Présentation de XML

Cours DAX pour Master-2 option TI

Présentation de XML

Définition, Historique, intérêts, syntaxe et structure….

Définition

XML (e**X**tended **M**arkup **L**anguage) un métalangage de balisage, conçu vers 1997 afin de faciliter l'échange de données via le Web.

- Langage orienté texte, formé de balises permettant d'organiser les données de manière structurée;
- Standard incontournable de l'informatique, utilisé aussi bien pour le stockage et le partage de documents, que pour la transmission de données entre applications;
- Simple, flexible et facile d'utilisation, format universel de texte brut, auto-descriptif, normalisé et ouvert;
- Adapté aux traitements automatiques, disponibilité d'outils génériques d'analyse et de transformation.

Historique

XML dérive du langage **SGML** (**S**tandard **G**eneralized **M**arkup **L**anguage), développé dans les années 80. Ce langage, servant à préciser la structure d'un document quelconque, a été conçu pour les documentations techniques de grande ampleur. Sa généralité la rendue difficile et complexe, ce qui a freiné son utilisation.

Le **HTML** (**H**yper**T**ext **M**arkup **L**anguage) une version allégée et adapté à l'écriture de documents pour Internet a été développée. Mais ce dernier reste limité malgré de nombreuses adaptations. C'est alors que fut créé le **XML**.

Le **XML** est un dérivé du **SGML**. Il se sert des principes de simplicité du **HTML** et de la souplesse **SGML**.

Intérêts

Le succès du XML est en grande partie dû à ses qualités qui sont essentiellement :

- Séparation stricte entre contenu et présentation
- Simplicité, universalité et extensibilité
- Format texte avec gestion des caractères spéciaux
- Structuration forte
- Modèles de documents (DTD et Schémas XML)
- Format libre

Séparation stricte entre contenu et présentation
Cela permet de séparer complètement l'information (le contenu) de son apparence (le contenant), et donc de fournir plusieurs types de sortie pour un même fichier de données (autre document **XML**, tableau, graphique, image, animation multimédia, fichier **HTML**, fichier **PDF**...)

Simplicité, universalité et extensibilité
Contrairement à HTML, le vocabulaire (ensemble des balises autorisées), n'est pas figé. Les noms des balises XML sont libres. Cette liberté dans les noms de balises permet de définir des vocabulaires particuliers adaptés aux différentes applications.

Modèles de documents (DTD et Schémas XML)
Il est nécessaire de fixer les règles que doivent respecter les documents pour pouvoir échanger et traiter de manière automatique ces documents.

Structuration forte
Bien que les données présentes dans un document XML soient fortement structurées.

Format texte avec gestion des caractères spéciaux
Le format XML est un format basé sur du texte. Un des atouts d'XML est sa prise en charge native des caractères spéciaux grâce à Unicode. De plus, il est possible d'utiliser les différents codages (UTF-8, Latin-1, …) possibles puisque l'entête d'un document spécifie le codage.

Format libre
Le langage XML est totalement libre développé par le W3C.

Langages apparentés

Un des atouts indéniables de XML est le nombre de technologies et de langages qui se sont développés autour de XML. La liste ci-dessous énumère les principaux langages qui font partie de l'environnement XML :

- XLink et XPointer (liens entre documents)
- XPath (langage de sélection)
- XQuery (langage de requête)
- Schémas XML (modèles de documents)
- XSLT (transformation de documents)

Dialectes

De très nombreux dialectes ont été définis pour appliquer XML à des domaines très variés. Le grand avantage est que ces différents dialectes partagent la même syntaxe de base et que tous les outils XML peuvent être utilisés pour spécifier et manipuler ces documents :

- RSS (Really Simple Syndication) : Abonnement à des flux de données
- XUL (XML-based User interface Language): Langage de description d'interfaces graphiques.
- SVG (Scalable Vector Graphics) : Description de dessins vectoriels
- SMIL (Synchronized Multimedia Integration Language): Description de contenus multimédia
- MathML (Mathematical Markup Language) : Description de formules mathématiques
- WSDL (Web Services Description Language) : Description de services WEB
- OpenStreetMap : Cartes libres
- XML Signature : Format pour les signatures électroniques
- SAML (Security Assertion Markup Language): Langage d'échange d'authentifications et

d'autorisations
- UBL (Universal Business Language) : Bibliothèque de documents standards pour les échanges commerciaux
- OpenDocument : Format de document pour les applications bureautiques.
- DocBook : Format de documentation technique

De nombreux projets informatiques, comme Ant ou Android utilisent XML pour le stockage de données et en particulier pour les fichiers de configuration.

Syntaxe et structure

Il y a, en français, l'orthographe et la grammaire. La première est constituée de règles pour la bonne écriture des mots. La seconde régit l'agencement des mots dans une phrase. Pour qu'une phrase en français soit correcte, il faut d'abord que les mots soient bien orthographiés et, ensuite, que la phrase soit bien construite. Il y aurait encore le niveau sémantique mais nous le laisserons de côté.

XML a également ces deux niveaux. Pour qu'un document XML soir correct, il doit être
- ***Bien formé*** : Un document bien formé doit respecter certaines règles syntaxiques propres à XML. Il s'agit, en quelque sorte, de l'orthographe d'XML.
- ***Valide*** : Un document valide doit respecter un *modèle de document* qui décrit de manière rigoureuse comment doit être organisé le document. Un modèle de documents peut être vu comme une grammaire pour des documents XML.

La différence essentielle avec le français est que la grammaire d'XML n'est pas figée. Pour chaque application, il est possible de choisir la grammaire la plus appropriée. Cette possibilité d'adapter la grammaire aux données confère une grande souplesse à XML.

Présentation de XML

Syntaxe de XML

Un fichier **XML** est composé d'un **prologue**, d'un élément **racine** et d'un **arbre**.

- Les premières lignes forment le **prologue**, constitué de la déclaration **XML**, puis éventuellement des instructions de traitement interprétées par les applications servant à traiter le document **XML** ;

  Exemple de déclaration :
  ```
  <?xml version="1.0" encoding="ISO-8859-1" standalone="yes"?>
  ```
  - **version** : version du XML utilisée dans le document;
  - **encoding** : le jeu de codage de caractères utilisé. Par défaut, **encoding** a la valeur **UTF-8**.
  - **standalone** : dépendance du document par rapport à une DTD ;

    standalone = **yes** : le processeur de l'application n'attend aucune DTD extérieure au document.

    standalone = **no** : le processeur attend une référence de déclaration de type de document. La valeur par défaut est **no**.

  Exemple d'instruction de traitement :
  ```
  <?xml-stylesheet type="text/xsl" href="biblio.xsl"?>
  ```

- Le second élément est une déclaration de type de document à l'aide d'un fichier annexe appelé **DTD** - *Document Type Definition*;
  Exemple de déclaration de DTD :
  ```
  <!DOCTYPE biblio SYSTEM "biblio.dtd">
  ```

- L'arbre est constitué d'éléments imbriqués les uns dans les autres (ayant une relation parent-enfant) et d'éléments adjacents.

La syntaxe de XML est relativement simple. Elle est constituée de quelques règles pour l'écriture d'un entête et des balises pour structurer les données. Ces règles sont très similaires à celles du langage HTML utilisé pour les pages WEB mais elles sont :

- Plus générales : car les noms des balises sont libres.
- Plus strictes : car elles imposent qu'à toute balise ouvrante corresponde une balise fermante.