Données du web

HMIN 103

Données du Web : XML Relationnel

Élèves :
Adel Terki
Emile Youssef

Enseignant :
Federico Ulliana

13 novembre 2018

Table des matières

1 Warm-up, Fonctionnalités Oracle pour XML : Stockage CLOB et Binary XML, Interrogation

```
DROP TABLE tweet_CLOB;
DROP TABLE tweet_binaryxml;
DROP TABLE film_binaryxml;

CREATE TABLE tweet_CLOB (xmlContent varchar(20), tweet CLOB);

CREATE TABLE tweet_binaryxml (xmlContent varchar(20), tweetXml XMLTYPE)
XMLTYPE COLUMN tweetXml STORE AS  BINARY XML;


CREATE TABLE film_binaryxml (xmlContent varchar(20), filmXml XMLTYPE)
XMLTYPE COLUMN filmXml STORE AS  BINARY XML;


INSERT INTO tweet_binaryxml (xmlContent,tweetXml)
VALUES ('tweet', sys.xmltype.createxml('
<tweets>
  <tweet id="t1" author="u1">
    <timestamp GMT="2"> 30 </timestamp>
    <message length="140" type="text" color="black" lang="en">
      <ref_user> example </ref_user>
      <text> absolutely smashed it at </text>
      <hashtags>
        <ref_tag>mtvlivelockdown</ref_tag>
        <ref_tag>I&lt;3XML</ref_tag>
      </hashtags>
      <text> ! Catch him at the official </text>
      <ref_user> clubmtvuk </ref_user>
      <text> after party tonight @10 pm </text>
    </message>
    <responses>
        <response>yo !<date>12/01/18</date></response>
        <response>waw!<date>14/02/18</date></response>
        <response>lets go <date>16/04/18</date></response>
    </responses>
    <gps>
      <latitude> 0 </latitude>
      <longitude> 0 </longitude>
      <altitude> 0 </altitude>
    </gps>
    <url_media>pic.com/TfXd2D3</url_media>
  </tweet>

  <tweet id="t2" author="u2">
    <timestamp GMT="2"> 31 </timestamp>
    <message length="140" type="text" color="black" lang="en">
      <ref_user> example </ref_user>
      <text> absolutely smashed it at </text>
      <hashtags>
        <ref_tag>mtvlivelockdown</ref_tag>
        <ref_tag>efforts</ref_tag>
      </hashtags>
      <text> ! Catch him at the official </text>
      <ref_user> clubmtvuk </ref_user>
      <text> after party tonight @10 pm </text>
    </message>
    <responses>
        <response>yo !<date>12/01/18</date></response>
        <response>waw!<date>14/02/18</date></response>
        <response>let s go <date>16/04/18</date></response>
    </responses>
    <gps>
      <latitude> 0 </latitude>
      <longitude> 0 </longitude>
      <altitude> 0 </altitude>
    </gps>
    <url_media>pic.com/TfXd2D3</url_media>
  </tweet>

  <tweet id="t3" author="u2">
    <timestamp GMT="2"> 32 </timestamp>
    <message length="140" type="text" color="black" lang="en">
      <ref_user> example </ref_user>
      <text> absolutely smashed it at </text>
      <text> ! Catch him at the official </text>
      <ref_user> clubmtvuk </ref_user>
      <text> after party tonight @10 pm </text>
    </message>
    <responses>
        <response>yo !<date>12/01/18</date></response>
        <response>waw!<date>14/02/18</date></response>
        <response>lets go <date>16/04/18</date></response>
    </responses>
    <gps>
      <latitude> 0 </latitude>
      <longitude> 0 </longitude>
      <altitude> 0 </altitude>
    </gps>
    <url_media>pic.com/TfXd2D3</url_media>
  </tweet>

  <tweet id="t4" author="u3">
    <timestamp GMT="2"> 33 </timestamp>
    <message length="140" type="text" color="black" lang="en">
      <ref_user> example </ref_user>
      <text> absolutely smashed it at </text>
      <text> ! Catch him at the official </text>
      <ref_user> clubmtvuk </ref_user>
      <text> after party tonight @10 pm </text>
    </message>
    <gps>
      <latitude> 0 </latitude>
      <longitude> 0 </longitude>
      <altitude> 0 </altitude>
    </gps>
    <url_media>pic.com/TfXd2D3</url_media>
  </tweet>

  <tweet id="t5" author="u3">
    <timestamp GMT="2"> 34</timestamp>
    <message length="140" type="text" color="black" lang="en">
      <ref_user> example </ref_user>
      <text> absolutely smashed it at </text>
      <text> ! Catch him at the official </text>
      <ref_user> clubmtvuk </ref_user>
      <text> after party tonight @10 pm </text>
    </message>
    <gps>
      <latitude> 0 </latitude>
      <longitude> 0 </longitude>
      <altitude> 0 </altitude>
    </gps>
    <url_media>pic.com/TfXd2D3</url_media>
  </tweet>

  <user id="u1" OS="Windows">
    <name> MTV Music </name>
    <city> Montpellier </city>
    <country> France </country>
    <url_avatar> null </url_avatar>
    <follow_count> 24 </follow_count>
    <sub_count> 94728472 </sub_count>
    <description> MTV Music UK </description>
  </user>

  <user id="u2" OS="Windows">
    <name> NRJ Music </name>
    <city> Paris </city>
    <country> France </country>
    <url_avatar> null </url_avatar>
    <follow_count> 29 </follow_count>
    <sub_count> 1597383 </sub_count>
    <description> MTV Music UK </description>
  </user>

  <user id="u3" OS="Windows">
    <name> W9 Hits </name>
    <city> Berlin </city>
    <country> Germany </country>
    <url_avatar> null </url_avatar>
    <follow_count> 41 </follow_count>
    <sub_count> 12042 </sub_count>
    <description> MTV Music UK </description>
  </user>

</tweets>
  ') );

SELECT EXTRACT(tweetXml, '//tweet/@id') FROM tweet_binaryxml;
SELECT EXTRACT(tweetXml, '//tweet/message') FROM tweet_binaryxml;
SELECT EXTRACT(tweetXml, '//user/@id') FROM tweet_binaryxml;

tweet_binaryxml
SELECT XMLQUERY('for $x in //user
  for $y in //tweet
    where $y/@author=$x/@id
      return
      <result>
        <pair
          tweetID ="{$y/@id}"
          author="{$x/name/text()}"
          />
      </result>'
    PASSING tweetXml RETURNING CONTENT)
FROM tweet_binaryxml;

SELECT XMLQUERY('
  for $x in //tweet
     return
    if  ($x/responses) then
    <retweet idtweet="{$x/@id}"><content>{$x/message}</content>
      {$x/responses/response/date}
      </retweet>
      else
      <nonRetwitted idtweet="{$x/@id}"/> '
    PASSING tweetXml RETURNING CONTENT)
FROM tweet_binaryxml;

SELECT XMLQUERY('
for $x in //user
for $y in //tweet
    where $y/@author=$x/@id
      return
      <result>
      <name>{$x/name/text()}</name>
      <dates>{$y/timestamp/text()}</dates>
      </result>'
    PASSING tweetXml RETURNING CONTENT)
FROM tweet_binaryxml;
```

WarmUp.sql

2 Stockage Monet

```
DROP TABLE presse;
DROP TABLE presse_journal;
DROP TABLE presse_journal_nom;
DROP TABLE presse_journal_directeur;
DROP TABLE presse_journal_directeur_nom;
DROP TABLE presse_journal_directeur_prenom;
DROP TABLE presse_journal_article;
DROP TABLE presse_journal_article_corps;
DROP TABLE presse_journalistes;
DROP TABLE presse_journalistes_journaliste;
DROP TABLE presse_journalistes_journaliste_anonymisation;


CREATE TABLE presse (
  node    VARCHAR(5),
  txtval  VARCHAR(5),
  numval  NUMERIC(5),
  CONSTRAINT PK_PRESSE PRIMARY KEY(node)
```

```sql
);

CREATE TABLE presse_journal(
  node  VARCHAR(5),
  txtval  VARCHAR(5),
  numval  NUMERIC(5)
  CONSTRAINT PK_PRESSE_JOURNAL PRIMARY KEY(node)
);

CREATE TABLE presse_journal_nom(
  node  VARCHAR(5),
  txtval  VARCHAR(5),
  numval  NUMERIC(5)
  CONSTRAINT PK_PRESSE_JOURNAL_NOM PRIMARY KEY(node)
);

CREATE TABLE presse_journal_directeur(
  node  VARCHAR(5),
  txtval  VARCHAR(5),
  numval  NUMERIC(5)
  CONSTRAINT PK_PRESSE_JOURNAL_DIRECTEUR PRIMARY KEY(node)
);


CREATE TABLE presse_journal_directeur_nom(
  node  VARCHAR(5),
  txtval  VARCHAR(5),
  numval  NUMERIC(5)
  CONSTRAINT PK_PRESSE_JOURNAL_DIRECTEUR_NOM PRIMARY KEY(node)
);

CREATE TABLE presse_journal_directeur_prenom(
  node  VARCHAR(5),
  txtval  VARCHAR(5),
  numval  NUMERIC(5)
  CONSTRAINT PK_PRESSE_JOURNAL_DIRECTEUR_PRENOM PRIMARY KEY(node)
);

CREATE TABLE presse_journal_article(
  node  VARCHAR(5),
  txtval  VARCHAR(5),
  numval  NUMERIC(5),
  CONSTRAINT PK_PRESSE_JOURNAL_ARTICLE PRIMARY KEY(node)
);

CREATE TABLE presse_journal_article_auteur(
  node  VARCHAR(5),
  txtval  VARCHAR(5),
  numval  NUMERIC(5),
  CONSTRAINT PK_PRESSE_JOURNAL_ARTICLE_AUTEUR PRIMARY KEY(node)
);

CREATE TABLE presse_journal_article_titre(
  node  VARCHAR(5),
  txtval  VARCHAR(5),
  numval  NUMERIC(5),
  CONSTRAINT PK_PRESSE_JOURNAL_ARTICLE_TITRE PRIMARY KEY(node)
```

```
);

CREATE TABLE presse_journal_article_corps(
   node  VARCHAR(5),
   txtval  VARCHAR(5),
   numval  NUMERIC(5)
   CONSTRAINT PK_PRESSE_JOURNAL_ARTICLE_CORPS PRIMARY KEY(node)
);

CREATE TABLE presse_journalistes(
   node  VARCHAR(5),
   txtval  VARCHAR(5),
   numval  NUMERIC(5)
   CONSTRAINT PK_PRESSE_JOURNALISTES PRIMARY KEY(node)
);

CREATE TABLE presse_journalistes_journaliste(
   node  VARCHAR(5),
   txtval  VARCHAR(5),
   numval  NUMERIC(5),
   CONSTRAINT PK_PRESSE_JOURNALISTES_JOURNALISTE PRIMARY KEY(NODE)
);

CREATE TABLE presse_journalistes_journaliste_idj(
   node  VARCHAR(5),
   txtval  VARCHAR(5),
   numval  NUMERIC(5),
   CONSTRAINT PK_PRESSE_JOURNALISTES_JOURNALISTE_IDJ PRIMARY KEY(node)
);

CREATE TABLE presse_journalistes_journaliste_pseudonyme(
   node  VARCHAR(5),
   txtval  VARCHAR(5),
   numval  NUMERIC(5),
   CONSTRAINT PK_PRESSE_JOURNALISTES_JOURNALISTE_PSEUDONYME PRIMARY KEY(node)
);

CREATE TABLE presse_journalistes_journaliste_anonymisation(
   node  VARCHAR(5),
   txtval  VARCHAR(5),
   numval  NUMERIC(5)
   CONSTRAINT PK_PRESSE_JOURNALISTES_JOURNALISTE_ANONYMISATION PRIMARY KEY(node)
);

INSERT INTO presse  (node,txtval,numval)
VALUES (n1,NULL,NULL);

INSERT INTO presse_journal  (node,txtval,numval)
VALUES (n2,NULL,NULL);

INSERT INTO presse_journal_nom  (node,txtval,numval)
VALUES (n3,NULL,NULL);

INSERT INTO presse_journal_directeur (node,txtval,numval)
VALUES (n4,NULL,NULL);
```

```sql

INSERT INTO presse_journal_directeur_nom (node,txtval,numval)
VALUES (n5,"Ulach",NULL);

INSERT INTO presse_journal_directeur_prenom (node,txtval,numval)
VALUES (n6,"Fred",NULL);

INSERT INTO presse_journal_article  (node,txtval,numval)
VALUES (n8,NULL,NULL);


INSERT INTO presse_journal_article_auteur (node,txtval,numval)
VALUES (n9,"Emile",NULL);

INSERT INTO presse_journal_article_corps (node,txtval,numval)
VALUES (n10,"faitDivers",NULL);

INSERT INTO presse_journal_article_titre (node,txtval,numval)
VALUES (n11,"viol",NULL);

INSERT INTO presse_journalistes (node,txtval,numval)
VALUES (n12,NULL,NULL);

INSERT INTO presse_journalistes_journaliste (node,txtval,numval)
VALUES (n13,NULL,NULL);

INSERT INTO presse_journalistes_journaliste_idj (node,txtval,numval)
VALUES (n14,NULL,164);

INSERT INTO presse_journalistes_journaliste_pseudonyme  (node,txtval,numval) VALUES (n15,"MCFred",NULL);

INSERT INTO presse_journalistes_journaliste_anonymisation (node,txtval,numval) VALUES (n16,"non",NULL);

--Xpath to SQL queries--

--/presse/journal/directeur/nom/text()
SELECT txtval FROM presse_journal_directeur_nom;

--/presse/journal/directeur/prenom/text()
SELECT txtval FROM presse_journal_directeur_prenom;

--/presse/journalistes/journaliste/@id
SELECT numval FROM presse_journalistes_journaliste_idj;
```

Monet.sql

3 Stockage schema-aware

3.1 Derivation du schéma relationel du DTD

3.1.1 Simplification des expressions régulières

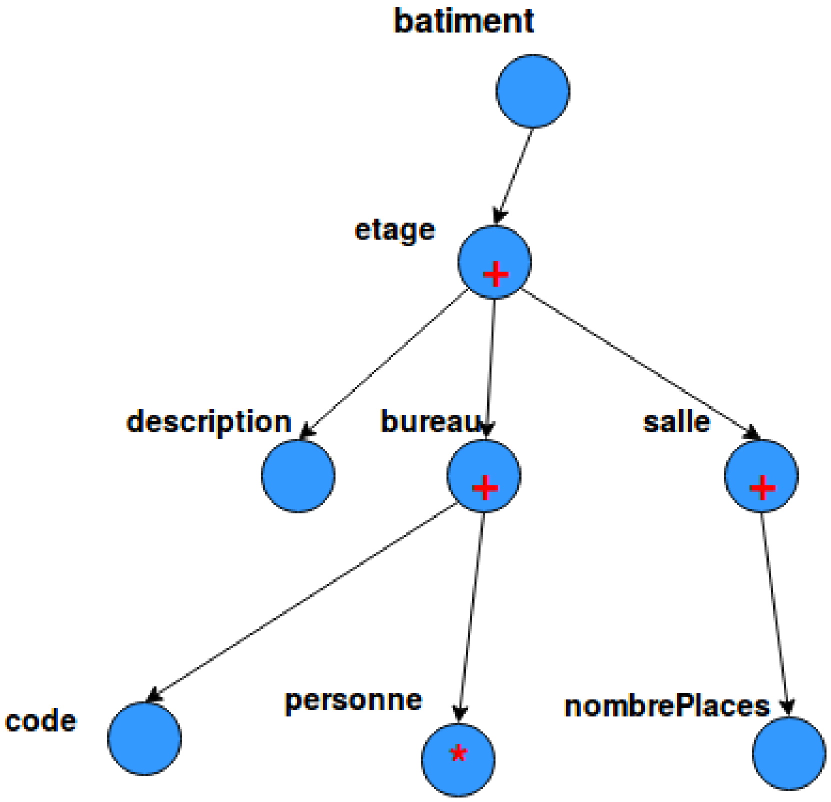
DTD initial :

```
<!DOCTYPE batiment [
 <!ELEMENT batiment (etage)+>
<!ELEMENT etage (description ,(bureau+|salle+))>
<!ELEMENT description (#PCDATA)>
<!ELEMENT bureau (code, personne*)>
<!ELEMENT code (#PCDATA)>
<!ELEMENT personne (#PCDATA)>
<!ELEMENT salle (nombrePlaces)>
<!ELEMENT nombrePlaces (#PCDATA)>]>
```

DTD transformé :

```
<!DOCTYPE batiment [
 <!ELEMENT batiment etage+ >
<!ELEMENT etage (description|(bureau+|salle+))>
<!ELEMENT description (#PCDATA)>
<!ELEMENT bureau (code|(personne*))>
<!ELEMENT salle (nombrePlaces)>]>
```

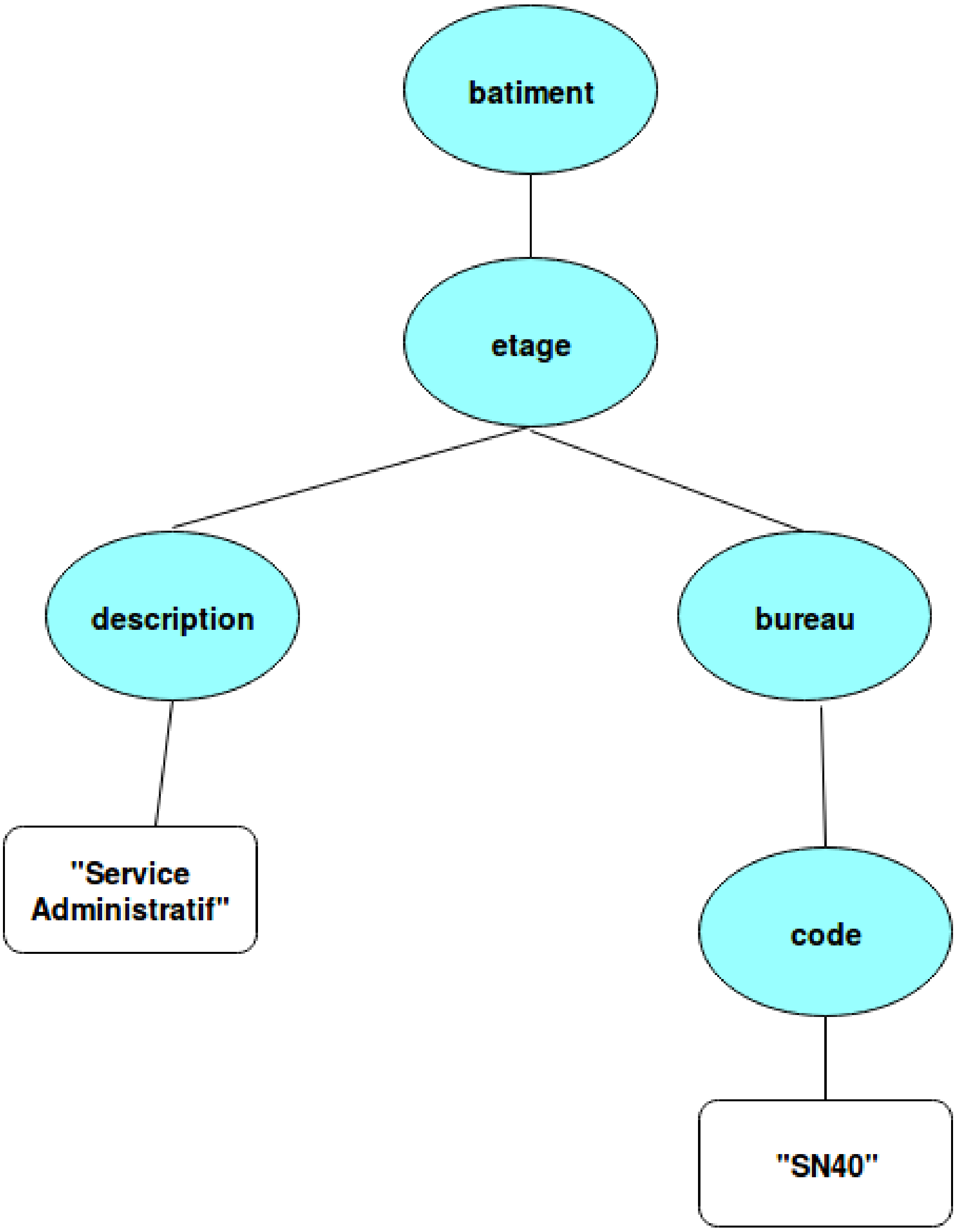
3.1.2 Graphe représantant le DTD obtenu



3.1.3 Creation des relations

batiment(batimentID)
etage(etageID,batimentID,description : string)
bureau(bureauID,etageID,code : string)
salle(salleID,etageID,nombrePlaces : string)
personne(personneID,bureauID,prs : string)

3.2 Remplissage des tables



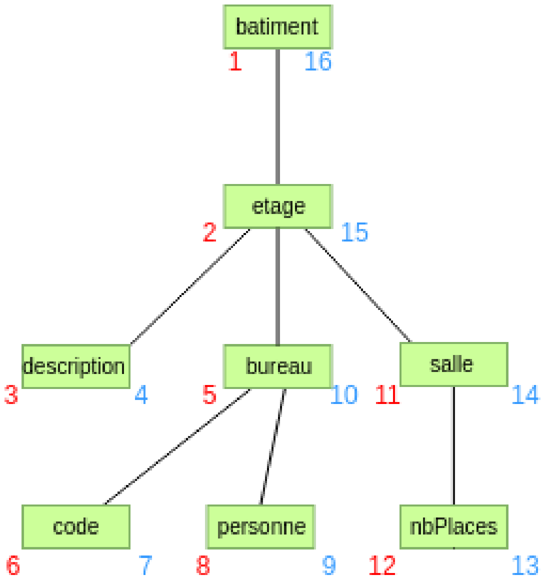
```
INSERT INTO batiment (batimentID) VALUES (1);
INSERT INTO etage (etageID,batimentID,description) VALUES (2,1,"Service
    Administratif");
INSERT INTO bureau(bureauID,etageID,code) VALUES (5,2,"SN40");
```

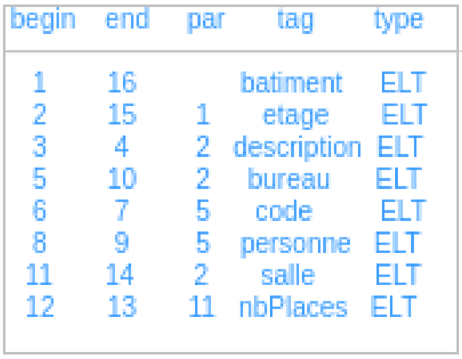
3.3 Traduction de requêtes en SQL

```
--/batiment/etage/description/text()
SELECT descrition FROM etage;
--/batiment/etage*
SELECT * FROM batiment,etage WHERE batiment.batimentID=etage.batimentID;
--/batiment/etage/bureau/code/text()
SELECT code FROM bureau ;
```

4 Interval-encoding avec SAX

4.1 Encodage begin/end





begin	end	par	tag	type
1	16		batiment	ELT
2	15	1	etage	ELT
3	4	2	description	ELT
5	10	2	bureau	ELT
6	7	5	code	ELT
8	9	5	personne	ELT
11	14	2	salle	ELT
12	13	11	nbPlaces	ELT