62-31.3 XML

[XML-01-P01-Introduction]

Cédric Benoit

Introduction (1)

XML: est l'abréviation de *eXtensible Markup Language*

Besoin : Commenter des documents texte à l'aide de balises

But : faciliter la recherche, l'archivage et la présentation en séparant l'information (contenu) et son contexte (métadonnées)

Principe : marquer des parties d'un texte en l'entourant de Balises. En XML, ces balises sont entre les symboles "<" et ">".

Introduction (2)

On peut traduire XML par *Langage à balises étendu*, ou *Langage à balises extensible*

C'est un langage qui permet de structurer des documents ou des données grâce à des balises qui se traduit en anglais par *markup*

La structure du document peut être ou non validée par un schéma

XML, un standard

XML est régi par le W3C (World Wide Web Consortium)

Lien: www.w3.org/XML/

"XML is the universal format for structured documents and data on the Web."

XML est un format de données universel, qui permet de gérer toutes les structures

- Fichiers plats
- Tables relationnelles
- Arbres
- Etc.

XML est un langage à balise

XML est un langage à balises, en imbriquant entre la balise d'ouverture <BALISE> et la balise de fermeture </BALISE>, d'autres balises et du texte encodant des données

Toute balise peut être porteuse d'un ou plusieurs attributs avec une valeur

Exemple d'un document XML

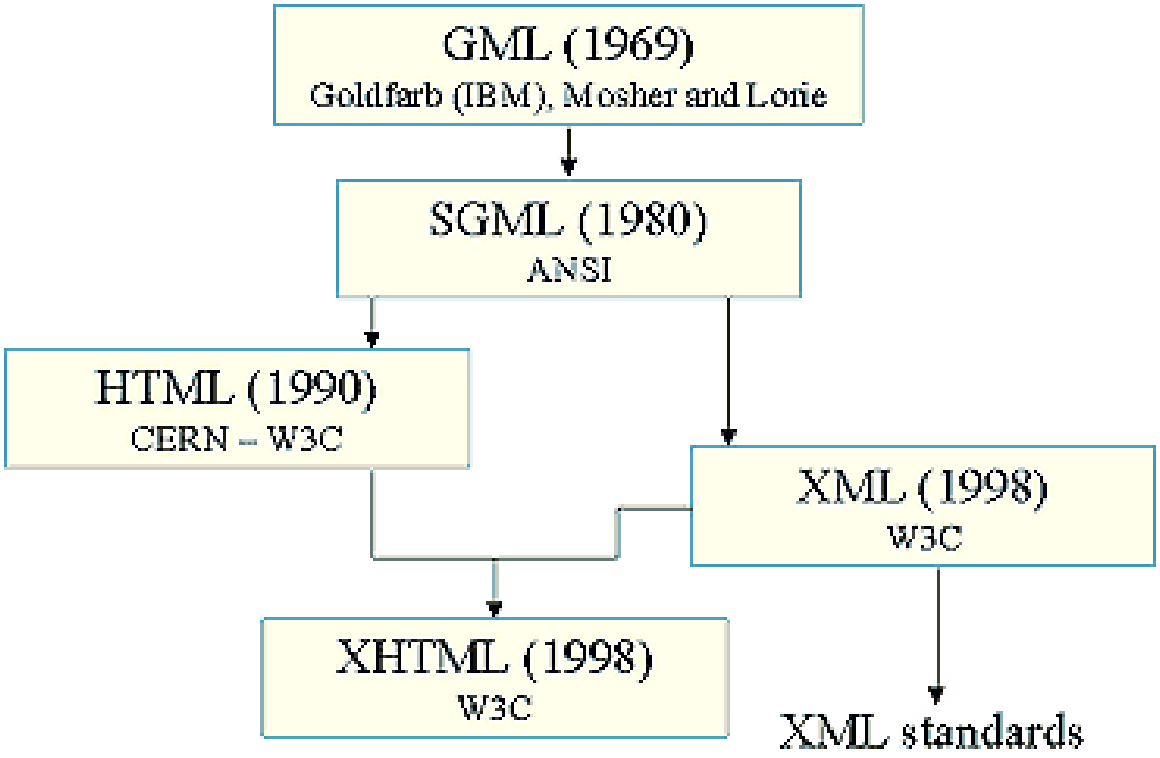
```
<? xml version="1.0" encoding="ISO-8859-1"?>
 <EMPLOYEE EMPNO="7566">
        <ENAME>JONES</ENAME>
        <JOB>MANAGER</JOB>
        <SAL>2975</SAL>
        <DEPARTEMENT>
                <DEPTNO>20</DEPTNO>
                <DESC>MARKETING</DESC>
                <CA>200000</CA>
        </DEPARTEMENT>
</EMPLOYEE>
```

Atouts de XML

- Lisible
  - Contenu sous forme texte particulièrement simple et lisible par l'humain
- Auto-descriptif
  - Permet de décrire sa propre structure
- Arborescent
  - C'est un arbre, et cela permet de modéliser la majorité des problèmes informatiques
- Universel et portable
  - les différents jeux de caractères sont pris en compte

Généalogie de XML

- Source: http://w4.uqo.ca/iglewski/ens/inf4533/src/xml/xml_intro.php



Structure d'un document XML (1)

```
<?xml version="1.0" encoding="UTF-8"?>


<!-- Ceci est un commentaire -->
<formulaire numForm="1">
    <titre>Voici du XML</titre>
</formulaire>
```

Structure d'un document XML (2)

Un document XML commence par un **prologue**

Suivi d'un arbre d'éléments

Tous les **éléments** sont à l'intérieur de l'**élément racine**

Une balise peut contenir des **attributs**

Entête du document

Le prologue est facultatif, mais il est fortement conseillé

- Ce n'est pas une balise XML `<?xml…`

```
<?xml version="1.0" ?>
```

De manière optionnelle on spécifie le jeu de caractères (*encoding*) utilisé dans le document

- Exemple : jeu ISO-8859-1 (jeu LATIN), jeu UTF
- Permet de prendre en compte les accents français avant l'UTF 8

```
<?xml version="1.0" encoding="ISO-8859-1"?>
```

Le prologue

Le prologue contient :

La déclaration XML indique :

- Version de XML utilisée
- Norme de codage des caractères
- Si le document est autonome

Les déclarations de type de document (facultatif) :

- Indique la grammaire à respecter (DTD, XSD, …)

Les déclarations commencent par : <!

Les instructions de traitement par : <?

L'arbre des éléments

- **Elément** : unité fondamentale d'un document XML, toutes les informations se trouvent dans des éléments.
- **Elément racine** : Premier élément du document. Il contient tous les autres, et forme la racine de l'arbre des éléments.
- **Balise** : marque le début et la fin d'un élément.
- **Attribut** : information additionnelle sur un élément. Présent dans sa balise de début, un élément peut avoir plusieurs attributs.
- **Entité** : Référence à un texte pouvant servir de raccourci. Pratique pour utiliser les caractères spéciaux (<, >, &, …)

Règles de syntaxe (1)

- Les balises sont délimités par des chevrons : **<** et **>**
- Toute balise doit être fermée : **<age>31</age>**
- Raccourci pour balise vide : **<vide/>**
- Balises bien imbriquées : **<out><in></in></out>**
- Respect de la casse : **<tag>** est différent de **<Tag>**
- Pas d'attribut sans valeur : **<tag num="1">**

Règles de syntaxe (2)

- Un élément est composé :
  - D'une balise d'ouverture
  - D'un contenu
  - D'une balise de fermeture

- Le nom de l'élément (balise) est composé :
  - Caractères alphanumériques, souligné, tiret et point
  - Pas de caractères d'espacement ou de fin de ligne
  - ":" est réservé pour les espaces de noms (namespace)
  - Premier alphanumérique ou "_"
  - Aucune balise ne peut commencer par "xml" (quelque soit la casse)

Validation d'un document XML

- Un document est **bien formé** s'il respecte la **syntaxe** XML
- Un document est **valide** s'il respecte les **règles grammaticales** de son modèle (schéma XML qui peut être type XSD, DTD, …)
- La syntaxe stricte de XML facilite le développement d'outils utilisables pour traiter tout langage issu de XML

Exemple d'un document bien formé

```
<?xml version="1.0" encoding="ISO-8859-1"?>

<!-- Document de description d'un film -->

<Film id="7566">
  <Titre>Alien</Titre>
  <Auteur>Ridley Scott</Auteur>
  <Acteurs>
     <Acteur>Tom Skeritt</Acteur>
     <Acteur>Sigourney Weaver</Acteur>
   </Acteurs>
  <Annee>1979</Annee>
  <Genre>Science-fiction</Genre>
  <Resume>Des aliens envahissent un vaisseau spatial et
 les cauchemars commencent pour  les personnes à bord</Resume>
</Film>
```

Quelques causes pour un document mal formé

```
<Film id="7566" id="7346" >
  <Titre>Alien</Titre>
  <Auteur>Ridley Scott</auteur>
  <Acteurs>
     <Acteur>Tom Skeritt</Acteur>
     <Acteur>Sigourney Weaver
   </Acteurs>
  <Annee>1979
  <Genre>Science-fiction </Annee> </Genre>
  <Resume du film>Des aliens envahissent un vaisseau spatial et
 les cauchemars commencent pour  les personnes à bord</Resume du
film>
</Film>
```

Domaines d'utilisation

En plus du balisage de documents :

- Web (xthml, AJAX, RSS, …)
- Services Web (SOAP, WSDL)
- Multimédia (SVG, SMIL)
- Interfaces graphiques (XUL, XAML, …)
- Interopérabilité (ODF, OpenXML, GPX, …)
- Domaines spécialisés (CML, MathML, …)
- Traitement de XML (XLink, XPointer, XPath, …)

Il est possible de combiner plusieurs dialectes de XML. Par exemple: **xhtml + MathML + SVG**

Simple, compréhensible, lisible,…

Une des grandes qualités de XML est d'être lisible et auto-documenté. La manière de concevoir un format XML doit chercher à tirer partie de cette qualité.

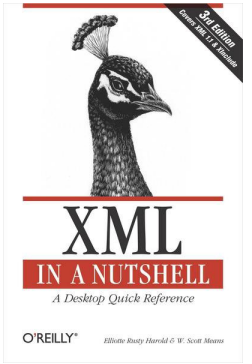
- Bon: <book><title/><author/></book>
- Moins bon: <f001><f002/><f003/></f001>

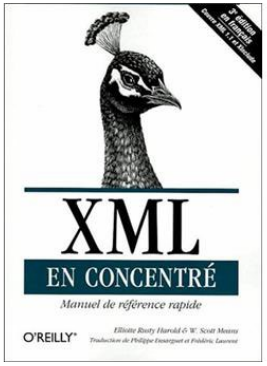
Isoler toutes les valeurs qui sont individuellement intéressantes dans des éléments ou attributs séparés.

- Bon: <pages><page>10</page><page>15</page></pages>
- Moins bon: <pages>10,15</pages>

Références

- La recommandation du W3C dont le lien: est: www.w3.org/XML/
- Le livre "XML en concentré" dont la version anglaise "XML in a Nutshell" peut être consultée en ligne sur O'RELLY (https://learning.oreilly.com/home/) avec la connexion SWITCHedu-ID
- Mon expérience dans le domaine,…





Merci pour votre attention !