació:

- Mòdul 2 de l'assignatura Arquitectura de Bases de Dades (únicament la secció 2 que fa referència l'extensió XML)
- Conjunt de dades dadesLleida.xml, dadesTarragona.xml, dadesBarcelona.xml i dadesGirona.xml
- Per resoldre la pràctica podeu utilitzar el processador BaseX (XPath, XQuery processor). El trobareu a : http://files.basex.org/releases/8.6.7/BaseX867.jar

Criteris de valoració

A l'enunciat de cada exercici s'indica el valor del mateix sobre la puntuació total de la pràctica. Aquesta activitat representa el 50% de la nota de pràctiques de l'assignatura i la seva realització és **obligatòria** per tal de superar l'assignatura.

No s'acceptaran ni es tindran en compte els lliuraments realitzats fora dels terminis indicats al calendari de l'aula.

Format i data de lliurament

Caldrà lliurar, a través de la bústia de lliuraments de l'assignatura de l'aula, un arxiu .zip que contingui els fitxers següents:

- Un document PDF que contingui les respostes dels exercicis 1, 2 i 3.
- Un fitxer de text per cada una de les sentències generades en els exercicis 2 i 3. Assegureu-vos que aquests siguin executables, que no esteu utilitzant variables d'entorn i que quan es faci referència als fitxers xml sigui sense paths:

```
doc ("dadesLleida.xml")
```

Tots els documents lliurats hauran de contenir el nom i cognoms de l'alumne, altrament no seran avaluats.

- Lliurament individual: Cal incloure el nom i cognoms a cadascun dels fitxers que lliureu, i cal indicar que es tracta de la PR2.
- Lliurament en parella: Cal incloure el nom i cognom dels dos estudiants a cadascun dels fitxers que lliureu, i cal indicar que es tracta de la PR2. Únicament un dels components ha de fer el lliurament al registre d'avaluació continuada (campus).

La data límit per lliurar la PR2 és el **08/11/2020**.