Présentation des données du Web - TD 1 : XML et DTDs

TD en binômes. Rendu facultatif possible sur Moodle, le 17/09.

1) Dès schémas aux données

Pour chaque DTD, donner 2 (deux) arbres XML valides.

DTD 1

```
<!DOCTYPE presse [
 <!ELEMENT presse (journal,journalistes) >
 <!ELEMENT journal (nom, directeur, article*) >
 <!ELEMENT article (corps) >
 <!ATTLIST article titre CDATA #IMPLIED >
 <!ATTLIST article auteur IDREF #REQUIRED >
 <!ELEMENT corps (#PCDATA) >
 <!ELEMENT journalistes (journaliste+) >
 <!ATTLIST journaliste idJ ID #REQUIRED >
 <!ELEMENT journaliste ((nom,prenom)|pseudonyme) >
 <!ATTLIST journaliste anonymisation (oui|non) ''non'' >
 <!ELEMENT pseudonyme (#PCDATA) >
 <!ELEMENT nom (#PCDATA) >  <!ELEMENT prenom (#PCDATA) >
 <!ELEMENT directeur (nom,prenom) > ]>
```

DTD 2

```
<!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) >]>
```

2) Dès données aux schémas

Pour chaque document XML, donner 2 (deux) DTDs permettant de valider le document.

XML 1

```
<F>
 <C><A/></C>
 <C><D/></C>
 <B><A/></B>
 <B><E/></B>
</F>
```

XML 2

```
<D>
 <C/><C/><E/>
 <B/><C/>
 <C/><C/><E/>
</D>
```

XML 3

```
<A>
 <C/><C/>
 <E/><E/>
 <E/><C/>
</A>
```

XML 4

```
<B>
 <C/><B/>
 <C/><B/><E/>
 <D/><D/><E/>
</B>
```

XML 5

```
<A/>
```

XML 6

```
<A>
 <C id = "p4" friend = "p4">Charles </C>
 <C id = "p5" friend = "p5">Bob</C>
 <C id = "p5" friend = "p1">Alice</C>
</A>
```

3) Modélisation : à quoi ressemble un Tweet ?

On veut définir une DTD pour stocker des tweets collectés du Web, tout en montrant l'intérêt ainsi que les possibles limitations des DTDs.

- Un tweet n'est pas juste un message de 140 caractères. C'est un objet complexe émis à une date precise (exprimée en secondes, par rapport à un fuseau horaire), qui a un identifiant unique. Si possible, les coordonnées géographiques, la ville, et le pays de l'émetteur du tweet, ainsi que une description de son système d'exploitation sont présentes. Images et videos intégrés dans un post sont référencés par un url.

- Le corps d'un message est composé par du texte libre mélangé avec des hashtags (e.g., `#I<3XML`) et des références d'utilisateur (e.g, `@timberners_lee`). On enregistre la taille, le type et le couleur du font du texte. Il est également important d'enregistrer la langue du message. Un post peut être retweeté (on enregistre le nombre de fois), ansi que apparaitre en réponse à un tweet précédent (dans ce cas, les réponses doivent être autorisées pour le tweet en question).

- On enregistre l'identifiant, le nom, ainsi que le lien vers le profile de l'auteur d'un message. Pour chaque utilisateur de la plateforme on a une description détaillée, une photo, le nombre d'utilisateurs qui le suivent et ceux dont il est abonné.