Avantages et inconvénients du XML, les différentes manières de parcourir un fichier XML, Modules python dédiés au XML

Introduction

- Structurer l'information dans des fichiers textes
- Définir une syntaxe de codage de documents qui soit facilement lisible aussi bien par les humains que par les machines
- Faciliter le traitement informatique tout en conservant un support texte lisible et éditable sans outil particulier

```
<BIBLIOTHEQUE>

        <ROMAN>
                <TITRE>Imajica</TITRE>
                <AUTEUR>Clive Barker</AUTEUR>
                <PRIX>6</PRIX>
        </ROMAN>

        <ROMAN>
                <TITRE>Dune</TITRE>
                <AUTEUR>Frank Herbert</AUTEUR>
                <PRIX>7</PRIX>
        </ROMAN>

        <MAGAZINE>
                <TITRE>Science et Vie</TITRE>
                <DATEPARUTION>2005-02-01</DATEPARUTION>
        </MAGAZINE>

        <ROMAN>
                <TITRE>Christine</TITRE>
                <AUTEUR>Stephen King</AUTEUR>
                <PRIX>5</PRIX>
        </ROMAN>

</BIBLIOTHEQUE>
```

Avantages et inconvénients de XML

Structurer l'information sous une forme plus robuste

- Naturellement structuré, facile à lire et à comprendre
- Pas de besoin d'un logiciel d'édition de code pour l'écrire
- Universel
- Extensible

```
<annuaire>
 <personne class = "etudiant">
   <nom>Desjardins</nom>
   <prenom>Jean-Philippe</prenom>
   <telephone>(819) 234 2343</telephone>
   <email>webmaster@usherbrooke.ca</email>
<!-- insertion de commentaires XML -->
 </personne>
 <personne>
   ...
 </personne>
</annuaire>
```

AVANTAGES	
Fichier texte	C'est un fichier texte, donc il sera toujours lisible dans des décennies. On garantit ainsi une meilleure pérennité de l'information
Le XML est standard	Cela signifie qu'il existe de nombreux outils informatiques qui permettent de lire ou d'écrire du XML. On trouve des librairies C, C++, java, PHP, perl, ... De plus en plus d'outils sont capables de lire des fichiers XML (Internet Explorer, Excel, Mathematica,...)
Le XML est strict	On ne peut pas écrire le XML n'importe comment; Vous êtes obligés de suivre une certaine syntaxe. Ca permet de garantir que le fichier soit toujours lisible. Pour vérifier la syntaxe d'un fichier XML, vous pouvez l'ouvrir dans Internet Explorer. Si le fichier est incorrect, Internet Explorer indiquera l'endroit de l'erreur.
Le XML est stucturé et hiérarchique	Le fichier contient des <BALISES> qui peuvent contenir d'autres balises et ainsi de suite (hiérarchie). L'ordre d'apparition des balises est conservé.
On peut ajouter des commentaires	Les commentaires sont des éléments prévus par la spécification. On peut en rajouter dans le fichier sans casser la structure. Ceci permet de commenter des fichiers afin de garantir une meilleure pérennité de l'information
INCONVENIENTS	
Le XML est verbeux	C'est vrai. Les fichiers XML sont plus gros que des fichiers binaires ou tabulaires. Mais on peut facilement les compresser pour le stockage (avec des outils OpenSource par exemple). On assure ainsi aussi une pérenité de l'information à très long terme
Le tabulaire est mieux compris par Excel	C'est vrai, mais les choses ne peuvent pas forcément toujours être décrites par des tableaux 2-dimensions. De plus on peut difficilement ajouter des commentaires dans les fichiers tabulaires

Le XML est verbeux

1. Utilise beaucoup de caractères

Exemple : <NOMBREEXEMPLAIRES>5</NOMBREEXEMPLAIRES>

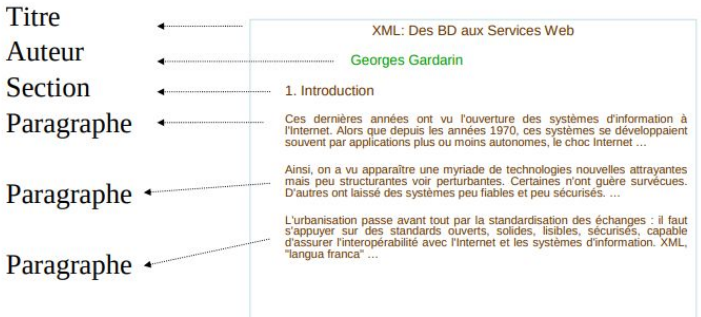
L'information elle-même n'occupe qu'un octet ("5"), mais on l'a encadré de 39 octets

- A terme, ça peut devenir problématique, ou du moins absorber une partie des progrès réalisés en matière de bande passante
- Augmente le trafic réseau

2. L’ordinateur doit tout convertir en binaire
3. Représentation hiérarchique / représentation d'autres types de données

Gestion d'un fichier XML

La syntaxe d'un fichier XML repose sur une chaîne de caractères structurée en deux niveaux, un pour le lecteur humain et un autre pour la machine. Un document XML prend la forme d'un arbre, donc le tronc sert de support à différent types d'éléments appelés "noeuds",comme des textes, attributs, commentaires, éléments et bien d'autres encore.

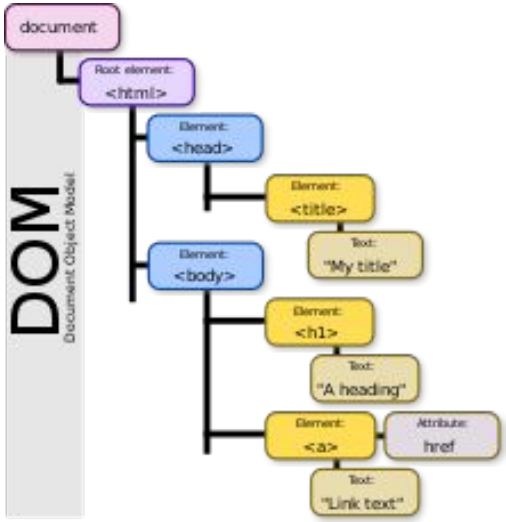


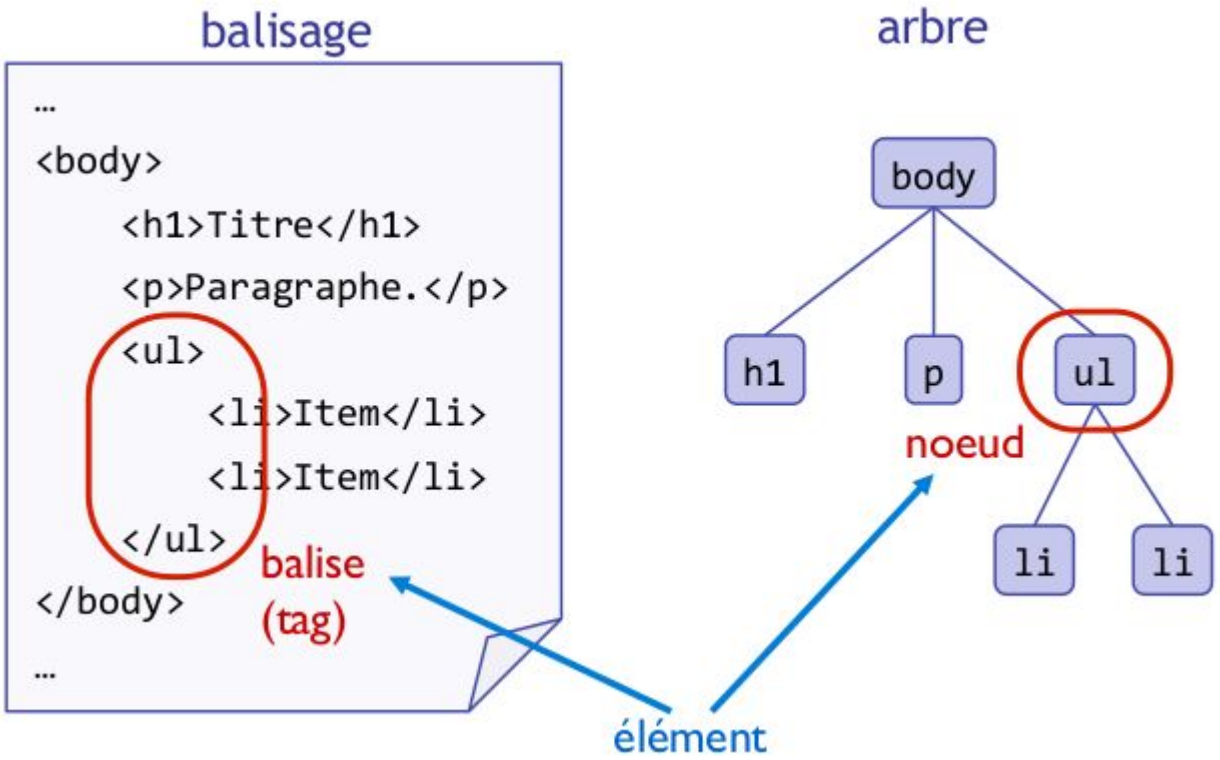

```
<Livre>
<Titre> XML : Des BD aux Services Web </Titre>
<Auteur>Georges Gardarin</Auteur>
<Section titre = "Introduction">
<Paragraphe>Ces dernières années ont vu l'ouverture des systèmes d'information à l'Internet. Alors
    que depuis les années 1970, ces systèmes se développaient souvent par applications plus ou
    moins autonomes, le choc Internet ... </Paragraphe>
<Paragraphe>Ainsi, on a vu apparaître une myriade de technologies nouvelles attrayantes mais peu
    structurantes voir perturbantes. Certaines n'ont guère survécues. D'autres ont laissé des systèmes
    peu fiables et peu sécurisés. ...</Paragraphe>
<Paragraphe>L'urbanisation passe avant tout par la standardisation des échanges : il faut s'appuyer
    sur des standards ouverts, solides, lisibles, sécurisés, capable d'assurer l'interopérabilité avec
    l'Internet et les systèmes d'information. XML, "langua franca" ... </Paragraphe>
</Section>
</Livre>
```

Gestion DOM d'un fichier XML

Le Document Object Model, DOM est une interface de programmation celle ci peut représenter le contenu d'un document XML sous forme d'objets.

Le document XML est représentée sous d'un jeu d'objet telle qu'une phrase ou style relié selon une structure en arbre, avec l'aide de DOM un document peut donc être modifier en ajoutant ou en supprimant des noeuds de l'arbre.



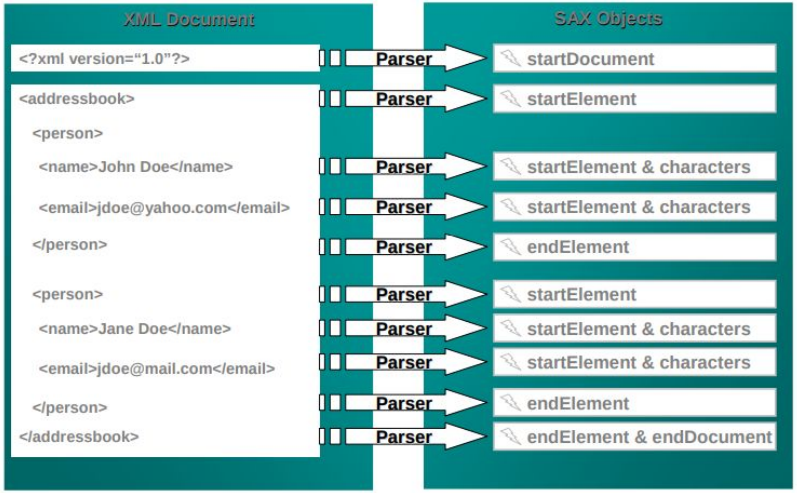

balisage
arbre
…
<body>
<h1>Titre</h1>
Paragraphe.

Item
Item

</body>
…
balise
(tag)
body
h1
p
ul
noeud
li
li
élément

Gestion SAX d'un fichier XML

SAX traite les documents élément par élément au fur à mesure qu'ils sont rencontré, comme pour chaque balise, commentaire ou texte une fonction de post-traitement est appelée. Pour cela que l'interprétation avec SAX utilise moins de mémoire car il n'accumule aucune donnée dans une structure.



Modules Python dédiés au XML

Les interfaces de Python de traitement de XML sont regroupées dans le paquet xml .

Les sous-modules de traitement XML sont :

- xml.etree.ElementTree
- xml.dom
- xml.dom.minidom
- xml.dom.pulldom
- xml.sax
- xml.parsers.expat

Source : docs.python.org

Remarques:

- ➔ les modules dans le paquet xml nécessitent qu'au moins un analyseur compatible SAX soit disponible. L'analyseur Expat est inclus dans Python, ainsi le module xml.parsers.expat est toujours disponible.
- ➔ la documentation des *bindings* des interfaces DOM et SAX se trouve dans xml.dom et xml.sax.

Conclusion

- On peut affirmer que le langage **XML** est la nouvelle base du document numérique, grâce à ses nombreux avantages.
- Il permet de normaliser les échanges de documents grâce à un balisage sémantique indépendant des plates-formes et des langages de programmation.