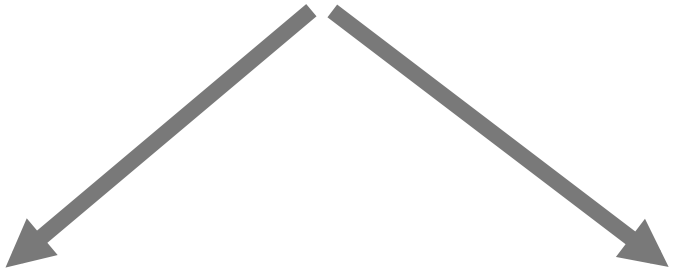
Atelier d’initiation à XML et XML-TEI

Ariane Pinche et Sarah Orsini

Un peu de vocabulaire...

SGML : Standard Generalized Markup Language
- Langage descriptif reposant sur l'usage de balises.



XML : eXtensible Markup Language
- **Contient et structure des données**
- Facilite la conservation et l'échange des données.

HTML : HyperText Markup Language
- **Affiche des données** notamment sur le Web.

Le XML est un Standard

Depuis 1998, XML est un langage libre et documenté. XML est également un *langage standard* respectant les recommandations du *W3C* (World Wide Web Consortium), il facilite :

Pourquoi un standard international ?

- Facilite la lisibilité, par machine ou par l'oeil humain
- Facilite l'échange de données (compatibilité)
- Facilite la migration vers d'autres plates-formes, d'autres logiciels, d'autres formats sans perdre de données
- Langage libre, documenté par ses créateurs et utilisateurs (communauté).

1)Généralités

XML est un format de données pur, très simple et documenté, conçu pour la *description* des documents textuels. XML ne possède pas de jeu de balises prédéfini.

EXEMPLE : Un post-it, version XML.

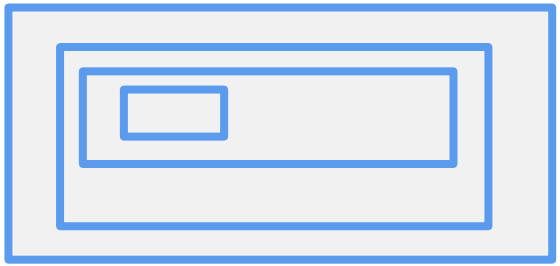
```
<de>Paul</de>
<a>Jacques</a>
<titre>Pense-bête</titre>
<texte>Prendre le dossier sur le
bureau</texte>
```

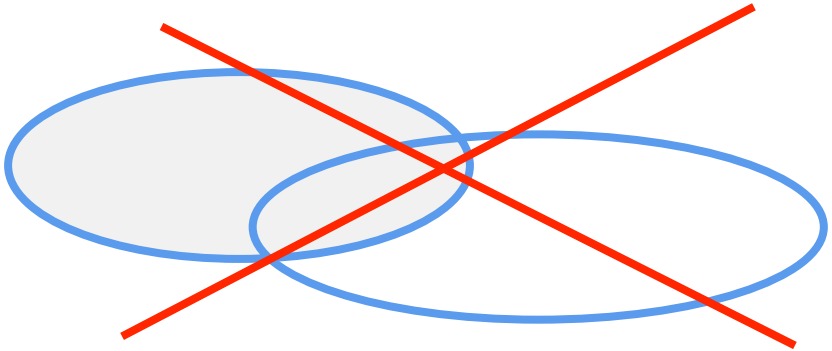
2) Le principe d’imbrication

Un **élément** “tutu” est composé

- D’une **balise ouvrante** <tutu>
- D’une **balise fermante** </tutu>

Les éléments s’imbriquent comme des poupées-russes : les éléments **enfants** héritent des propriétés des éléments **parents**.





Structure générale du XML

Les données sont incluses dans le document XML sous forme de chaînes de caractères délimitées par un balisage les décrivant. L'unité de base qui comprend données et balisage est appelée élément.
Exemple : <nomElement>chaineCaracteres</nomElement>

Un élément XML peut être un élément vide
Exemple : <elementVide/>

Un élément XML peut avoir des attributs
Exemple : <MiseEnValeur rendu=" rouge italique" position=" centrePage">texte</ MiseEnValeur>

Quelques règles importantes

- À chaque balise de début doit correspondre une fin de balise;
- Les éléments peuvent être imbriqués, mais ils ne doivent pas se recouvrir;
- Il ne doit y avoir qu'un seul élément racine;
- Un élément ne doit pas avoir deux attributs avec le même nom.

Un document XML qui suit ces grands principes est dit bien formé.

Exercice

Bien formé ou pas ?

- <paragraphe>du texte</paragraphe>
- <paragraphe><article>du</article><nom>texte</nom></paragraphe>
- <paragraphe><article>du <nom></article> texte</nom></paragraphe>
- <paragraphe type="texte">du texte</paragraphe>
- <paragraphe type=texte>du texte</paragraphe>
- <paragraphe type="texte">du texte<paragraphe/>
- <paragraphe type="texte">du texte<nomPersonnage>nom de personnage</paragraphe>
- <paragraphe type="texte">du texte</Paragraphe>
- du texte</paragraphe>

Exercice

J'ai un texte, je dois signaler l'ensemble texte, les chapitres, les titres de chapitre, les paragraphes contenus dans les chapitres, les noms de personnages trouvés dans le texte, des notes de bas de page. Comment vais-je imbriquer les balises suivantes ?

<texte>
<chapitre>
<titreChapitre>
<paragraphe>
<nomPersonnage>
<noteDeBasDePage >

Blanche-Neige mange la pomme empoisonnée

Il était une fois Blanche-Neige[1] que sa belle-mère détestait.

Blanche-Neige attend le nain Charmant

Un jour Charmant[2] arriva sur son cheval blanc et la sauva.

1 Elle porte ce nom, car son visage est blanc comme la neige
2 Nain qui sauve Blanche-Neige, il est rare que les variantes du conte mentionnent sa petite taille.

3) XML-TEI

TEI = Text Encoding Initiative.

C’est une communauté qui a fixé des standards pour l’édition numérique. La version actuelle date de 2007, on l’appelle TEI-P5.

Pour que notre document soit compréhensible et échangeable au sein de la communauté des éditeurs, il doit être **valide,** c’est-à-dire respecter une **grammaire et un jeu de balises défini par la TEI.**

Cette grammaire s’appelle les **TEI Guidelines,** qui définissent les balises les unes par rapport aux autres.

Exercice 3 :

Encoder le corrigé de l'exercice 2 en remplaçant les balises par celles qui suivent les prescriptions de la TEI.

Jeu de balises :

<text>
<body>
* *
<p>
<head>
<persName>
<note>

Structure d’un document XML-TEI : deux parties principales

1. **Un TeiHeader :**
    - les métadonnées (informations de publication, auteur, localisation dans l’archive, description du document encodé…). => **<fileDesc>**
    - Entrées de glossaire, listes de personnages, des mains qui ont écrit le document => **<profileDesc>**

1. **Une section Text :**
    - Préface **=> <front>**
    - Le texte **=> <text>**
    - Les éléments de structure, d’analyse (apparats, notes) et d’enrichissement (liens vers d’autres textes)
    - Épilogue **=> <back>**

Exemple à ouvrir avec Oxygen : document xml.base

Liens Utiles

Tutorats et exercices :

- https://www.w3schools.com/xml/
- http://teibyexample.org/

Guides :

- https://www.xml.com/axml/axml.html
- http://www.tei-c.org/Guidelines/