



XML TEI

M2 TNAH – 2021/2022



Segolene-Albouy/XML-TEI_M2TNAH

[segolene.albouy\[at\]gmail.com](mailto:segolene.albouy[at]gmail.com)

tinyurl.com/XMLTEI-TNAH1

Déroulé des séances

| | |
|--------------------|---|
| 13 octobre | Encoder le texte : les langages à balises |
| 18 octobre | Contraindre le XML : la <i>Document Type Definition</i> |
| 25 octobre | Décrire les documents patrimoniaux : la TEI |
| 1 novembre | Anatomie d'un document TEI : parties et composants |
| 15 novembre | Les TEI guidelines : modules, modèles, macros, etc. |
| 22 novembre | Localisation dans le document : XPath |

Déroulé des séances

| | |
|--------------------|---|
| 29 novembre | Éditer un texte : initiation à l'édition scientifique |
| 6 décembre | Éditer un texte : description des sources manuscrites |
| 13 décembre | Éditer un texte : encodage d'une édition critique |
| 3 janvier | Les schémas XML : initiation à l'ODD |
| 10 janvier | Les schémas XML : ODD personnalisations avancées |
| 17 janvier | Les schémas XML : Mise en pratique |

Évaluation

Devoir sur table

29 novembre

Vérification des bases de la structure XML et des principes de la TEI

Coefficient 1

Projet individuel

À rendre avant le 24 janvier

Édition en XML TEI d'un texte accompagné d'une ODD documentée

Coefficient 3

Objectifs

**Maîtriser la
structure XML**


**Créer une édition
numérique**

**Décrire un
document textuel**

**Documenter ses
choix d'encodage**

Oxygen

oxygenxml.com/xml_editor



Séance 1

Encoder le texte

Les langages à balises

Le texte informatisé



Texte brut

Le contenu textuel,
dépourvu de toute
structure et mise en
forme



Texte balisé

Le contenu textuel,
structuré à l'aide de
balises



Texte formaté

Le texte mis en forme à
l'aide d'un éditeur de
texte

Les formats textuels

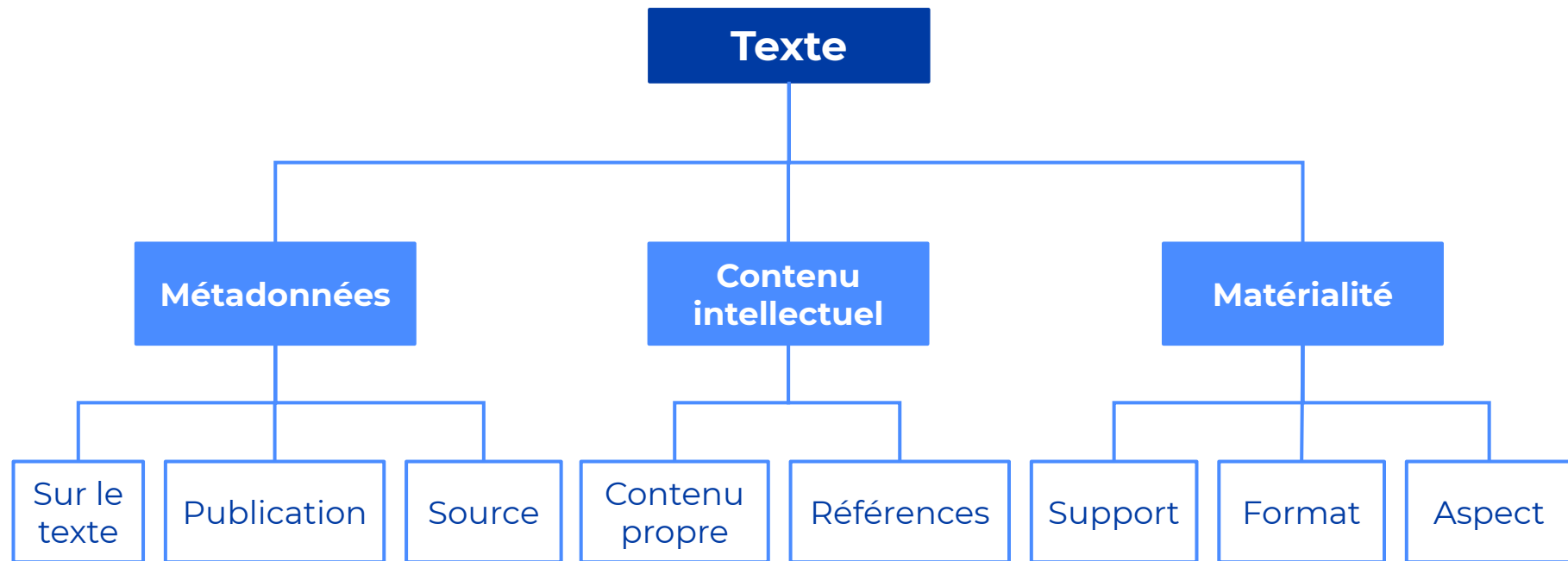


A decorative graphic on the left side of the slide consisting of two blue squares. The top square is a lighter shade of blue and is positioned above the bottom square, which is a darker shade of blue. They are aligned to the left of the main text area.

Qu'est-ce qu'un “texte” ?

Quelles peuvent être ses caractéristiques ?
Comment les retranscrire numériquement ?

Caractéristiques d'un texte



Les langages à balises

```
{poeme}  
  {meta}  
    {type}Haïku{type}  
    {auteur}Takahama Kyoshi{auteur}  
  {meta}  
  {contenu}  
    {vers}Le serpent s'esquiva{vers}  
    {vers}Mais {gras}le regard{gras} qu'il me lança{vers}  
    {vers}Resta dans l'herbe.{vers}  
  {contenu}  
{poeme}
```

Différents type d'encodage

| | Typographique | Sémantique |
|-------|---|--|
| LaTeX | <code>\emph{ad hoc}</code> | <code>\selectlanguage{latin}{ad hoc}</code> |
| HTML | <code><i>ad hoc</i></code> | <code><i lang="la">ad hoc</i></code> |
| TEI | <code><hi rend="i">ad hoc</hi></code> | <code><foreignxml:lang="la"> ad hoc</foreign></code> |

Exercice

Relever les différences d'encodage

`./Exemples_encodage/Mon_reve_familier`

A decorative graphic on the left side of the slide consisting of two blue squares. The top square is a lighter shade of blue and is positioned above the bottom square, which is a darker shade of blue. They are aligned to the left of the main text area.

Caractéristiques du XML

Format de données simple, documenté
Conçu pour la description des documents textuels
Pas de jeu de balise prédéfini

Standard international

Depuis 1998, XML est un langage libre et documenté. C'est un standard respectant les recommandations du W3C, 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.

A decorative graphic on the left side of the slide consisting of two overlapping squares. The front square is a medium blue and is positioned slightly higher and to the right. The back square is a darker blue and is positioned slightly lower and to the left.

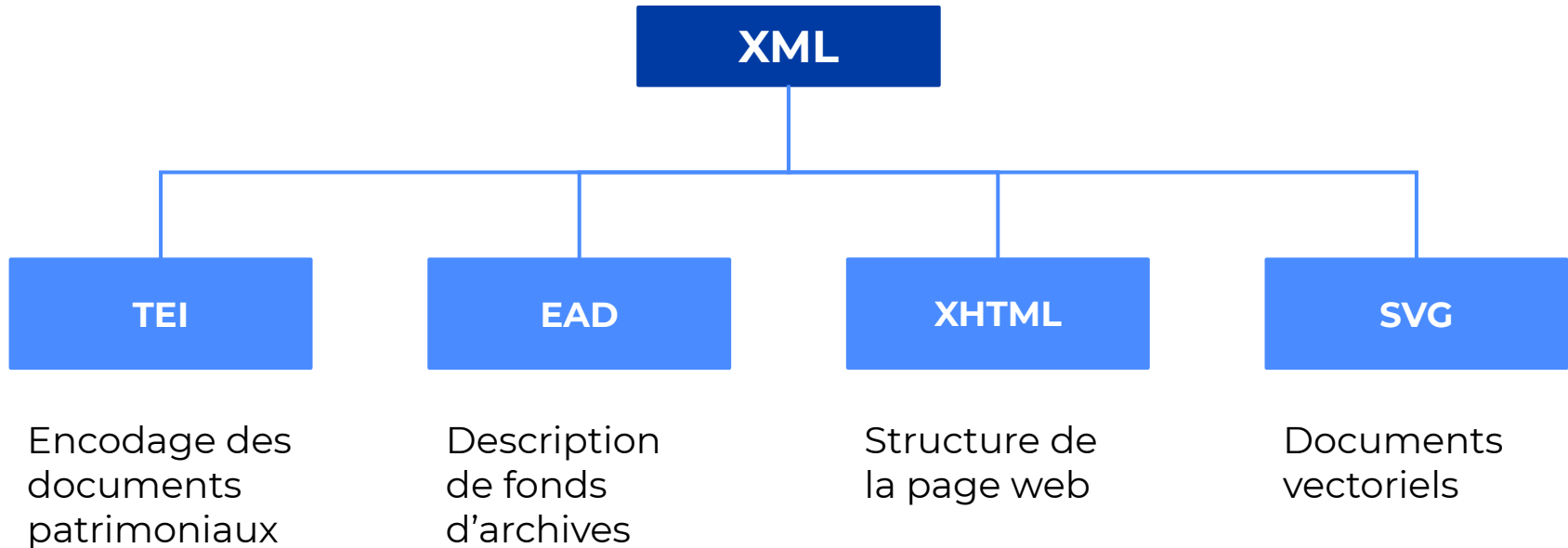
Historique

1970 : SGML, *Standard Generalized Markup Language* ;

1989-1992 : HTML, *HyperText Markup Language* ;

1998 : XML, *eXtensible Markup Language*.

Les langages XML



Principes fondamentaux

Balises

Les données textuelles sont repérées de part et d'autres par des "étiquettes"

Imbrication

Les éléments sont contenues les uns dans les autres et ne se chevauchent pas

Héritage

Les éléments enfants héritent des éléments parents

```
</desc>
<fileDesc>
  <langUsage>
    <language ident="fre">français (langue de l'édition,<0xa0>langue par défaut)</lang>
    <language ident="frm">français moyen (langue principale du document édité)</language>
    <language ident="lat">latin (langue utilisée dans le document édité)</language>
  </langUsage>
</fileDesc>
<revisionDesc>
  <change who="#KF #FC">
    <date when="2009-07">juillet 2009</date>
    <desc>nouvelle conversion en TEI,<0xa0>intégrant notamment un index enrichi,<0xa0>
      corrections manuelles</desc>
  </change>
  <change who="#FC">
    <date when="2010-01">août 2009-janvier 2010</date>
    <desc>enrichissement manuel du document TEI,<0xa0>en particulier pour l'indexation
  </change>
  <change who="#KF #DB #FC">
    <date when="2010-05">mai 2010</date>
    <desc>corrections dans les index,<0xa0>établissement de nouveaux liens et renvois
  </change>
  <change who="#KF #FC">
    <date when="2010-12">décembre 2010</date>
    <desc>nouvelle conversion en TEI à partir d'une édition au traitement de texte incluant
      un index complet de toutes les entités nommées,<0xa0>et intégrant les dernières
      corrections de transcription</desc>
  </change>
  <change who="#KF #FC">
    <date when="2014">2014</date>
    <desc>relecture et corrections de l'édition et des index,<0xa0>intégration de
      l'introduction</desc>
  </change>
  <change who="#FC">
    <date when="2015-04">avril 2015</date>
    <desc>corrections de l'édition,<0xa0>développement du site web définitif<0xa0>

```

Observer un document XML

./Christofle-septembre2016.xml

L'année 1437 dans la pratique de Pierre Christofle, notaire du Châtelet d'Orléans

ACCUEIL | INTRODUCTION | ÉDITION DES NOTES | TYPES D'ACTES | LIEUX ET PERSONNES



Reproduction de Pierre Christofle sur un document
portant le sceau de la prévôté d'Orléans
ADL, H Dépôt2, 1B54, 1^{er} août 1442)

Pierre Christofle fut notaire royal à Orléans de 1423 à 1450, attaché à la prévôté pour qui il rédigeait des contrats portant le sceau de l'institution. Comme ses confrères, Pierre Christofle inscrivait dans un registre et en une forme abrégée les conventions qu'il attestait. Ces notes, au nombre de 387, rédigées entre le 1^{er} janvier et le 29 décembre 1437, conservées aux Archives départementales du Loiret sous la cote 3E 10144, sont ici éditées dans leur ensemble. L'édition, indexée et prochainement téléchargeable, permet l'enquête sur le lexique des actes. Elle accompagne et complète une version imprimée, publiée sous le titre *Promettre, confesser, s'obliger devant Pierre Christofle, notaire royal d'Orléans [1437]*, dans la collection « Mémoires et documents de l'École des chartes », qui propose un essai interprétatif de cette documentation.

La mise en ligne autorise l'accès rapide au contenu par le biais de plusieurs outils de consultation, une vue en parallèle des images numérisées du registre et de leur transcription afin de mieux comprendre (voire d'étudier) la mise en scène de l'écrit notarié, et, bientôt, la possibilité de télécharger l'ensemble des notes.

L'édition électronique intégrale des notes du notaire Pierre Christofle pour l'année 1437 est un projet d'abord mené à l'université d'Ottawa par Kouky Fianu, professeur à l'université d'Ottawa, avec la collaboration d'Anne Fortier, étudiante de maîtrise de l'université d'Ottawa, puis développé dans sa version électronique par Florence Clavaud, conservateur en chef du patrimoine

Observer le résultat d'une édition XML

elec.enc.sorbonne.fr/christofle/index.html

Exercice

**Relever les différents
“composants” du XML**



Les “composants” du XML

Éléments

Le “nom” des balises

Attributs

Les “caractéristiques”
d’une balise

Entités

Appel à un caractère ou
une séquence prédéfinie

Commentaires

Pour extraire des
passages du XML

Les éléments

```
<element>texte</element>
```

```
<elementVide/>
```

≠ Balise

Un élément désigne les balises ainsi que leur contenu

Cloture

Un élément ouvert doit être fermé obligatoirement

Imbrication

Un élément ne peut pas en chevaucher un autre

Le prologue

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

Il indique la version de la recommandation XML qui sert de base à l'écriture du fichier (il n'y a eu à cette date qu'une seule version, la version 1.0), ainsi que l'encodage de caractères utilisé.

Les attributs

```
<miseEnValeurposition="centrePage">  
    texte  
</miseEnValeur>
```

Décrire

Utilisés pour décrire le contenu d'un élément

Situation

Placé les uns à la suite des autres dans la balise ouvrante

Répétabilité

Un même attribut ne peut pas être réutilisé dans la même balise

Les entités

&entité;

Caractère

Est utilisée à la place d'un caractère interdit en XML comme le "&"

Remplacement

Mise à la place d'une séquence de code définie dans une DTD

Conversion

hapax.qc.ca/conversion.fr

Les commentaires

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

Ajout

Destinés à ajouter une note généralement pour préciser le contexte à un lecteur humain

Hors XML

Pour baliser une chaîne de caractère non contrainte par le standard XML

Règles à retenir

**Balise ouvrante =
balise fermante**

1 seul élément racine

**Respect de
l'imbrication**

**1 attribut du même
nom par balise**

= encodage

bien formé

Exercice

**Encoder un court
passage**



Encoder

À l'aide des balises suivantes, structurer le texte proposé :

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

Blanche-Neige mange la pomme empoisonnée

Il était une fois Blanche-Neige¹ que sa belle-mère détestait. [...]

Blanche-Neige attend le prince Charmant

Un jour Charmant² arriva sur son cheval blanc et la sauva. [...]

¹ Elle porte ce nom, car son visage est blanc comme la neige

² Nom du prince qui sauve Blanche-Neige.

Exercice

**Bien formé ou mal
formé ?**



Quelques règles importantes

<paragraphe>du texte</paragraphe>

<paragraphe><article>du</article><nom>texte</nom></paragraphe>

<paragraphe><article>du <nom></article>texte</nom></paragraphe>

<paragraphe><article>du <nom>texte</nom></article></paragraphe>

<paragraphe type="texte">du texte</paragraphe>

<paragraphe type=texte>du texte</paragraphe>

<paragraphe type="texte">du texte<paragraphe/>

<paragraphe type="texte">du texte<nom>nom</paragraphe>

<paragraphe type="texte">du texte</Paragraphe>

<segment type="texte" type="nombre">du texte</paragraphe>