Ontologies et Web Sémantique
Master 2

Chapitre I:XML

I.1.Introduction à XML

Mr Bougherara.S

Plan du cours

- Introduction
- Limites HTML
- Vers XML
- Conséquences XML
- Langages apparentés
- conclusion

Introduction

- Le langage XML dérive de SGML (Standard Generalized Markup Language) et de HTML (HyperText Markup Language).

- SGML qui a été introduit en 1986 par C. Goldfarb. SGML a été conçu pour des documentations techniques de grande ampleur. Sa grande complexité a freiné son utilisation en dehors des projets de grande envergure

- Comme ces derniers, XML est un langage orienté texte et formé de *balises* qui permettent d'organiser les données de manière structurée

Introduction

- Le langage XML (eXtended Markup Language) est un format général de documents orienté texte. Il s'est imposé comme un standard incontournable de l'informatique. Il est aussi bien utilisé pour le stockage de document que pour la transmission de données entre applications.
- Sa simplicité, sa flexibilité et ses possibilités d'extension ont permis de l'adapter à de multiples domaines allant des données géographiques au dessin vectoriel en passant par les échanges commerciaux

HTML

- HTML est un **langage de balises** comme SGML :
  - Balises: normalisation **d'un ensemble figé** de symboles encadrés par "<" et ">".
- Les balises HTML définissent des directives de mise en page d'un texte encadré par un couple balise ouvrante, balise fermante: <H2> xxxx </H2> un titre

- HTML peut rester très longtemps utilisé comme langage graphique pour les navigateurs WEB

Limites HTML

- **HTML: un langage non extensible**
  - Le langage HTML a été conçu pour être **assez simple** de sorte que le nombre et la signification des balises est **limité**.
    - Ex: Il n'existe pas de balisage pour la représentation des données en chimie (molécules, formules, valeurs numériques)
  - Le langage HTML est **figé**.
    - Toutes les balises utilisables sont définies au départ ce qui est intenable si l'on voulait prendre globalement en compte les besoins d'un grand nombre de métiers.

Limites HTML

- **HTML: un langage de description de documents non structurés**
  - HTML permet de définir de façon beaucoup trop **limitée** la **structure** d'un document.
  - Il n'y a en fait pas de **vérification d'une structure** pour le document que l'on peut définir.
    - Ex: on peut créer un document commençant par une tête de chapitre H2 et poursuivant par une tête de chapitre H1.
- **HTML définit en fait un univers de documents plats.**
  - Une recherche doit considérer un document HTML comme **une chaîne de caractères**.
  - Pas de moyen de partager entre communicants **une structure de document** préétablie.

Limites HTML

- **Non séparation du document et de sa présentation graphique.**
- **HTML non-orienté contenu : pas d ’info sémantique**
- **Inadapté** à l’échange de données

Vers XML (Extensible Markup Language)

- **SGML** riche, lourd, mal adapté au Web.
- **HTML** adapté à la présentation graphique, mais limité par un ensemble de balises figé, non extensible et sans typage.
- **Groupe de travail XML** à partir de août 1996 qui est composé essentiellement de membres du groupe de travail SGML.
  - **Recherche d'un langage assez simple** mais présentant une richesse proche de SGML.
  - **XML : un sous-ensemble de SGML**, qui élimine des points trop ciblés sur certains besoins.
  - Un document XML est **conforme** SGML.

Principes XML

- **séparer la structure d'un document de sa présentation**

•distinguer le contenu d'un document et la présentation qui en est donnée. Un même contenu peut être rendu de façons très différentes. (**html, pdf……**)

Principes XML

- **Une structuration forte du document**

HTML	XML
Olivier Carton 175, rue du Chevaleret 75013 Paris <tt>Olivier.Carton@liafa.jussieu.fr</tt>	<personne> <nomComplet> <nom> Carton </nom> <prenom> Olivier </prenom> </nomComplet> <dresse> 75013 Paris </dresse> <email> Olivier.Carton@liafa.jussieu.fr </email> </personne>

Principes XML

- **Une structuration forte du document**

XML

```
<personne>
        <nomComplet>
                <nom> Carton </nom>
                <prenom> Olivier </prenom>
        </nomComplet>
        <dresse> 75013 Paris </dresse>
        <email> Olivier.Carton@liafa.jussieu.fr  </email>
</personne>
```

Les balises (<personnes> ,<Nom> …) sont des éléments syntaxiques destinés à structurer le contenu.

Principes XML

- **Simplicité, universalité et extensibilité**

- les noms des balises XML sont libres. Il appartient aux auteurs de documents de fixer les balises utilisées.

- il est seulement nécessaire que les auteurs s'entendent sur le vocabulaire, c'est-à-dire la liste des balises utilisées, lorsque des documents sont échangés.

- Cette liberté dans les noms de balises permet de définir des vocabulaires particuliers adaptés aux différentes applications.

Principes XML

- **Interopérabilité**

- Pouvoir échanger et traiter une donnée en utilisant de nombreux types de logiciels.

Principes XML

- **Validation**

•La liberté dans le choix des noms de balises implique une contrepartie. Il devient nécessaire de fixer des règles que doivent respecter les documents.

•Sans ces règles, il n'est pas possible d'échanger et de traiter de manière automatique ces documents.

Ces règles doivent d'abord fixer le vocabulaire mais aussi les relations entre les balises.

- Les règles peuvent, par exemple, imposer qu'une balise <nomComplet> contiennent exactement une balise <nom> et une balise <prenom> sans pour autant fixer l'ordre de ces deux balises.

Langages apparentés

- plusieurs technologies et de langages se sont développés autour de XML. Ceux-ci enrichissent les outils pour la manipulation des documents XML:
- Xpat,xpointer,xquery, Xpath, XSLT …

Langages apparentés

- XLink et XPointer : (liens entre documents)

un mécanisme pour matérialiser des liens entre des éléments d'un document

- .XPath (langage de sélection)XPath est un langage d'expressions permettant de sélectionner des éléments dans un document XML. Il est la pierre angulaire du langage XSLT pour la transformation de documents

Langages apparentés

- XQuery est un langage permettant d'extraire des informations à partir d'un ou plusieurs documents XML et de synthétiser de nouvelles informations à partir de celles extraites. Il s'apparente à un langage d'interrogation de bases de données et joue le rôle de SQL pour les documents XML.

Langages apparentés

- Schémas XML (modèles de documents)

Pour la validation

- XSLT est un langage permettant d'exprimer facilement des transformations complexes des documents XML en d'autres formats comme PDF ou des pages HTML. Il s'appuie sur la structuration forte des documents XML vus comme des arbres. Chaque transformation est décrite par des règles pour chacun des éléments du document.