XML et Données Semi Structurées

Structure d’un document XML

Structure d'un document XML

Un document **XML** est composé:

1) d'un prologue
2) d'un élément racine du document qui est lui-même composé d'éléments et des données textuelles
3) des commentaires
4) d'instructions de traitements destinées aux applications traitant le document

Structure d'un document XML

Exemple:

```
<?xml version="1.0" encoding="UTF-8"?>                    (1)
<!-- ceci est un commentaire -->                          (3)
<employees>                                               (2)
   <employee>                                             (2 )'
      <employee_id>120</employee_id>
      <last_name>Weiss</last_name>
      <salary>8000</salary>
   </employee>
</employees>
<?gifPlayer size="300,100" ?>                             (4)
```

Structure d'un document XML

Il peut être découpé en entités dans un ou plusieurs fichiers.

- **Le prologue**: Il s'agit de la première ligne d'un document XML. Il sert à donner les caractéristiques globales du document c'est-à-dire:
  - la version XML (soit 1.0 ou 1.1),
  - le jeu de caractère employé (encoding),
  - l'indépendance du document (standalone)

Structure d'un document XML

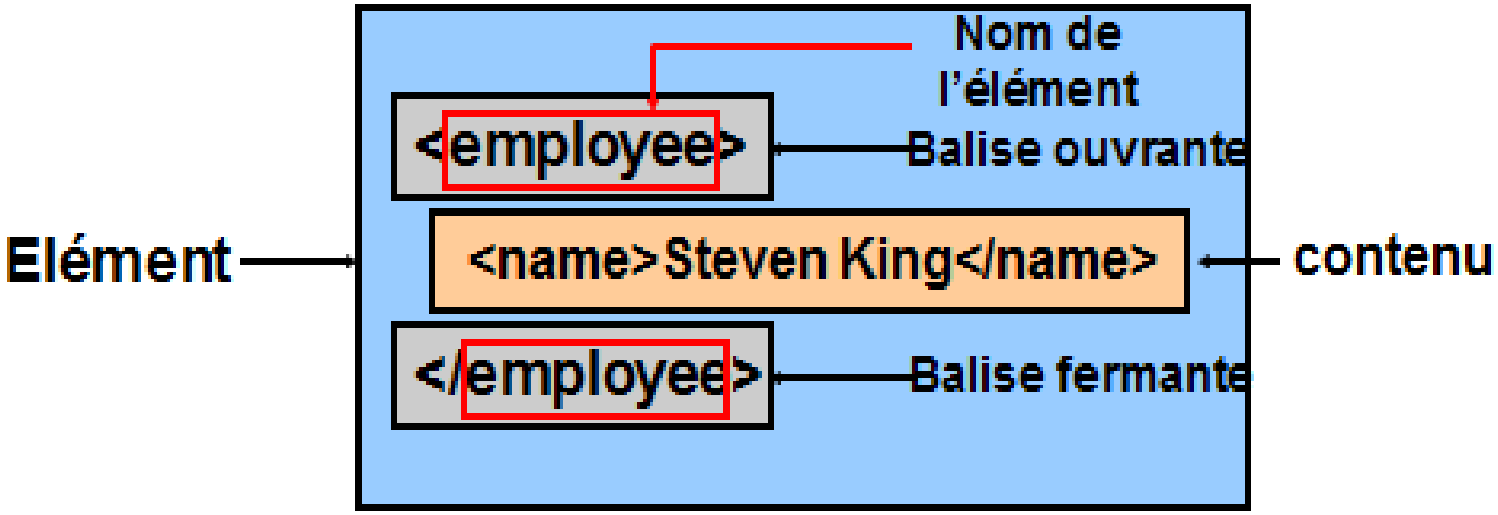
Exemple:

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="no"?>
```

NB: dans un document XML, le prologue n'est pas obligatoire. Par défaut, tout document XML dispose les caractéristiques suivantes: (version 1.0 et encoding utf-8)

Structure d'un document XML

- **Un élément XML**: il est composé d'une balise ouvrante (qui contient le nom de l'élément et éventuellement ses attributs), d'un contenu et d'une balise fermante



Structure d’un document XML

- **Une balise** doit :
  - ✓toujours aller de pairs (utiliser la syntaxe abrogée pour les éléments vide),
  - ✓respecter les minuscules et les majuscules
  - ✓commencer par une lettre, un '_' ou un ':' suivi d’une combinaison de lettres, de chiffres, '_' , '-' , ':' ou '.'

Structure d’un document XML

- **Un attribut** doit :
  - ✓ toujours posséder une valeur (cette valeur est toujours définie entre guillemets ou apostrophes. Elle peut être vide ou non vide),
  - ✓ respecter les minuscules et les majuscules
  - ✓ commencer par une lettre, un '_' ou un ':' suivi d’une combinaison de lettres, de chiffres, '_' , '-' , ':' ou '.'

Structure d'un document XML

➢ **Exemple**:

```
<?xml version="1.0" encoding="utf-8"?>
<employees>
  <employee id="100"    name='Rachael O&apos;Leary'>
    <salary>1000</salary>
  </employee>
</employees>
```

Structure d'un document XML

- **Le contenu** peut être soit:
  - ✓ vide: **<name></name>** ou **<name/>**
  - ✓ composé d'éléments:

```
<employee>
      <id>100</id>
      <last_name>King</last_name>
      <salary>24000</salary>
</employee>
```

Structure d’un document XML

- ✓ mixte: mélange de texte et d’éléments

```
<employee>
   <id/>, identifiant vide.
   <last_name>King</last_name>
   <salary>24000</salary>
</employee>
```

Structure d'un document XML

- **<u>Les données textuelles</u>:** elles sont contenues dans certains éléments et dans la valeur des attributs. **NB**: Les caractères < et & doivent respectivement être codés comme suit < et &. Lorsqu'un texte contient des caractères qui jouent un rôle de délimiteur dans la syntaxe XML, il est parfois nécessaire de pouvoir inhiber ce rôle. Ceci peut être fait en insérant le texte contenant les délimiteurs dans une **section CDATA** sous la forme suivante.

Structure d'un document XML

- **Section CDATA:** <![CDATA[*texte contenant des délimiteurs*]]>

Le texte inséré peut contenir n'importe quel caractère excepté la chaine **]]**. Une section CDATA ne peut donc pas contenir une autre. Par exemple la phrase « L'expression <name> Dior </name> est un élément XML » peut être représenté par l'élément suivant:

<phrase>

L'expression <![CDATA[<name> Dior </name>]]> est un élément XML

</phrase>

Structure d'un document XML

- **<u>Les références de caractères</u>**: permettent de de coder les caractères non directement supportés par le type d'encodage. Elles utilises le code unicode du caractère.

  - **<u>La syntaxe</u>:**
    - Code décimal du caractère précédé de **&#** et suivi de **;**
    - Code hexadécimal du caractère précédé de **&#x** et suivi de **;**

  Par exemple: `<first_name>Am&#233;lie</first_name>`

Structure d’un document XML

- **Les références d’entités**: font référence à une entité (à titre de rappel, une entité est une unité de stockage qui contient une partie du document). Le contenu des entités est appelé texte de remplacement. Elles (entités) peuvent être internes ou externes.
  - **La syntaxe: $nom_entité;**

Structure d’un document XML

- **Les entités prédefinies en XML**:

&amp ;	&
&lt ;	<
&gt ;	>
&quot ;	" (guillemet)
&apos ;	' (apostrophe)

Structure d'un document XML

- **Les commentaires**: ce sont des phrases qui ont la formes suivantes `<!-- texte du commentaire -->`
  - Ils peuvent contenir n'importe quel caractère excepté --
  - Ils ne peuvent donc pas inclure d'autres commentaires
  - Ils peuvent être inclus dans le contenu d'un élément mais pas à l'intérieur d'une balise.

Structure d’un document XML

- **Les instructions de traitements**: elles sont destinées aux parseurs qui vont exploiter le document. Elles commence par <?suivi d'un nom et se termine par ?>. Elles ne font pas partie du document. Elles peuvent varier d’un parseurs à un autre. Par exemple:

```
<?xml-stylesheet  type="text/xsl"  href="style.xsl"?>
<?xml-stylesheet  type="text/css"  href="style.css"?>
```

Structure d'un document XML

- **NB**: Respecter la syntaxe XML ne suffit pas pour qu'un document soit utilisable. Deux notions essentielles sont à distinguer avec les documents XML:
  - Les documents XML **bien-formés**: ils sont syntaxiquement correctes et respectent un certains nombres de règles précises.
  - Les documents XML **valide**: ils sont bien-formés et respectent une DTD (ou un schéma)

  Il est primordial qu'un document XML soit bien formé, il n'est pas nécessaire qu'il soit valide.

Structure d'un document XML

- **<u>Les document xml bien-formés</u>**: pour qu'un document XML soit bien-formé il faut que:
  - le document commence par un prologue correctement formé
  - le document contient un seul élément appelé **élément racine**
  - le contenu de cet élément respecte un certain nombre de règles de contraintes qui sont les suivantes:
    - tous les éléments doivent être correctement imbriqués;
    - un même attribut ne peut apparaitre qu'une seule fois dans une balise;

Structure d’un document XML

- pas de référence dans un attribut vers une entité externe;
- pas de référence dans un attribut vers une entité dont le texte contiendrait un < ;
- les références de caractères doivent représenter un caractère reconnu dans **unicode**;
- les références d’entités doivent faire référence à une entité existant qui doit être analysable;
- une entité ne peut pas se faire référence en elle-même,
- les entités doivent être déclarées dans la DTD avant leur utilisation.

Structure d’un document XML

- **<u>Les document xml valide</u>**: un document xml bien-formé ne garantit pas que son contenu est correct. Pour ce faire, on doit lui associer une DTD (**D**ocument **T**ype **D**efinition) qui définit une sorte de grammaire que devra respecter le document.