Notes du cours

XML (eXtensible Markup Language)

Introduction

- XML a été approuvé par le World Wide Web Consortium (W3C) en février 1998. On peut dire que XML n'est pas réellement un nouveau langage, c'est un méta-langage qui est utilisé pour définir d'autres langages.

- XML est un langage de description de documents structurés. Comme SGML dont il est issu, il permet de décrire la structure d'un document sans tenir compte de sa description physique (mise en page, choix de polices de caractères, en-tête, marges ...), toutes ces propriétés physiques sont définies dans un autre document, la feuille de style. Le point commun le plus important avec SGML est le fait que tout document XML peut être basé sur une DTD ou un Schéma (cette association n'est pas obligatoire, un fichier XML peut se suffire à lui même.

- XML est un langage balisé comme HTML. Mais alors que HTML utilise des balises prédéfinies, XML utilise les balises seulement pour délimiter les éléments de données et laisse l'entière interprétation des données à l'application qui les lit. XML possède des règles de syntaxe beaucoup plus strictes que HTML.

- XML est en format texte; cela permet de consulter les données sans le programme qui les a produites. Ceci permet aux programmeurs ou utilisateurs avancés de déboguer facilement des applications.

- XML utilise des balises pour délimiter les données, donc les fichiers sont souvent volumineux. Il s'agit d'une décision prise en toute conscience par les développeurs de XML. L'espace disque n'est plus aussi coûteux qu'auparavant, de plus les protocoles de communication peuvent compresser des données à la volée, ceci permet de faire transiter du texte aussi efficacement qu'un fichier en format binaire, et en l'occurrence d'économiser la bande passante.

- Autour de la spécification XML qui définit les balises et les attributs, il existe de nombreux modules facultatifs qui fournissent des balises et des attributs comme Xlink (norme pour l'ajout des liens hypertextes dans les documents XML), XSLT (langage de transformation de documents XML), schémas XML (aide à définir de nouveaux formats basés sur XML, etc.

Quelques acronymes :

SGML : Standard Generalized Markup Language est bien adapté et utilisé dans l'industrie pour créer des documents techniques ou lexicographiques très volumineux (encyclopédies, spécifications techniques d'un système), XML est une version abrégé de SGML qui permet de définir ses propres types de documents plus facilement.

CSS : Cascading Style Sheet : la réalisation physique d'un document XML est pilotée par une feuille de style. Une feuille de style CSS est composée de règles de style qui s'appliquent aux divers éléments du document traité.

DTD : Document Type Definition : Une DTD est un fichier écrit en XML, qui contient une définition formelle d'un type particulier de document. Elle définit les noms qui peuvent être utilisés pour les types d'éléments (<section>, <para>, ...) où ils peuvent apparaître et comment ils s'organisent les uns par rapport

aux autres. La définition de document type est donc un mécanisme de spécification de structure, elle permettra en outre de vérifier si le document XML répond à ces spécifications (notion du document valide). Valider un document consiste à le soumettre à un analyseur syntaxique (parseur) qui vérifie si : le document respecte bien la syntaxe du langage XML et si le document respecte bien la structure définie par la DTD.

XSLT (Extensible Style Language Transformations) est, comme son nom l'indique, un langage destiné à *transformer* un fichier XML en quelque chose d'autre. Ce quelque chose d'autre sera le plus souvent un fichier XML ou HTML. Mais ce pourra être tout aussi bien un fichier d'un autre format : par exemple du texte pur, ou du *Rich Text Format*...

Composition du document XML

Un document XML se compose :

- d'un prologue (facultatif mais conseillé) : il peut contenir une déclaration XML, une déclaration de type de document

- d'un arbre d'éléments : il constitue le contenu du document XML. Chaque élément est composé d'une balise d'ouverture, d'un contenu d'élément et d'une balise de fermeture :

  Exemple 1. Élément paragraphe

```
<para>A gauche on trouve l'ouverture de la balise para. L'ensemble du
texte constitue le contenu de l'élément. A droite, on trouve la
fermeture de la balise para </para>.
```

- de commentaires (facultatifs) : ils permettent une maintenance facile du document (mise à jour, modification par un autre utilisateur).

  Exemple 2. Commentaire

```
<!-- ceci est un commentaire -->
```

Si vous utilisez le mode xml sous XEmacs[1]., vous pouvez placer les balises grâce à un menu contextuel accessible avec le bouton droit de la souris.

Génération des documents au format HTML:

Pour générer les documents de sortie au format HTML, il suffit de lancer la commande comme s'est décrit ci-dessous :

Exemple 3.

Le fichier source XML se trouve dans /fichiers/xml_doc

Les fichiers de sorties au format HTML seront générés dans /fichiers/xml_doc/xml2html

[1] Pour lancer XML avec Xemacs, appuyer sur M-x puis dans le mini buffer situé en bas de la fenêtre, lancer le mode XML avec la commande xml-mode.
- C-c C-p : lance la lecture de la *dtd* déclarée au début du document.
- C-c C-v : valide la conformité du document avec la *dtd*, il est possible de se passer de cette validation, *jade* le post-processeur qui générera le fichier de sortie l'assure aussi.

On dispose d'un fichier mapage_html.dsl qui se trouve dans le répertoire /fichiers/xml_doc . En se plaçant dans le répertoire cible fichiers/xml_doc/xml2html, lancer la commande :

```
jade -d ../mapage_html.dsl -t sgml /usr/lib/sgml/declaration/xml.decl
../essai1.xml
```

Suite de cette commande on observe qu'il existe des fichiers *.png dans le répertoire cible. Il existe aussi un fichier style.css dans le répertoire cible. Et que le post-processeur jade, a généré n documents au format HTML dans ce même répertoire.

Exemple 4: une bibliographie

```
<!-- Prologue -->
<?xml version="1.0" encoding="ISO-8859-1"?>
<!-- Élément racine -->
<biblio>
  <!-- Premier enfant -->
  <livre>
    <!-- Élément enfant titre -->
    <titre>Les Misérables</titre>
    <auteur>Victor Hugo</auteur>
    <nb_tomes>3</nb_tomes>
  </livre>
  <livre>
    <titre>L'Assomoir</titre>
    <auteur>Émile Zola</auteur>
  </livre>
  <livre lang="en">
    <titre>David Copperfield</titre>
    <auteur>Charles Dickens</auteur>
    <nb_tomes>3</nb_tomes>
  </livre>
</biblio>
```

L'élément-racine (en anglais : document element) est, comme son nom l'indique, la base du document XML. Il est unique et englobe tous les autres éléments. Il s'ouvre juste après le prologue, et se ferme à la toute fin du document. Dans l'exemple ci-dessus, l'élément racine est biblio.

Les premières lignes forment le prologue, constitué dans l'exemple précédent de la déclaration XML, et éventuellement d'une déclaration de type de document (une DTD) ;

L'élément biblio est notre élément racine (en anglais : document element) ; il est constitué de trois éléments livre. Dans chacun d'entre eux nous retrouvons la même composition, c'est-à-dire : un élément titre, un élément auteur et éventuellement un élément nb_tomes. L'élément livre, de plus, a un attribut lang ;

Même s'il est simple de comprendre ce code, on s'aperçoit mieux d'une éventuelle erreur lorsqu'on visualise ce même fichier dans un navigateur.

Une instruction de traitement est une instruction interprétée par l'application servant à traiter le document XML. Elle ne fait pas totalement partie du document. Les instructions de traitement qui servent le plus souvent sont la déclaration XML ainsi que la déclaration de feuille de style. Exemple d'instruction de traitement :

```
<?xml-stylesheet type="text/xsl" href="biblio.xsl"?>
```

Dans cet exemple, l'application est xml-stylesheet, le processeur de feuille de style du XML. Deux feuilles de style différentes peuvent être utilisées, les XSL (propres au XML) ainsi que les CSS (feuilles de style apparues avec le HTML). L'attribut type indique de quel type de fichier il s'agit (text/css pour les feuilles de style CSS, par exemple) et l'attribut href indique l'URL du fichier. Cette instruction de traitement est notamment utilisée par les navigateurs Internet pour la mise en forme du document.

Les éléments forment la structure même du document : ce sont les branches et les feuilles de l'arborescence. Ils peuvent contenir du texte, ou bien d'autres éléments, qui sont alors appelés "éléments enfants", l'élément contenant étant quant à lui appelé logiquement "élément parent".

Exemple d'élément contenant du texte :

```
<titre>Les Misérables</titre>
```

Exemple d'élément contenant d'autres éléments :

```
<livre>
  <titre>L'Assomoir</titre>
  <auteur>Émile Zola</auteur>
</livre>
```

Les attributs

Tous les éléments peuvent contenir un ou plusieurs attributs. Chaque élément ne peut contenir qu'une fois le même attribut. Un attribut est composé d'un nom et d'une valeur. Il ne peut être présent que dans la balise *ouvrante* de l'élément (par exemple, on n'a pas le droit d'écrire `</livre lang="en">`).

Exemple d'utilisation d'un élément avec attribut :

```
<instrument type="vent">trompette</instrument>
```

Exemple d'utilisation d'un élément vide avec attributs :

```
<img src="ours.gif" alt="Gros ours" width="56" height="100" />
```

Les entités

Il existe deux sortes d'entités, définissables et définies. Elles peuvent être analysables ou non, internes ou externes. La déclaration des entités s'effectue au sein de la `DTD`. Elles peuvent être utilisées aussi bien dans la `DTD` que dans le document `XML`. Nous reviendrons plus en détails sur les entités et leur utilisation ultérieurement.

Certains caractères ayant un sens précis en `XML`, il est nécessaire de leur trouver un remplaçant lorsque l'on a besoin de les insérer dans un document. On a recours dans ce cas à des entités prédéfinies. Ces entités sont :

Caractère	***Entité***
&	&amp;
<	&lt;
>	&gt;
"	&quot;

'	&aquot;

Table 1. Liste des entités prédéfinies

Il n'existe pas d'entité prédéfinie pour les lettres accentuées ou pour les alphabets latins. Il faut utiliser à la place les entités numériques du type `&#n;` (où `n` est une valeur décimale). La valeur numérique correspond au code ISO 10646 ; par exemple le caractère *é* est codé par l'entité numérique `&#233;`. Il est néanmoins possible d'importer des entités en provenance d'une autre `DTD`, notamment celle du `HTML`.

Les sections CDATA

Une section `CDATA` est une section pouvant contenir toute sorte de chaîne de caractères. Une section `CDATA` permet de définir un bloc de caractères ne devant pas être analysés par le processeur `XML`. Ceci permet entre autres de garder dans un bloc de texte un exemple de code à afficher tel quel.

Exemple d'utilisation de `CDATA` :

<![CDATA[Une balise commence par un < et se termine par un >.]]>

Règles de composition:

Un certain nombre de règles de base doivent être respectées :

1. Un nom d'élément ne peut commencer par un chiffre. Si le nom n'est composé que d'un seul caractère, ce doit être une lettre comprise entre "a" et "z" pour les minuscules, "A" et "Z" pour les majuscules. S'il est composé d'au moins deux caractères, le premier peut être "_" ou ":". Le nom peut ensuite être composé de lettres, chiffres, tirets, tirets bas et deux points. La syntaxe `XML` est sensible à la casse (le format distingue majuscules et minuscules).
2. Toutes les balises portant un contenu non vide doivent être fermées. La balise de début, la balise de terminaison et le contenu entre deux sont globalement appelés *élément* ;
3. Les balises n'ayant pas de contenu doivent se terminer par /> (voir la balise `<img ...>` ci-dessus) ;
4. Les noms d'attributs sont *en minuscules* ;
5. Les valeurs d'attributs doivent être *entre guillemets* ;

Un document respectant ces critères est dit *bien formé* (well formed).

Il est aussi possible de définir des règles plus strictes indiquant quelles sont les séquences et imbrications de balises autorisées. Cela se fait à l'aide d'une `DTD` ou d'un `Schéma`. Il est alors possible d'effectuer une "validation" des documents faisant référence à une `DTD` pour s'assurer qu'ils respectent bien les règles qui y sont mentionnées. Un document bien formé dont la syntaxe est conforme aux règles stipulées dans une DTD ou un Schema XML est dit *valide*. Nous reviendrons sur cette notion ultérieurement.

Langages "orientés-contenu" et "orientés-présentation"

XML, est, de même que son grand ancêtre SGML, un langage de balisage universel. Il peut donc, comme SGML, servir à encapsuler toutes sortes de données -- à la seule condition qu'elles soient représentables sous forme d'arborescence. En particulier il peut parfaitement servir à encapsuler des données relatives à la manière de présenter des informations sur un support. C'est donc un raccourci un peu inexact de dire que

XML est orienté-contenu et non pas orienté présentation -- puisque l'orienté-présentation est seulement un cas particulier de l'orienté-contenu ! Rappelons au passage que le langage de présentation favori du Web, HTML, est lui-même une application particulière de SGML (ce qui le rend à quelques détails près conforme à la syntaxe XML -- son successeur XHTML le sera complètement). Et qu'une myriade de nouveaux langages de présentation sont en train d'apparaître (XHTML, XSL-FO, SVG, X3D...) qui seront conformes à la syntaxe XML.

XML et feuilles de style

Ces clarifications apportées, il reste qu'un fichier XML n'est pas, en général, un fichier affichable/présentable en l'état. Il faut donc lui ajouter quelque chose pour que cet affichage soit possible. Ce quelque chose a été appelé "feuille de style", par une analogie un peu boiteuse avec les feuilles de styles *stricto sensu* -- comme les feuilles de style CSS ou les styles de MS Word -- qui servent à associer (de manière centralisée) des caractéristiques typographiques (marges, alignements, polices et tailles de caractères, couleurs, etc.) à un contenu déjà orienté-présentation.

En XML une feuille de style *stricto sensu* n'est bien entendu pas suffisante. Si votre XML contient, par exemple, une bibliographie, vous pouvez certes l'associer directement à un feuille de style CSS qui vous permettra, par exemple, d'associer à l'élément *auteur* la police Verdana 14 points et la couleur *teal*. Mais une telle feuiIle de style CSS ne vous permettra pas de spécifier :

> que vous voulez que la bibliographie soit présentée sous la forme d'un tableau, ou sous la forme d'une liste ;
> qu'elle doit être classée selon tel ou tel critère ;
> que les différentes informations relatives à un même livre (auteur, titre, éditeur...) devront apparaître dans tel ou tel ordre, avec tels ou tels séparateurs, etc.

On voir par cet exemple que pour qu'un fichier XML puisse être affiché de manière réellement intéressante, il nous faut pouvoir spécifier :

> non seulement les objets de présentation *génériques*, tels que listes et tableau -- et à l'intérieur de ceux-ci l'italique, les sauts de ligne, etc -- qui vont être mis en oeuvre pour afficher son contenu ;
> mais encore, et surtout, la façon dont les parties constitutives du contenu (en l'occurence les livres et à l'intérieur de ceux-ci les auteurs, les titres etc.) vont être *distribuées* à l'intérieur de ces objets génériques -- dans quel ordre, selon quel classement, etc.

Dans le premier cas on parlera d'objets de formattage ou *formatting-objects*. Et nous constatons qu'HTML (ou plutôt le couple HTML + CSS) nous fournit d'ores et déjà de tels objets de formattage, à peu près suffisants tout au moins pour l'affichage sur écran.

Dans le second cas on parlera de *transformation*.

Cette analogie boiteuse qui avait été faite au départ avec les feuilles de styles *stricto sensu* explique que dans les premiers projets de spécification du W3C le langage de transformation propre à XML que nous appelons aujourd'hui XSLT a pu être mélangé, dans un projet des spécification unique baptisé à l'époque XSL (*Extensible Style Language*), à un tout autre langage. Cet autre langage étant, lui, destiné à définir des objets de formattage plus riches que ceux de HTML puisque destinés à de présenter un contenu XML sur les supports les plus variés (écran, mais aussi papier...)

Désormais les choses sont beaucoup plus claires, puisque les deux langages ont été séparés. L'un est devenu XSL Transformations (XSLT) et l'autre XSL-Formatting Objects (XSL-FO). Le premier seul est à l'heure actuelle arrivé au stade de spécification du W3C.

Les avantages de XSLT: Ils sont énormes. Par la grâce de *XML* les fichiers de données d'une part, et les documents d'autre part, deviennent une seule et même chose. Par la magie de *XSLT* les uns et les autres peuvent être manipulés à volonté de façon automatique, et ce grâce à un langage certes complexe mais néanmoins accessible au non programmeur, puisque seulement *déclaratif*. Ce qui veut dire que nous en avons désormais fini avec les tâches répétitives effectuées manuellement sur nos documents ! Et que tout fichier "hérité", quel que soit son format d'origine -- sous réserve qu'il puisse être d'abord transformé au format HTML (la transformation en XML "clone" n'étant ensuite qu'une formalité), ou au format "comma separated values" va pouvoir :

1. être transformé en XML "propre" (c'est-à-dire reflétant la structure *intrinsèque* de l'information qu'il contient et non plus une présentation plus ou moins arbitraire de cette information)
2. être secondairement, selon les besoins, transformé en fichiers affichables sur quelque support que ce soit (papier, microordinateur, téléphone portable...)

Comme il vient d'être dit, XSLT est en lui-même un langage très puissant et accessible au non programmeur. Mais, dans la mesure où ce langage comporte encore quelques déficiences, ou ne traite pas certains cas particuliers, les programmeurs pourront continuer à se faire plaisir en profitant des extensions propriétaires proposées par les moteurs de transformation du marché qui en élargissent encore les possibilités -- en particulier en y ajoutant des possibilités de scriptage...

Les caractéristiques de XSLT

Les deux caractéristiques principales de XSLT sont les suivantes :

c'est un langage *déclaratif* et non *procédural*. Ce qui revient à dire qu'à la différence d'un langage de programmation classique, il ne spécifie pas le *comment ?* (les algorithmes) : il se contente de déclarer le *quoi ?* Par exemple :

- que tout ou partie des balises <para> présentes dans le XML source sont à remplacer dans le HTML cible par des balises <p>
- que telle partie de l'arbre XML source doit être reproduite telle quelle dans l'arbre XML résultat, ou bien déplacée, ou bien encore dupliquée...

il est lui-même écrit en XML. Ce qui veut dire qu'il pourra être à son tour transformé par une nouvelle feuille de style XSLT, et ainsi de suite, à l'infini ! Ou bien encore qu'il pourra être manipulé à l'aide de tout langage de programmation qu'on voudra, pourvu que ce langage implémente l'interface *Document Object Model* (DOM)...

A côté de sa syntaxe propre, XSLT fait aussi appel à un second langage, déclaratif lui aussi : XPath. XPath sert à spécifier des*chemins de localisation* à l'intérieur d'un arbre d'éléments XML (ainsi que des *expressions* booléennes, numériques ou "chaîne de caractères" construites à partir de ces chemins), et fait l'objet d'une spécification distincte du W3C.

L'espace de nom (*namespace*) de XSL(T)

XSLT constitue un bel exemple d'utilisation de la philosophie des "espaces de nom" XML (*XML namespaces*). Toute feuille de style XSL(T) débute en effet (après la processing instruction xml) par une déclaration de l'espace de nom xsl :

```
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
version="1.0">
```

ou `<xsl:transform xmlns:xsl="http://www.w3.org/1999/XSL/Transform" version="1.0">` (synonyme)

Deux avantages :

> Le préfixe xsl: va permettre de différentier à l'intérieur de la feuille de style les éléments qui appartiennent au langage XSLT de ceux qui devront apparaître dans le résultat final (les "éléments de résultat litéral")
> L'URI spécifiée dans la déclaration de l'espace de nom va éventuellement permettre de distinguer plusieurs implémentations de XSLT.

Note. Le préfixe xsl: est le préfixe couramment utilisé mais n'est pas obligatoire. L'important est l'URI spécifiée dans sa déclaration. Ce qui suit serait donc valide :

```
<toto:transform xmlns:toto="http://www.w3.org/1999/XSL/Transform"
version="1.0">
....
</toto:transform>
```

La philosophie des "règles modèle" (*templates*)

Qu'est-ce qu'une règle modèle (*template*)?

Comparaison avec les règles CSS

En dépit de ce qui vient d'être dit sur la différence entre CSS et XSLT, il est utile de partir de l'exemple des feuilles de style CSS pour bien comprendre comment fonctionnent les règles modèles XSLT.

On se rappelle qu'une feuille de style CSS se compose d'un certain nombre de **règles** (*rules*). Chacune de ces règles se composant :

1. d'un **sélecteur** (ex. p.commentaire em, qui signifie "les balises contenues dans une balise <p> de classe *commentaire*")
2. d'une ensemble de **déclarations** du type **propriété** (ex. color) - **valeur** (ex. yellow)

```
p.commentaire em {

    color: yellow

}
```

L'effet d'une règle CSS est que si un élément (une balise) du source HTML se trouve satisfaire à la *condition* exprimée par le sélecteur de cette règle, l'objet de formattage correspondant (paragraphe pour une balise <p>, cellule de table pour une balise <td>, etc.) se verra affecté des caractéristiques spécifiées par les déclarations de la règle (dans notre exemple la couleur jaune).

Mais il est important de comprendre qu'une feuille CSS ne fait que *décorer* un arbre HTML ; elle ne le modifie pas. Une règle CSS ne fait qu'*ajouter* des caractéristiques nouvelles (positionnement, couleurs, polices de caractères, etc.) à un objet de formattage qui aurait été généré de toutes façons, CSS ou pas. En conséquence une régle CSS vide (ne contenant aucune déclaration) sera simplement sans effet : lorsqu'une balise satisfera aux conditions exprimées par son sélecteur, l'objet de formattage associé sera généré exactement comme si la règle n'existait pas.

Autre caractéritistique des feuilles de style CSS : l'ordre d'apparition des règles dans la feuille de style est indifférent.

Enfin, lorsqu'une balise satisfait aux conditions exprimées par les sélecteurs de *plusieurs* régles, l'objet de formattage associé va être affecté de *la somme* des caractéristiques spécifiées par les déclarations listées dans ces différentes règles. En cas de conflit entre des déclarations, un mécanisme de priorité entrera en jeu : la déclaration qui est contenue dans la règle la plus spécifique (ex. p.commentaire em est plus spécifique que em) va l'emporter.

A première vue les règles modèles de XSLT ressemblent fort aux règles CSS :

1. elles comportent un sélecteur (ici un chemin de localisation XPath) qui va définir à quel(s) *noeud(s)* (éléments, attributs...) la règle s'applique ;
2. leur ordre d'appartion dans la feuille de style est indifférent ;
3. il existe un mécanisme de priorités pour régler les conflits entre règles s'appliquant à un même noeud.

```
<xsl:template match="para[@type='commentaire']//important">

    <span class="insister"><xsl:apply-templates /></span>

</xsl:template>
```

Mais il y a toutefois qqs différences :

1. Une régle modèle n'a pas pour fonction d'ajouter des caractéristiques nouvelles à un résultat prédéterminé : le résultat dépend *entièrement* de la règle modèle. Ainsi une régle modèle *vide* ne va générer aucun résultat. Lorsqu'un noeud (élément, attribut...) satisfera aux conditions exprimées par son sélecteur, *aucun* contenu correspondant (ni balise, ni attribut, ni texte) ne sera généré dans le fichier résultat : la règle modèle se comportera alors comme un filtre.
2. Caractéristique très importante : pour produire un résultat une règle modèle doit impérativement avoir été *invoquée* par une autre règle modèle.
3. Enfin, en cas de conflit *une seule* règle modèle s'appliquera (par application également d'un mécanisme de priorités), à l'exclusion de toutes les autres.

Toutes ces différences aboutissent à une première conséquence remarquable : alors qu'une feuille de style CSS vide est sans effet (le résultat est le même que si aucune feuille de style n'était associée), une feuille de style XSLT *véritablement* vide aurait, elle, pour effet de produire un *résultat nul* (un fichier vide).

Note. En réalité il n'existe pas de feuille de style XSLT *véritablement* vide, puisque la spécification XSLT a prévu qu'un certain nombre de règles modèles, dites régles modèles "internes", doivent exister par défaut dans toute feuille de style. Ces règles modèles, explicitées plus bas, vont faire qu'une feuille de style "vide" va néanmoins produire un résultat.

De ce qui précède on retiendra deux choses essentielles :

> La feuille de style XSLT *ressemble* à la feuille de style CSS en ce qu'elle est constituée comme elle d'une suite de déclarations, d'ordre indifférent, dont chacune comporte un *filtre* (sélecteur pour CSS, chemin de localisation XPath pour XSLT) servant à distinguer les constituants du fichier source (éléments HTML pour CSS, noeuds XML pour XSLT) auxquels la déclaration s'applique de ceux auxquels elle ne s'applique pas.
> La feuille de style XSLT *diffère* de la feuille CSS en ce que ses différentes déclarations (règles modèles) doivent *s'appeler* les unes les autres, exactement comme dans un langage de programmation procédural un programme principal peut appeller des sous-programmes, qui eux-

mêmes peuvent appeler des sous-programmes, etc. C'est ce mécanisme qui va permettre à la feuille XSLT de générer un arbre résultat éventuellement très différent de l'arbre source, alors que la feuille CSS ne sait que décorer un arbre source dont elle est bien incapable de modifier la structure.

Note. Deux conséquences importantes à ce mécanisme d'appel des règles modèles entre elles :

Il faut bien une régle modèle "principale", qui soit appelée en premier et qui puisse ensuite donner la main aux autres. Cette règle principale est celle qui est associée à la racine (notée "/" dans le langage XPath) du document XML (nous dirons que cette règle est *positionnée* sur la racine du document). Cette première règle va en général en appeler d'autres qui auront toutes pour caractéristiques d'être elles aussi positionnées sur un noeud, ou un ensemble de noeuds précis du document XSLT. Il faut bien en effet avoir présent à l'esprit qu'à tout moment de l'exécution d'une feuille XSLT le moteur de transformation va se trouver positionné sur *deux* documents/fichiers différents :

1. dans le *programme* que constitue la feuille de style : dans une règle modèle précise, et sur une instruction précise de ladite règle modèle ;
2. dans les *données* que constitue le fichier XML source : sur un noeud (élément, attribut...) précis.

Contrairement à ce qui se passe dans un langage procédural, on a ici un mécanisme "piloté par les données" (*data driven)*. En ce sens que l'appel n'est pas un appel direct à une règle modèle (Note. un tel mécanisme existe en XSLT mais il y joue un rôle secondaire : c'est celui des *règles modèle nommées*), mais passe par deux étapes successives:

Rechercher dans le fichier XML les noeuds satisfaisant une condition précise (`<xsl:apply-templates select="`*`motif XPath`*`"/>`). Par défaut (absence de l'attribut "`select`") on recherche les éléments *fils* du noeud sur lequel on se trouve actuellement positionné.
Note. Le motif XPath consiste essentiellement en un adressage, soit absolu soit relatif au noeud XML sur lequel on se trouve actuellement positionné, assorti éventuellement de conditions (dites *prédicats*).

- Constituer la liste de ces noeuds qui devient la *liste de noeuds courante*, et éventuellement ordonner cette liste selon certains critères.
- **Se positionner** successivement sur chacun des noeuds de cette liste (qui devient alors provisoirement le *noeud courant)*, et pour chacun de ces noeuds :
    - rechercher les règles modèles s'appliquant à ce noeud ;
    - s'il y en a plusieurs déterminer la plus prioritaire ; et enfin
    - invoquer cette règle.

Bien entendu les règles ne font pas que se passer le relai les unes aux autres. De temps en temps il leur faut bien travailler! Un travail qui consiste tout naturellement à insérer du contenu dans le fichier cible :

- à l'emplacement du fichier cible où l'on se trouvait lorsque le modèle a été invoqué
- en exploitant le fichier source -- en adressage absolu ou relatif à partir du noeud courant.

Le prélèvement d'information sur le noeud contextuel sélectionnés pourra être :

- soit direct (`<xsl:value-of...>`)
- soir par appel d'une éventuelle règle modèle susceptible de s'y appliquer (`<xsl:apply-templates.../>`)

La dualité fichier source/fichier cible est ce qu'il y a de plus difficile à comprendre lorque vous mettez au point une feuille XSLT. Le fichier source a l'avantage d'être préexistant donc physiquement visible. Le fichier cible, quant à lui, est en gestation, et vous devez donc l'imaginer mentalement, ce qui n'est pas toujours facile, surtout s'il doit différer beaucoup dans sa structure du fichier source... Une chose à bien comprendre en tous cas est que le fichier source *ne change pas* au cours de la transformation XSLT. Le fichier cible, par contre, se construit progressivement, et vous devez, pour la construction de chaque règle

modèle, imaginer où il en sera de sa construction lorsque la règle modèle sera invoquée. Pour cette raison, il est recommandé de construire les feuilles de style de manière incrémentale, règle modèle après règle modèle, en visualisant à chaque fois le résultat obtenu.

Qu'est ce qu'un noeud XML ?

Vu de XSLT, "noeud" (*node*) est le terme générique qui désigne sept objet différents parmi ceux que l'on rencontre à l'intérieur d'un fichier XML. A savoir

1. la *racine* du document -- l'objet qui englobe l'ensemble du document : à ne pas confondre avec l'élément (balise) de plus haut niveau, qui en est le fils, et qui est parfois appelé *élement racine*. En XPath la racine du document est représentée par le symbole /. Si une feuille XSLT déclare des règles modèles il est obligatoire qu'elle en déclare au moins une ayant pour cible la racine (mais il est possible de construire des feuilles XSLT ne déclarant aucune règle modèle et se comportant comme une unique règle modèle).
2. les *éléments* ou balises -- de loin les noeuds les plus importants. Ce sont les "branches" de l'arbre XML. En XPath ils sont symbolisés individuellement par leur nom, tout simplement, et collectivement par * (si le *spécificateur d'axe* qui précède n'est pas attribute:: ou namespace::)
3. les *noeuds textuels* -- les "feuilles" de l'arbre XML, où réside l'information "de base". Les noeuds textuels sont ce qui reste à l'intérieur d'une balise quand on a retiré les balises filles et leur contenu. Ils sont eux aussi considérés comme des noeuds "fils" de la balise qui les contient. En XPath ils sont symbolisés par text(). Note. Les valeurs d'attribut ne sont pas des noeuds textuels
4. les attributs -- les informations complémentaire inscrites (sous la forme étiquette="valeur") à l'intérieur même des balises. En XPath ils sont symbolisés individuellement par leur nom précédé de attribute:: ou de @, et collectivement par attribute::node() ou attribute::*, ou @*. Ils ne sont pas considérés comme des noeuds fils de l'élément qui les contient.
5. enfin, et pour mémoire : les *instruction de traitement* ("processing instructions"), les *commentaires*, et les *espaces de nom* ("namespaces"). En XPath ils sont symbolisés par processing-instruction('*cible*'), comment() et namespace::*préfixe*

A noter qu'en XPath node() désigne collectivement tous les types de noeuds.

Note. On lit parfois que node() désignerait collectivement tous les types de noeud à l'exception des attributs. C'est inexact. La confusion vient du fait que dans une *étape de localisation* node() non précédé d'un *spécificateur d'axe* est une abréviation de child::node(), et désigne donc tous les noeuds *fils* du noeud contextuel. Or les attributs n'en font pas partie. Pour les désigner collectivement il convient donc d'écrire attribute::node() ou attribute::* ou @*

Comment les règles modèles accèdent aux données XML : les *chemin de localisation* ou *motifs* XPath

Les règles modèles passent leur temps à recherche des noeuds dans le fichier source, on vient de le voir. Pour ce faire elles ont à leur disposition un outil de recherche puissant : le "chemin de localisation" (*location path*), parfois appelé aussi "motif" (*pattern*).

Qu'est-ce que le chemin de localisation/motif ? C'est, si vous voulez, l'objet que vous donnez à flairer à votre limier pour qu'il piste un gibier bien particulier... Ou si vous préférez des lunettes dont vous chaussez votre moteur de transformation et qui le rendent aveugle à tous les noeuds à l'exception de ceux, bien spécifiques, qui devront être sa proie unique. Le chemin de localisation s'écrit à l'aide du langage XPath déjà mentionné.

La syntaxe des chemins de localisation : les "étapes de localisation"

Un *chemin de localisation* XPath est constitué de :

une ou plusieurs *étapes de localisation* XPath, séparées entre elles par le symbole /. A son tour chacune de ces étape de localisation est constituée de, dans l'ordre :

1. un *axe* (par défaut l'axe implicite est child), suivi du séparateur :: , suivi de
2. un *test de noeud*, suivi de
3. de zero à *n prédicats*, encadrés chacun par [ ].

Un chemin de localisation XPath, lorsqu'il est écrit en syntaxe abrégée, ressemble beaucoup au chemin qui nous sert à spécifier l'emplacement d'un fichier sur le disque dur de notre ordinateur. Ils désigne une position (relative ou absolue) à l'intérieur d'une hiérarchie, en l'occurence une hiérarchie de noeuds, en faisant appel éventuellement à des "jokers".

Exemple de chemin de localisation absolu : //biblio/*/livre/auteur va rendre le moteur de transformation aveugle à tout ce qui n'est pas une balise <auteur>, fille d'une balise <livre>, elle-même petite fille d'une balise <biblio>", laquelle pourra se trouver n'importe où à l'intérieur de l'arbre XML.
Exemple de chemin relatif : ../adresse va limiter les recherches aux balises <adresse> qui sont filles de la balise mère de la balise "contextuelle" (celle sur laquelle on est actuellement positionné) -- autrement dit les soeurs de la balise contextuelle qui seraient des balises <adresse>, et la balise contextuelle elle-même au cas où elle serait une balise <adresse>.

Symboles et abréviations utilisés dans la syntaxe XPath des chemins de localisation

Symbole	Valeur
/	Séparateur d'étapes de localisation. Si ce symbole est en tête d'un chemin le chemin part de la racine du document (adressage absolu). Si ce symbole n'est pas suivi d'un spécificateur d'axe, il est équivalent à un séparateur parent-enfant (l'axe child étant alors implicite)
//	"Joker vertical". Abréviation de /descendant-or-self::node()/. Si ce symbole est en tête d'un chemin le chemin part de la racine du document (adressage absolu)
.	Désigne le noeud contextuel. Abréviation de self::node().En tête d'un chemin ./ est facultatif.
::	Sépare un spécificateur d'axe d'un test de noeud
..	Désigne l'élement parent du noeud contextuel. Abréviation de parent::node()
*	"Joker horizontal". Peut désigner l'ensemble d'une fratrie d'éléments (*) ou d'attributs (@* ou attribute::*), ou l'ensemble des espaces de nom en vigueur (namespace::*).
@	Préfixe des attributs. Abréviation de attribute::.
:	Séparateur de préfixe d'espace de nom (dans les noms d'élément ou d'attribut).
( )	regroupe des opérations qu'il rend prioritaires
[ ]	Encadre un prédicat
[ ]	Encadre un indice à l'intérieur d'une collection (= cas particulier de prédicat)
|	Opérateur booléen. Entre deux étapes de localisation, spécifie l'une **ou** l'autre de ces étapes de localisation

Les spécificateurs d'axe

Axe	Valeur	Abréviation correspondante
ancestor	les ancêtres du noeud contextuel	aucune
ancestor-or-self	idem, plus le noeud contextuel	aucune
attribute	les attributs du noeud contextuel	@ (équivaut strictement à attribute::)
child	les enfants du noeud contextuel	rien (axe par défaut)
descendant	les descendants du noeud contextuel	aucune
descendant-or-self	idem, plus le noeud contextuel	// (équivaut en fait à /descendant-or-self::node()/)
following	les éléments qui suivent le noeud contextuel (dans l'ordre du document)	aucune
following-sibling	idem, limité à la même fratrie	aucune
namespace	les noeuds "espace de nom" du noeud contextuel	aucune
parent	le parent du noeud contextuel	.. (équivaut en fait à parent::node())
preceding	les éléments qui précèdent le noeud contextuel	aucune
preceding-sibling	idem, limité à la même fratrie	aucune
self	le noeud contextuel	. (équivaut en fait à self::node())

Les tests de noeud

Le test de noeud sert à spécifier un noeud, ou un ensemble de noeuds dans la collection de noeuds désignée par l'axe (explicite ou implicite) qui le précède. Le test de noeud peut être un nom (d'élément, d'attribut, ou d'espace de nom) ou un joker :

Joker	Valeur
*****	tous les noeuds du *type de noeud principal* de l'axe (explicite ou implicite) qui précède. Autrement dit tous les noeuds de type : *attribut* pour l'axe attribute ; *espace de nom* pour l'axe namespace ; *élément* pour tous les autres axes
node()	tous les noeuds quel que soit leur type
text()	tous les noeuds textuels
comment()	tous les noeuds de type commentaire
processing-instruction()	tous les noeuds de type instruction de traitement

Les filtres ou prédicats

Les prédicats vont ajouter des restrictions supplémentaires aux chemins. Ils vont rendre les motifs encore plus sélectifs. Ainsi //biblio/*/livre[@sujet="xml" and datepub="2000"]/auteur va restreindre la recherche précédente aux auteurs de livres dont le sujet (désigné par un attribut de la balise livre) est XML et l'année de publication (désignée par une balise fille du nom de "datepub") est l'an de grâce 2000.

Quelle est la syntaxe d'une règle modèle ?

Elle est la suivante :

```
<xsl:template match="motif"> <!-- Spécifie à quels noeuds la règle est
applicable -->

    <!-- Insertions dans le fichier cible de données prélevées/calculées à
    partir des données du fichier source. Appel éventuel d'autres règles
    modèles -->

</xsl:template>
```

Que fait une règle modèle une fois qu'elle est activée ?

Par défaut elle ne fait rien -- ce qui revient à dire que ce sur quoi elle est positionnée (et qui peut être un sous-arbre) sera absent (ne sera pas reproduit) dans le fichier cible. Ainsi, si vous avez écrit une règle modèle qui capture toutes les balises <para> et qui est vide, les balises `<para>` et tous leurs contenus (qui sont peut-être des sous-arbres) ne seront pas exploités/reproduits dans le fichier cible. (Si l'on accepte l'idée que celui-ci est une image, plus ou moins déformée, du fichier source, on peut dire que les balises `<para>` et tout leur contenu auront été effacés.)

Si vous souhaitez que le modèle fasse quelque chose, ne serait-ce que rendre la main à d'autres modèles, eh bien il faut le lui demander explicitement !

Vous pouvez, par exemple lui demander de simplement remplacer les balises `<para>` et tout ce qu'elles contiennent par le texte "`trouvé` !". En ce cas il vous suffira d'écrire :

```
<xsl:template match="//para">

    trouvé !

</xsl:template>
```

Si vous désirez que ce texte "`trouvé` !" apparaisse à l'intérieur d'une nouvelle balise, par exemple `<p>`, alors vous écrirez :

```
<xsl:template match="//para">

    <p>trouvé !</p>

</xsl:template>
```

Si vous voulez que le texte qui se trouvait précédemment à l'intérieur de la balise `<para>` apparaisse maintenant à l'intérieur de la balise `<p>`, vous écrirez :

```
<xsl:template match="//para">

    <p><xsl:value-of select="."></p>
```

```
</xsl:template>
```

Mais ceci ne va reproduire que les noeuds *textuels* descendants de la balise <para>. Que se passera-t-il si la balise <para> contient des balises "descendantes" -- par exemple une balise <important> ? Eh bien, ces balises descendantes seront ignorées et seul leur contenu textuel appararaîtra dans le résultat final.

Alors, comment faire ? C'est ici qu'apparaît tout l'intérêt de l'instruction "reine" de XSLT : <xsl:apply-templates />.

Que fait cette instruction ? Eh bien, comme on l'a vu, elle permet à la règle modèle active de demander à la cantonade si, par hasard, d'autres règles modèles ne voudraient pas reprendre le travail là où elle l'a laissé... Et si il ne s'en trouve pas ? Eh bien tant pis, elle s'en désintéresse !

Prenez par exemple l'instruction que nous avons écrite ci-dessus dans notre modèle :

```
<p><xsl:value-of select="."></p>
```

Eh bien dans la philosophie coopérative qui est celle de XSLT, elle est inutile. On peut la remplacer par <xsl:apply-templates /> et créer par ailleurs une règle modèle générique unique chargée de la reproduction des noeuds textuels de tout un fichier XML. Cette règle modèle générique "reproductrice de feuilles" (qui fait fort opportunément partie des règles modèle internes de XSLT) aura l'allure suivante :

```
<xsl:template match="text()">

    <xsl:value-of select=".">

</xsl:template>
```

L'avantage de cette façon de travailler est que le <xsl:apply-templates /> en question va pouvoir faire face à tous les cas possibles :

- le cas où la balise <para> ne contient que du texte
- le cas ou elle contient des balises filles -- auquel cas il faudra bien entendu que des règles modèles spécifiques aient été définies pour traiter ces balises, du genre :

```
<xsl:template match="important">

    <em><xsl:apply-templates /></em>

</xsl:template>
```

Note. Les règles modèles capturant des balises petites-filles ne prendront pas la main. A moins que :

- des règles modèles capturant spécifiquement les balises filles existent et leur passent la main en faisant à leur tour <xsl:apply-templates />
- ou bien que l'on ait défini une règle modèle générique qui, faute de de modèles plus spécifique, va capturer les balises filles et leur faire passer la main à leurs propres filles et ainsi de suite. Cette règle modèle générique "qui parcourt l'arbre" à l'allure suivante :

```
<xsl:template match="*">

    <xsl:apply-templates />
```

```
    </xsl:template>
```

Une telle règle modèle est vraiment utile. Si elle n'existe pas et que vous voulez effectuer un traitement pour un noeud donné, disons changer le nom des balises `<toto>` en `<tata>`, il va vous falloir :

- soit avoir des règles modèles qui ciblent tous les noeuds intermédiaires entre la racine du document source et ce noeud particulier, même si vous n'avez aucun traitement particulier à effectuer sur ces noeuds, auquel cas ces règles modèles ne contiendront que le fameux `<xsl:apply-templates />`
- soit effectuer un saut direct à partir d'une autre règlemodèle, en utilisant l'instruction `<xsl:apply-templates select="`*chemin*`" />`

Les règles modèles "internes"

Cette règle modèle générique "parcoureuse d'arbre" est tellement utile que la spécification XSLT a décidé de vous dispenser d'avoir à l'écrire en spécifiant que les moteurs de tranformation XSLT devront la considérer comme existant toujours par défaut, sous la forme encore plus générique suivante :

```
<xsl:template match="/ | *">

    <xsl:apply-templates />

</xsl:template>
```

Cette règle modèle spécifie que, à defaut de spécification contraire, l'arbre XML source doit être parcouru dans son intégralité, à partir de sa racine / et en traversant, dans l'ordre hiérarchique, tous les éléments `(*)` -- et non pas les attributs `(@*)`

La spécification XSLT demande encore que deux autres règles modèles génériques soient considérée elles aussi comme existant toujours par défaut : il s'agit d'une part d'une forme encore plus générique (puisqu'elle reproduit aussi bien les valeurs d'attributs que les noeuds textuels) de la règle modèle "reproductrice de feuilles" évoquée plus haut :

```
<xsl:template match="text() | @*">

    <xsl:value-of select=".">

</xsl:template>
```

Et d'autre part d'une règle modèle assurant la non reproduction des instructions de traitement et des commentaires :

```
<xsl:template match="processing-instruction() | comment()" />
```

A elles trois ces règles modèles par défaut, dites "règles modèle internes", assurent que une feuile de style XSLT *vide* va reproduire tout le *contenu* du fichier XML source, dans l'ordre. Autrement dit elle va dépouiller le fichier de tout ce qui est purement XML (balises -- y compris les attributs et leurs valeurs, processing instructions, commentaires).

Note 1. Les valeurs d'*attribut* ne seront pas reproduites parce que la *première* règle ne fait pas parcourir les attributs. Il y faudrait un `<xsl:apply-templates select="@* | node()" />`. Puisque :

1. `<xsl:apply-templates />` est équivalent à `<xsl:apply-templates select="node()" />` ;
2. node() est en fait l'abréviation de child::node() ;

3. les attributs ne sont pas considérés comme "enfants" de l'élément qui les comporte.

Note 2. Les contenus des noeuds textuels seront, eux, bien reproduits, puisque :

1. comme rappelé ci-dessus, `<xsl:apply-templates />` est équivalent à `<xsl:apply-templates select="child::node()" />`;
2. les noeuds textuels sont considérés comme "enfants" de l'élément qui les comporte.

Ce ne serait pas le cas si dans la *première* règle, on avait un <xsl:apply-templates select="*" />. En ce cas la main ne serait pas donnée aux noeuds textuels ; la *deuxième* règle ne pourrait donc pas leur être appliquée ; donc leur contenu ne serait pas reproduit.

La règle modèle de transformation "à l'identique"

En revanche une règle modèle particulière, dite règle modèle de transformation "à l'identique" (voir aussi http://www.cafeconleche.org/books/bible2/chapters/ch17.html), va permettre de reproduire l'intégralité du fichier XML source, y compris les balises et leurs attributs, les processing intructions, les commentaires, etc. Cette régle modèle est la suivante :

```
<xsl:template match="@* | node()">

        <xsl:copy>
        <xsl:apply-templates select="@* | node()" />
        </xsl:copy>

</xsl:template>
```

Cette règle modèle garantit que tout les noeuds de l'arbre XML source sont parcourus (y compris les attributs), que leurs étiquettes sont copiées (par la grâce de l'instruction `<xsl:copy>`) et leurs valeurs reproduites également (par l'instruction `<xsl:copy>` également lorsque la règle modèle arrive au niveau des feuilles)

Question. Pourquoi "`@* | node()`" et non pas seulement `select="node()"` ? Réponse ci-dessus.

Noeud courant, liste de noeuds courante, et noeud contextuel

Lorsqu'une règle modèle "appelante" passe le relai "à la cantonnade" en spécifiant un chemin de localisation XPath (`<xsl:apply-templates select="..."/>`), l'ensemble des noeuds qui satisfont à l'ensemble des conditions spécifiées par ce chemin de localisation va devenir pour le moteur de transformation XSLT la *liste de noeuds courante*. Le moteur va devoir s'intéresser successivement à chacun des noeuds de cette liste qui deviendra alors provisoirement le noeud courant (caractérisé par une certaine *position* à l'intérieur de la liste de noeuds courante), et pour chacun tâcher de trouver une règle modèle qui lui soit applicable, et passer le relai à cette règle. Ce noeud courant va constituer ensuite le premier contexte de travail de la règle modèle "appelée". Autrement dit, toutes les instructions de travail qui figurent à l'intérieur de cette règle modèle vont utiliser l'emplacement du noeud courant comme base d'adressage à chaque fois qu'elles feront appel à une adresse relative dans le fichier source.

Rappelons que ces instructions "de travail" peuvent être de deux sortes :

> Instructions d'*exploitation directe* du fichier cible. Ainsi, dans l'exemple précédent, l'instruction `<xsl:value-of select=".">` va reproduire les noeuds textuels *descendants* de la balise <para> trouvée par le modèle. Ces instructions ne modifient en général pas la liste de

noeuds courante et le noeud courant. Une exception est l'instruction de répétition `<xsl:for-each...>`. Cette instruction va avoir pour effet de créer provisoirement une nouvelle liste de noeuds courante, dont les éléments vont devenir, chacun à leur tour, le nouveau noeud courant. A la fin de ce processus la liste de noeud courante et le noeud courant vont redevenir ce qu'ils étaient au début.
Instruction d'invocation d'autres règles modèles. De même, si dans l'exemple précédent, l'instruction `<xsl:value-of select=".">` est remplacée par l'instruction `<xsl:apply-templates />`, celle-ci va passer la main à d'autres règles modèles à la condition que celles-ci capturent (s'appliquent à) des noeuds *fils* de la balise `<para>` trouvée par le modèle. Chaque règle modèle ainsi invoquée va bien entendu, à son tour, constituer sa propre liste de noeuds courants et itérer à l'intérieur de celle-ci.

Dans l'un et l'autre cas, l'utilisation de chemins de localisation XPath peut amener à considérer provisoirement, dans les différentes étapes de localisation, des noeuds distincts du noeud courant, et éventuellement à les comparer à celui-ci. Par exemple, supposons que nous soyons dans la régle modèle qui traite chacun des noeuds *livre* d'une bibliothèque et que nous voulions, pour chaque livre, établir une liste des autres livres du même auteur. Nous écririons une instruction du type de :

```
<xsl:apply-templates select="//biblio/livre[auteur/nom =
current()/auteur/nom]/titre[. != current()/titre]" mode="liste" />
```

Dans le chemin de localisation XPath utilisé par cette instruction, nous considérons tour à tour tout une kyrielle de noeuds sur lesquels nous pratiquons éventuellement des tests. Pour distinguer du noeud courant chacun de ces noeuds au moment où nous nous intéressons à lui, nous l'appelons le *noeud contextuel*. Dans l'exemple qui précède, la fonction `current()` nous permet de garder la mémoire du noeud courant et de comparer le titre et le nom de l'auteur du noeud *contextuel* avec le titre et le nom de l'auteur du noeud *courant*.