Travaux dirigés : documents XML et méta-données

1 Objectif

Apprendre quelques bases des langages XML. Voir comment insérer des méta-données dans un document.

2 Bases d'XML

2.1 La syntaxe des balises XML

XML (*eXtensible Markup Language*) est une syntaxe de **balises**. Les balises XML s'utilisent sous la forme `<balise> contenu </balise>`. Une balise sans contenu `<balise> </balise>` s'écrit plus simplement `<balise/>`. Le contenu entre deux balises peut être en général d'autres balises ou du texte simple. Dans un document, tout élément ouvert doit être fermé. De plus, les éléments doivent être correctement imbriqués : `<h1>Résumé de la <em>Critique de la raison pure</h1></em>` n'est pas correct, on doit écrire `<h1>Résumé de la <em>Critique de la raison pure</em></h1>`.

Une balise XML peut avoir des **attributs** associés, sous la forme `nom="valeur"`, qui la raffinent. Par exemple, on pourrait avoir une balise `<chapitre classe="bibliographie"> contenu </chapitre>` pour préciser que le chapitre en question est une bibliographie.

On utilise des entités particulières pour représenter les caractères réservés : `&lt;` pour < (*less than*), `&gt;` pour > (*greater than*), `&amp;` pour & (*ampersand*).

Cette syntaxe est particulièrement adaptée aux données hiérarchiques. On peut aisément représenter un document XML sous la forme d'un arbre.

2.2 Utilisation pour XHTML

Il est possible d'écrire des documents HTML dans une syntaxe XML : on parle alors de document XHTML. Le document sur la page `http://www.lsv.ens-cachan.fr/~schmitz/teach/2011_web/` est écrit en XHTML. Enregistrez-le sous le nom `test.html` dans le répertoire `~/public_html` de votre machine (si vous travaillez sur votre portable, il faudra le copier par `ssh` sur la machine `acces1.rip.ens-cachan.fr` à chaque fois que vous voudrez tester votre fichier).

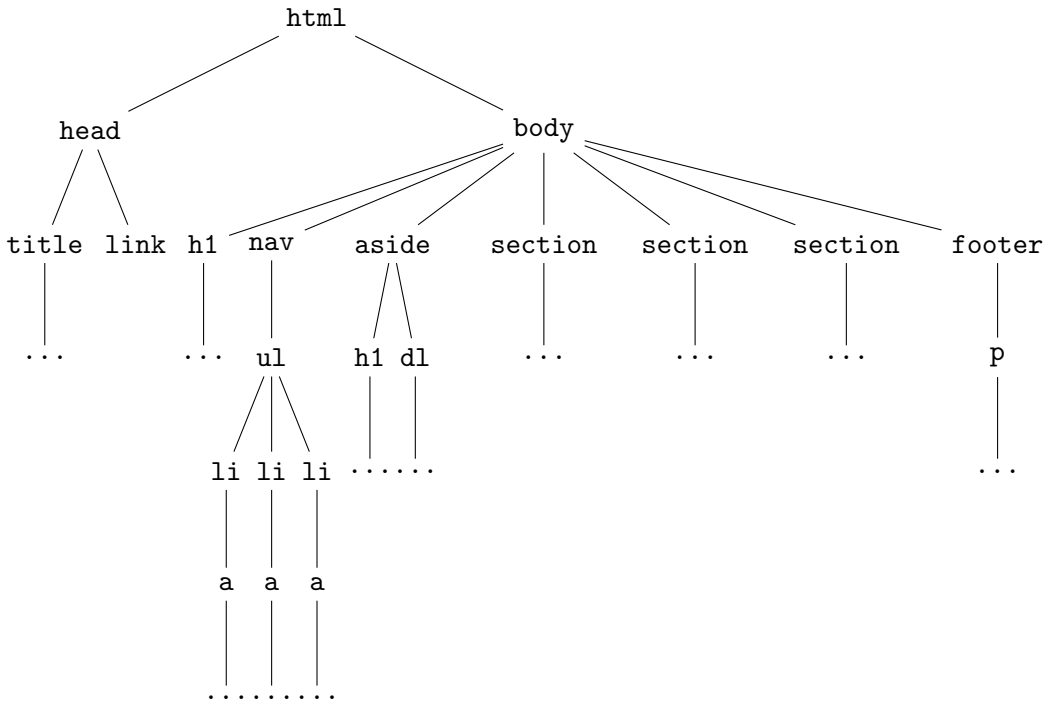
Ouvrez le document : il commence par une ligne de **déclaration**

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

qui annonce qu'il s'agit d'un document XML et son codage de caractères (cette déclaration est inutile si vous travaillez en UTF-8). Puis vient la première balise **racine** du document,

```
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="fr">
```

qui dénote un document dans le dialecte HTML, en français. Cette balise contient deux balises filles, `<head>` et `<body>`, qui elles-même contiennent d'autres balises ou du texte etc :



Le langage XHTML est l'un des nombreux **dialectes** XML : il s'agit d'un ensemble standard de balises avec leurs règles d'utilisation (quelles balises peuvent contenir quelles autres balises, en quel nombre, dans quel ordre, avec quels attributs). Vous pouvez essayer de déchiffrer le sens des balises qui apparaissent dans le document : qu'est-ce qu'un `<h1>` ? un `<nav>` ? un `<ul>`, `<ol>` ou `<li>` ? Le document est en fait rédigé en suivant les règles (encore incomplètes) de la version 5 d'HTML ; voir `http://dev.w3.org/html5/spec/spec.html` pour une référence complète.

3 Méta-données Dublin Core

La norme de méta-données du Dublin Core est un ensemble standard de **termes**, développé en particulier pour échanger des méta-informations entre archives ouvertes, et décrit à l'adresse `http://dublincore.org/documents/dcmi-terms/`. On trouvera par exemple dans cet ensemble les termes *creator*, *title*, *language*, etc.

Ces méta-données peuvent être insérées dans des documents XML en utilisant les possibilités de **mélange** de dialectes (qui justifie le nom *eXtensible*). Pour des raisons de compatibilité avec HTML, on a néanmoins l'habitude de les ajouter dans la balise `<head>` sous la forme de balises `<meta>`. Concrètement, on ajoute une balise

```
<meta name="nom du terme" content="méta-donnée correspondante" />
```

On peut par exemple ajouter une méta-donnée de titre à un document XHTML via

```
<head>
  <title>Mon titre</title>
  ...
  <link rel="schema.DC" href="http://purl.org/dc/terms/" />
  <meta name="DC.title" content="Mon titre" />
</head>
```

D'autres méta-données peuvent être ajoutées de la même manière. Complétez les méta-données de votre document `~/public_html/test.html`.

4 Micro-données

Les méta-données Dublin Core, si elles sont bien lues par les moteurs de recherche, sont assez faiblement utilisées par ceux-ci dans leur établissement du rang des pages : comme elles ne sont pas vues par les visiteurs de la page, elles peuvent en être totalement déconnectées. Les principaux moteurs de recherche font depuis peu la promotion de **micro-formats**, où les méta-données sont insérées dans des attributs dans le contenu principal de la page – cela impose que ce contenu soit réellement consulté.

Nous allons voir comment utiliser les micro-formats de `http://schema.org` dans un document XHTML. Plusieurs schémas d'annotations sont définis sur ce site, par exemple Event, EducationalOrganization, ou encore Person. L'annotation se fait en délimitant la portion du contenu concernée par les attributs `itemscope=""` et `itemtype="URL du schéma"`, tandis que les différentes données sont introduites par des attributs `itemprop="nom de propriété"`. Plus concrètement, voici comment annoter le document en tant que WebPage et renseigner sa date de modification :

```
<body itemscope="" itemtype="http://schema.org/WebPage">
  ...
  <footer>
    <p>Dernière modification :
      <time itemprop="dateModified">2012-05-09</time></p>
  </footer>
</body>
```

Il a été nécessaire d'ajouter une balise `<time>` pour identifier la portion du contenu contenant la propriété `dateModified` ; plus généralement, on utilisera des balises `<span>` pour délimiter des portions de texte (on aurait pu écrire

```
<span itemprop="dateModified">2012-05-09</span>
```

cela aurait été moins informatif).

Annotez autant que possible le document `~/public_html/test.html`. Utilisez l'outil de validation Google pour vérifier que vos annotations sont bien interprétées : `http://www.google.com/webmasters/tools/richsnippets` ; l'URL à fournir pour vos tests est `http://rip.ens-cachan.fr/~fcXX/test`.