Qu'est-ce qu'un texte ?

Matérialité du texte

- Le texte brut (.txt) / *plain text* se compose de caractères pour lesquels ils existent plusieurs répertoires :
  - ASCII (American Standard Code for Information Interchange)
  - ISO 8859-1 (Latin 1)
  - UTF-8 (Universal Character Set Transformation Format - 8 bits)

Pour aller plus loin : https://www.youtube.com/watch?v=MijmeoH9LT4

- Le texte « stylé » / *fancy text*: balises et mise en valeur typographique :
  - balisage LaTeX
  - balisage HTML

La notion « texte »

- Quels peuvent être les différents aspects d'un texte ?
- Comment décrire les différents aspects d'un texte ?
    - Le balisage sémantique permet d'expliciter certains aspects du texte. **XML** (eXtended Markup langage) est très adapté pour ce type d'usage.

Exemple de balisage typographique et sémantique

Langage	Balise typographique	Balise sémantique
LaTeX	\emph{ad hoc}	\selectlanguage{latin}{ad hoc}
HTML5	`<i>ad hoc</i>`	`<i lang="la">ad hoc</i>`
XML-TEI	`<hi rend="i">ad hoc</hi>`	`<foreign xml:lang="la">ad hoc</foreign>`

Pour réviser : https://github.com/architexte/cours-TEI/blob/master/1-text-balises.md

- **Quel est l'intérêt d'un balisage sémantique ?**
    - Exemple d'édition scientifique : *The Shelley-Godwin Archive*, http://shelleygodwinarchive.org/sc/oxford/frankenstein/volume/i/#/p1/mode/rdg
    - Pour aller plus loin : *DH in Practice - Digital Scholarly Editions* par E. Pierazzo, https://www.youtube.com/watch?v=0DB51fblNWI

Introduction à XML

Observer un document XML : *L'année 1437 dans la pratique de Pierre Christofle, notaire du Châtelet d'Orléans*, http://elec.enc.sorbonne.fr/christofle/index.html

Définition

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.

Un standard international

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 :

- la lisibilité par les machines ou par l'œil humain;
- l'échange de données;
- la migration vers d'autres plates-formes, d'autres logiciels, d'autres formats.

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>`

Les éléments XML suivent un principe d'arborescence par imbrication.

Exemple :

```
<elementParent>
  <elementEnfant>chaineCaracteres</elementEnfant>
</elementParent>
```

Ainsi les éléments *enfants* héritent des propriétés des éléments *parents*

Un peu d'Histoire...

- SGML (1970), Standard Generalized Markup Language;
  - HTML, HyperText Markup Language: affiche des données notamment sur le Web;
  - XML, eXtensible Markup Language: contient et structure des données textuelles.

Les éléments XML

Éléments et attributs

Les éléments

`<element>texte</element>` ou `<elementVide/>`

Les éléments doivent tous strictement respecter le principe d'imbrication.

Les attributs

```
<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 encodage qui respecte ces grands principes du XML est dit **bien formé**.

Les commentaires

```
<!-- texteCommentaire -->
```

Les entités

```
&entité;
```

Les entités sont des appels pour insérer dans le XML des caractères interdits ou bien des séquences de code définies au préalable dans une DTD.

Astuce : convertisseur de caractères (pour obtenir le code decimal), http://hapax.qc.ca/conversion.fr.html

Instruction de traitement et déclaration XML

```
<?xml version="1.0" encoding="UTF-8"?>
```

Les instructions de traitement sont un autre moyen de fournir des informations aux applications auxquelles est destiné le document. Une instruction de traitement commence par "< ?" et se termine par "? >".

Ces dernières sont des balises et pas des éléments. Elles doivent donc être en dehors d'une balise.

Les instructions de traitement les plus courantes sont l'appel d'une feuille de style, d'un schéma et l'appel d'une version de XML. Ces appels doivent être placés avant l'élément racine.

Pour réviser :

Why do we encode : E. Pierazzo

https://www.youtube.com/watch?v=R0ncI_rr1z4&list=PL77mHK9JuenN9NXeXQbVcUORz7HZk-9Pv&index=2

Liens de téléchargement :

Oxygen Editor :
https://www.oxygenxml.com/xml_editor/download_oxygenxml_editor.html?os=Linux
XMLmind : https://www.xmlmind.com/xmleditor/download.shtml