

Fiche XML n°2

Description d'un document XML

Éléments de correction pour les exercices

Exercice 1. Répertoire

Soit la DTD "repertoire.dtd" ci-contre qui décrit des documents permettant de mémoriser des répertoires d'adresses.

Question. Le document ci-dessous n'est pas valide. Corriger les erreurs du document XML ci-dessous. Vous indiquerez la localisation de l'erreur, la raison de l'erreur et ce qu'il faut modifier pour la corriger.

```

01 : <?xml version="1.0" encoding="ISO-8859-1"?>
02 : <!DOCTYPE repertoire SYSTEM "repertoire.dtd">
03 : <repertoire>
04 :   <catégorie nom="1-perso" >
05 :     <contact nom="desvignes" prénom="nicole">
06 :       <bureau téléphone="0234542312"
07 :         adresse="2 rue de l'espoir 44566 Chignole"/>
08 :       <maison téléphone="0870754566"
09 :         adresse="2 rue de l'espoir 44566 Chignole"/>
10 :     </contact>
11 :     <contact nom="fonteau" prénom="jérémie">
12 :       <maison téléphone="0445879044"
13 :         adresse="6bis avenue Jean Charcot 65322 St Mazan"/>
14 :       <bureau téléphone="0465345622"
15 :         adresse="54 bd de la liberté 65444 jalibert">
16 :         Batiment 6, Porte 5, Bureau 304
17 :       </bureau></contact>
18 :   </catégorie>
19 :   <catégorie nom="2-boulot">
20 :     <contact nom="martin" prénom="paul">
21 :       <bureau téléphone=0245664442 adresse="5 rue de l'épine 45665 Astorie"
22 :         courriel="paul.martin@gmail.com" fax="0245664412"/>
23 :     </contact>
24 :     <contact nom="batiset" prénom="raymond">
25 :       <maison téléphone="0248666445" adresse="5 rue de l'épine 45665 Astorie"
26 :         courriel="rbatisset@free.fr"/>
27 :       <contact nom="fonteau" prénom="paul">
28 :         <bureau téléphone="0245678966" adresse="34 chemin de la fraise 56789 Ghéno"
29 :           email="paul-fonteau@yahoo.fr"/>
30 :         <maison téléphone="0256780032" adresse="11 rue du petit plaisir 56888 Trigonet"/>
31 :       </contact>
32 :     </contact>
33 :   </catégorie>
34 : </repertoire>

```

```

<?xml version="1.0" encoding="ISO-8859-1"?>
<!ELEMENT repertoire (catégorie)*>
<!ELEMENT catégorie (contact)*>
<!ATTLIST catégorie
  nom ID #REQUIRED>
<!ELEMENT contact (bureau, maison?)>
<!ATTLIST contact
  nom ID #REQUIRED
  prénom CDATA #REQUIRED>
<!ELEMENT bureau EMPTY>
<!ATTLIST bureau
  téléphone CDATA #REQUIRED
  adresse CDATA #REQUIRED
  courriel CDATA #IMPLIED>
<!ELEMENT maison EMPTY>
<!ATTLIST maison
  téléphone CDATA #REQUIRED
  adresse CDATA #REQUIRED
  courriel CDATA #IMPLIED>

```

Solution.

- en entête, il manque «standalone='no'»
- lignes 3 et 33, manque un accent sur le premier «e»
- ligne 24-25 : manque l'élément «bureau»
- inverser les lignes 12/13 et 14/15
- ligne 21, manque les guillemets autour du numéro de téléphone
- lignes 4 et 19, ce ne sont pas des noms XML
- lignes 26-27 manque la balise fermante «</contact>»
- ligne 16, la balise «bureau» doit être vide
- ligne 22, l'attribut «fax» n'existe pas
- ligne 27, le nom ne peut pas être réutilisé (ID)
- ligne 29, remplacer «email» par «courriel»

Exercice 2. Formules Mathématiques

Soit la DTD ci-contre.

Question 1. Une erreur s'est glissée dans cette DTD : la corriger.

Solution 1. Évidemment, «formule» utilise un élément «différence» (absent de la DTD) alors que l'élément «diff» n'est pas utilisé. Il faut donc renommer ce dernier.

Question 2. Écrire un document XML valide (conforme à la DTD ci-contre) destiné à transmettre l'énoncé suivant :

```
<!-- DTD pour décrire un texte contenant des formules mathématiques -->
<ELEMENT textemath (texte | formule)+ >
<ELEMENT texte (#PCDATA) >
<ELEMENT formule (valeur | somme | différence
    | produit | fraction | racine | puissance) >
<ELEMENT fraction (op1, op2)>
<ELEMENT valeur (#PCDATA)>
<ELEMENT somme (op1, op2)>
<ELEMENT diff (op1, op2)>
<ELEMENT produit (op1, op2)>
<ELEMENT racine (op1)>
<ATTLIST racine ordre CDATA #IMPLIED>
<ELEMENT puissance (op1)>
<ATTLIST puissance exposant CDATA #REQUIRED>
<ELEMENT op1 (valeur|formule)>
<ELEMENT op2 (valeur|formule)>
```

Calculer la valeur de l'expression $\frac{x^3 + \sqrt[3]{5}}{2\sqrt{3}}$ pour $x = 4$

Solution 2. Le document selon la DTD présenté est le suivant :

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<!DOCTYPE textemath SYSTEM "math.dtd">
<textemath> <texte>Calculer la valeur de l'expression</texte>
<formule>
  <fraction>
    <op1>
      <formule>
        <somme>
          <op1>
            <formule> <puissance exposant = "3"><op1><valeur>x</valeur></op1></puissance> </formule>
          </op1>
          <op2>
            <formule>
              <racine ordre = "3"><op1><valeur>5</valeur></op1></racine>
            </formule>
          </op2>
        </somme>
      </formule>
    </op1>
    <op2>
      <formule>
        <produit>
          <op1><valeur>2</valeur></op1>
          <op2>
            <formule> <racine><op1><valeur>3</valeur></op1></racine> </formule>
          </op2>
        </produit>
      </formule>
    </op2>
  </fraction>
</formule>
<texte>pour x = 4</texte>
</textemath>
```

Cette solution n'est pas satisfaisante, car elle utilise beaucoup de balises inutiles («op1», «op2», «formule»...). D'autres solutions équivalentes, mais moins volumineuses (en nombre de balises), peuvent être proposées. Nous en verrons dans d'autres fiches.

Question 3. Faire en sorte que la partie «x = 4» soit aussi écrite par «formule», c'est-à-dire que l'opérateur «=» soit aussi pris en compte.

Solution 3. Il suffit d'introduire le fait que «textemath» peut aussi être une égalité ou que «formule» propose aussi «egal» qui est défini par «(op1, op2)»...

Question 4. Modifier la DTD pour introduire la notion de variable (comme «x») et de constante (comme «2», «3»...).

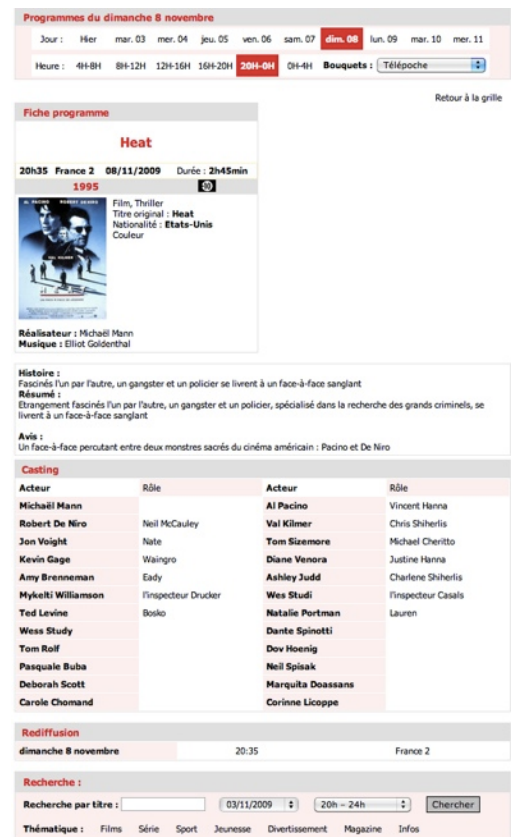
Solution. Il faut remplacer «valeur» par deux éléments : «constante» et «variable».

Question 5. Que se passe-t-il lorsque l'on veut mettre une variable en exposant ou dans l'ordre de la racine ? Modifier la DTD en conséquence.

Solution. Les attributs «ordre» et «exposant» deviennent alors des éléments.

Exercice 3. Programme télé

Le responsable d'un site Web spécialisé se propose de présenter les programmes de télévision par semaines. Pour cela, il envisage de mémoriser les informations dans un document XML. La figure suivante présente le programme télé de d'une semaine¹ :



Ci-contre, une solution de DTD est proposée.

Question 1. Proposer un document XML valide pour cette semaine en utilisant cette DTD. Afin de ne pas faire trop long, vous ne ferez que le programme de ce jour sur France 2. Toutes les parties omises ou inconnues seront représentées par des points de suspension en commentaire.

Question 2. A votre avis, quels sont les inconvénients de cette DTD ? Permet-elle de représenter tout ce qui apparaît dans l'exemple précédent ?

Solution 2. Les principaux inconvénients :

- impossibilité de fixer le nombre de semaine ou le nombre de jours dans la semaine
- il peut ne pas y avoir de programme dans une journée,

```
<?xml version="1.0" encoding="UTF-8"?>
<ELEMENT programmation (semaine)* >
<ATTLIST programmation annee CDATA #REQUIRED>

<ELEMENT semaine (journee)* >
<ATTLIST semaine no_semaine CDATA #REQUIRED>

<ELEMENT journee (programme)* >
<ATTLIST journee
    nom (lundi | mardi | mercredi | jeudi | vendredi | samedi | dimanche) #REQUIRED>

<ELEMENT programme (creneau-horraire, description, commentaire?) >
<ATTLIST programme
    titre CDATA #REQUIRED
    chaine CDATA #REQUIRED
    type CDATA "fiction"
    avis-redaction (deconseille | sans-avis | correct | bon | excellent) 'sans-avis'
    accord-parental (tout-public | moins10 | moins12 | moins16 | moins18) 'tout-public'>

<ELEMENT description (#PCDATA) >
<ELEMENT commentaire (#PCDATA) >

<ELEMENT creneau-horraire EMPTY>
<ATTLIST creneau-horraire
    debut CDATA #REQUIRED
    fin CDATA #IMPLIED
    duree CDATA #IMPLIED>
```

¹ <http://www.telepoche.fr/listing.html>

- dans une semaine, dans une année
- Comme toujours avec les DTD, problèmes de contrôle de types (horaires, numéro de semaine...).
- impossible de contrôler la cohérence du créneau horaire ou ne s'obliger qu'à un seul choix entre «durée» et «fin».
- ordre contraint dans le programme
- «CDATA» pour le type peut être remplacé par une liste de possibilités comme pour «avis-redaction» ou «accord-parental».
- n'apparaît pas l'histoire et le résumé.
- les structures d'émissions sont limitées...

Exercice 4. Annotation d'entités nommées

En traitement automatique du langage naturel, la reconnaissance d'entités nommées est une sous-tâche de l'activité d'extraction d'information dans des corpus documentaires. Elle consiste à rechercher des objets textuels (c'est-à-dire un mot, ou un groupe de mot) catégorisables dans des classes telles que noms de personnes, noms d'organisations ou d'entreprises, noms de lieux, quantités, distances, valeurs, dates, etc.²

Émeutes de Birmingham

Les **émeutes de Birmingham**, aussi connues sous le nom d'**émeutes de Priestley**, se déroulent du 14 au 17 juillet 1791 à Birmingham, en Angleterre. Les Dissidents de l'Église d'Angleterre et, en particulier, le théologien et philosophe politique Joseph Priestley, sont la cible principale des émeutiers. Les événements locaux et nationaux, qui suscitent la passion des foules, vont d'un désaccord sur l'achat de livres par la bibliothèque publique, jusqu'à la controverse à propos des tentatives des Dissidents pour l'obtention de droits identiques à ceux des autres citoyens du royaume, en passant par leur soutien à la Révolution française.

Les émeutes commencent par l'attaque d'un hôtel où se tient un banquet organisé pour célébrer le second anniversaire de la prise de la Bastille. Puis, commençant par l'église et la maison de Priestley, les émeutiers attaquent ou incendient quatre chapelles des Dissidents, vingt-sept maisons et plusieurs commerces. Nombre d'entre eux se saoulent avec l'alcool trouvé dans les lieux pillés ou celui qu'on leur offre pour qu'ils n'incendient pas tel ou tel bâtiment. Les émeutiers brûlent, non seulement les maisons et les chapelles des Dissidents, mais aussi les foyers des gens connus pour être favorables à leur cause, comme les membres de la communauté scientifique de la *Lunar Society*.

Même si les émeutes n'ont pas été organisées par le gouvernement du Premier ministre William Pitt, il est très lent à répondre à l'appel à l'aide des Dissidents. Des officiels locaux de Birmingham ont sans doute été impliqués dans la préparation des émeutes et vont plus tard se montrer réticents à en poursuivre les meneurs. L'industriel James Watt écrit que les émeutes « divisèrent [Birmingham] en deux camps qui se haïssaient mortellement ». Ceux qui ont été victimes des émeutiers quittent progressivement Birmingham, abandonnant une ville devenue plus conservatrice qu'elle ne l'a été tout au long du XVIII^e siècle.



Soit le texte ci-dessous, début de la page http://fr.wikipedia.org/wiki/Émeutes_de_Birmingham sur Wikipédia.

Peinture représentant les émeutes de Birmingham
(Auteur inconnu, fin du XVIII^e siècle)

http://fr.wikipedia.org/wiki/Fichier: Priestley_Riots_painting.jpg

Question. Proposer une DTD afin de repérer les entités nommées présentes dans un texte et l'appliquer sur le document ci-contre.



Cette création est mise à disposition selon le Contrat Paternité-Pas d'Utilisation Commerciale-Pas de Modification 2.0 France disponible en ligne <http://creativecommons.org/licenses/by-nc-nd/2.0/fr/> ou par courrier postal à Creative Commons, 171 Second Street, Suite 300, San Francisco, California 94105, USA.

² http://fr.wikipedia.org/wiki/Entités_nommées