Gestion Électronique de Documents et XML

Master 2 TSM

Introduction

Les formats de données

Formats de donnée

- **Format de donnée** : manière de représenter des informations dans un document informatique (sous forme binaire)

Exemple des images :

- Bitmap (.bmp)
- JPEG (.jpg)
- GIF (.gif)
- ...

Un logiciel doit connaître le format d'un fichier pour pouvoir l'exploiter !

Création de documents : exemple de la musique

Chacun crée sa musique selon des formats « standards » (CD, MP3, ...)

- Formats bien connus, répondant parfaitement aux besoins
- Tous les appareils savent les lire parfaitement
- Un nouveau CD est lu sur une chaîne ancienne sans problèmes et inversement
- Je sais lire parfaitement vos CD !
- Les traitements automatiques sont « simples »

ET TOUT CELA PARAIT NORMAL !

Pour les autres documents ?
Exemple : les feuilles de comptabilité

- Pas de format standard, ouvert...
- vous ne savez pas lire mes fichiers
- je ne sais pas lire vos fichiers

Sauf si nous avons exactement le même logiciel (même version, ...)

Création de documents

Situation «classique» :

- Un logiciel propriétaire (Word ?)
- Chaque «auteur» a son style, sa présentation, voire son logiciel

Problèmes :

Non homogénéité de la présentation

Le Traitement automatique de l'information est TRES difficile :

- Réunir l'information,
- la synthétiser,
- La stocker de manière unique
- **Interopérabilité** faible !

une première solution ? l'uniformisation !

Un seul **logiciel** pour tous !

- Simplifie le traitement
- Oui, mais si on veut en changer ? (coût, complexité, ...)

La même présentation pour tous !

- Oui, mais si on en change ? (les styles ?)

Et les traitements automatiques ?

- Achat de solutions propriétaires (=onéreuses et pas souvent existante)

Solution moderne :

Un seul **FORMAT** pour tous !

Un format **OUVERT** :

- on sait ce qu'il y a dedans !
- Il ne dépend pas du logiciel
- À l'opposé de format *propriétaire* ou *fermé*
- Les traitements sont «simples»
- En général pérenne, et si non, facile à changer

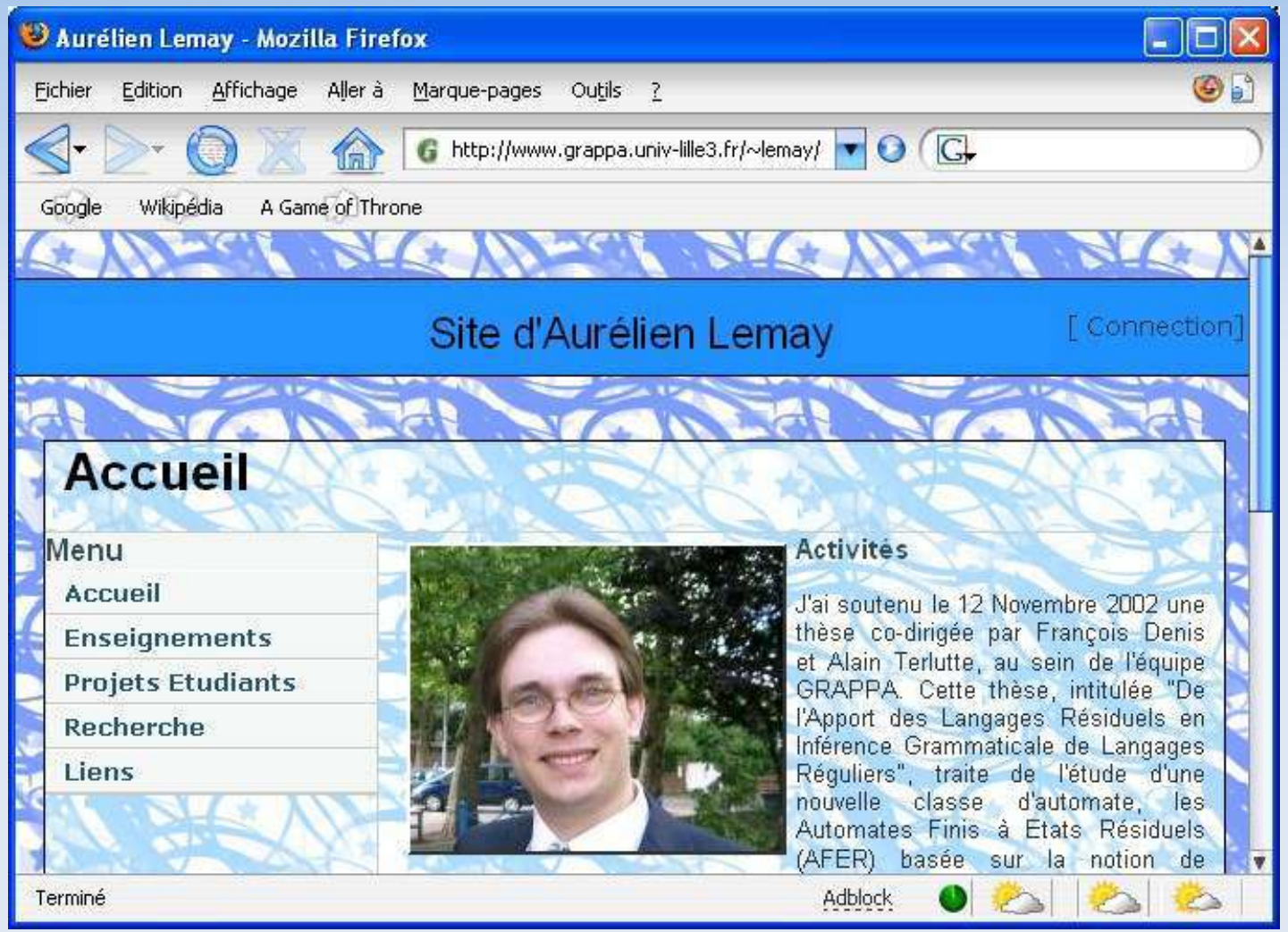
Et pour la présentation ?

Présentation SEPAREE du contenue !

Un document est composée de

- Un contenu (les *données*)
  - de l'information, du texte
  - Des images, ...
- Des règles de présentation (un style)
  - L'écriture, la couleur, ...
  - Ou sont mises les images, comment sont écrits les titres, les sous-titres...

Un bon exemple : HTML + CSS



HTML + CSS

```
<html>
<head>
  <title> Ma page </title>
  ...
</head>
<body>
  <div id=title>Le site ...
  </div>
  <img src=moi.jpg>
  <h1>  Activités </h1>
  J'ai soutenu le 12 Nov....
...
```

Un fichier HTML (la page web)

```
body {
  background-image :
    url(image.jpg);
}

#title {
  color: #00000;
  background-color:
    #1E90FF;
}
...
```

Un fichier CSS (feuille de style)

le XML : un format universel ?

Le XML : eXtensible Markup Language

- **Simple : métaformat** à base de balises

<truc> ... </truc>

- **Lisible** par des humains : en texte...

- **Puissant** :Tout document informatique peut être représenté en XML !

Syntaxe XML : en quatre points

- des balises ouvrantes et fermantes

<gras> Bonjour **</gras>**

- Des attributs

<texte **gras='oui'**> Bonjour </texte>

- balise sans contenu

<texte/>

équivaut à :

<texte></texte>

- Commentaires

<!-- bla bla bla --!>

XML Partout

- Du texte :

```
<livre>
   <titre> Oui Oui à la plage </livre>
<chapitre>
  <titre> A la plage ! </titre>
  <texte>
Un matin, le nain Potiron arrive en courant à la petite maison
de son ami Oui-Oui. <citation> Si nous allions à la plage ?
</citation>  propose-t-il. ...
</texte>
</chapitre>
...
</livre>
```

XML Partout

- de la musique :

```
<musique>
<instrument>
<type> Piano >/type>
<partition>
<note><hauteur>5</hauteur><durée>1</durée></note>
<note><hauteur>6</hauteur><durée>1</durée></note>

...
```

XML Partout

- des dessins :

```
<dessin>
  <forme>
    <type> Rectangle</type>
   <x> 10 </x> <y> 20 </y>
  <largeur> 30 </largeur> <hauteur> 20 </hauteur>
 <couleur> rouge </couleur>
</forme>
...
</dessin>
```

XML n'est pas un format !

- plusieurs manières de représenter la même chose :

```
<cv>    <nom>Lemay</nom>
<prenom>Aurélien</prenom>
<emploi> Maitre de conférence </emploi>   ...
```

```
<curiculum>
<identity name= 'Lemay' firstname='Aurélien' />
<background>
<job from='2002'>Maître de conférence</job> ...
```

Il me faut connaître l'organisation et la signification des balises !

Formats XML

Il faut préciser :

- les balises à utiliser
- les attributs
- comment les combiner ...
- Etc, etc...

Un format XML : règles d'utilisation du XML, décrit dans une **DTD** ou un **schéma XML**.

quelques formats XML

- Texte : Docbook, TEI, EAD, Opendocument (Openoffice, ...), Word (depuis 2003)
- Musique : musicXML
- Syndication : RSS, Atom
- Bibliographie : MODS, Mets
- formules mathématiques : MathML
- Dessin : SVG
- Web : XHTML
- ...

Exemple : le XHTML

- XHTML : format XML utilisé pour créer des pages WEB.
- Très proche cousin du HTML

Même ensemble de balises...

- Différent du HTML ! plus strict.

Par exemple, en HTML, il n'est pas obligatoire de fermer certaines balises

Exemple de XHTML

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
    "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">

<html xmlns="http://www.w3.org/1999/xhtml">

<head> <title>Exemple XHTML 1</title> </head>

<body> <ul>

<li>Tous les éléments doivent être explicitement balisés.</li>

<li>Les balises fermantes ne sont pas optionnelles.</li>

<li>Les noms d'éléments et d'attributs <em
    class="important">doivent</em> être en minuscules.</li>

<li>Tous les attributs doivent avoir une valeur explicite <table
    border="1"><tr><td>x</td></tr></table>.</li>

<li>Les guillemets sont <em class="important">toujours</em> obligatoires
    autour des valeurs d'attribut.</li>

<li>Les éléments vides doivent être fermés <img src="i.png" alt="i"/>.</li>

</ul> </body> </html>
```

Limitations de XML

- très verbeux (taille énorme de documents)

Mais se compresse extrêmement bien

- les traitements peuvent être long

du fait de la longueur des documents

- Lisibilité parfois relative

Contenu texte mais que comprendre de :

<data> HtJK0£28DC30F </data>

- XML seul ne définit pas grand chose !

Nécessité d'avoir une DTD ou un schéma !

Mini-projet

Fonctionnement d'une base de données XML

- Une base de données XML = un ensemble de documents XML répondant à une DTD

Exemple :

un ensemble de recettes de cuisines

un ensemble de livres

...

Comment utiliser la base ?

- La questionner ? (requêtes Xquery)

Ex : Combien d'oeufs pour un gateau au yahourt

- L'afficher ? (format XHTML)
- L'imprimer (PDF ?)
- Récupérer les données pour les mettre dans une autre base de données ?

➔Transformer les données !

Transformation des données ?

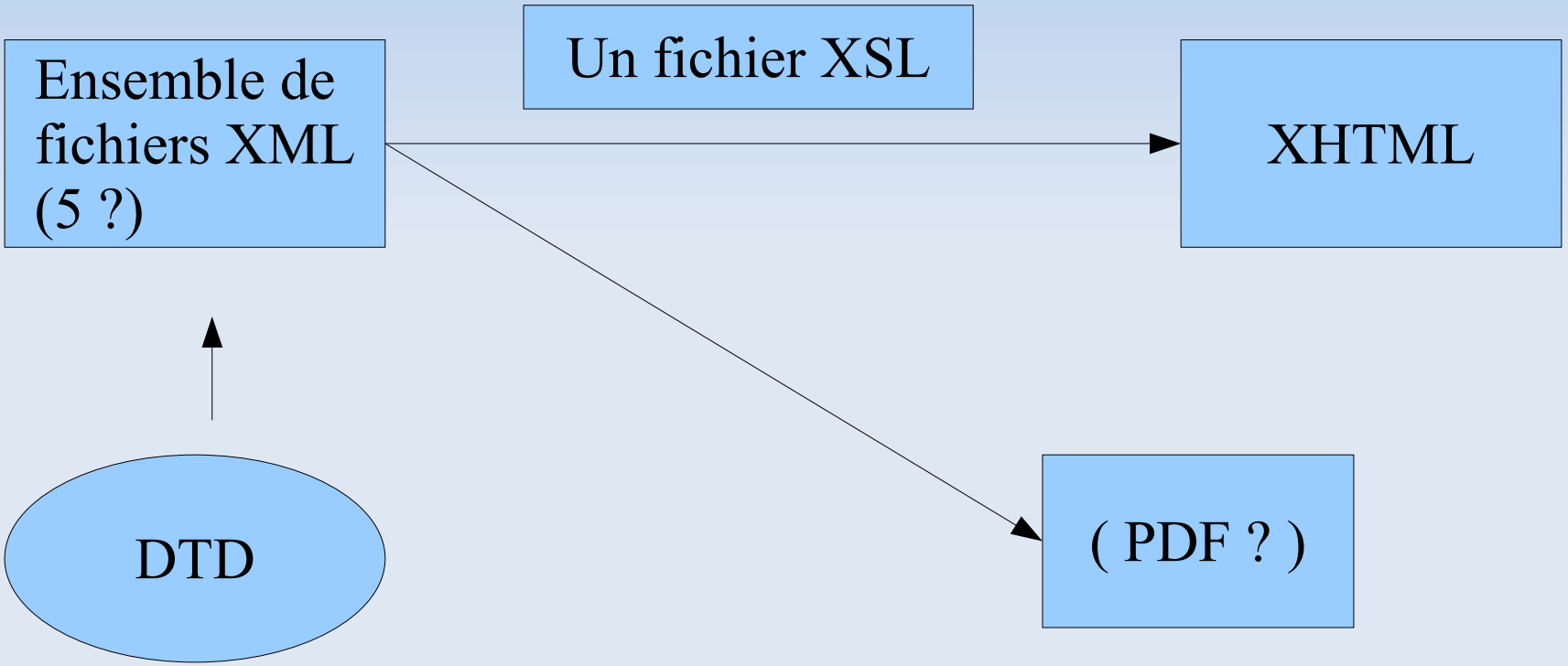
- Transformer des documents d'un format à un autre

Ex : transformer mes fiches cuisines en pages web (XHTML)

Pour transformer des documents XML :

feuille de transformation **XSL**

Mini-projet



Mini-projet

- Par groupes de 2 personnes
- Choix d'un type de données

Fiches de films, d'acteurs, de voyages, de recettes de cuisine, ...

- Ecriture de la DTD
- Créer des documents (5?) respectant la DTD
- Créer un exemple de fichier HTML (avec CSS)
- Créer le fichier de transformation XSL