Modéliser et exploiter des corpus textuels

-

Initiation au XML-TEI et à l'analyse outilée de documents XML

-

Atelier organisé par :
Léa Saint-Raymond (ENS-PSL) et Paul Kervegan (INHA)

ENS-PSL, 31 mars 2023

Modéliser et exploiter des corpus textuels (Léa Saint-Raymond, Paul Kervegan, ENS-PSL, 30.03.2023)
Le programme de la journée

Matin : introduction au XML
- Du SGML, au XML, à la TEI : comment en est-on venu.e.s à modéliser du texte ?
- Exercice pratique : l'encodage sans ordinateurs
- Présentation de la correspondance Matsutaka

Après-midi : analyse automatique d'un corpus en XML-TEI
- Enrichissement automatique (via des API, entres autres)
- Reconnaissance d'entités nommées
- Analyse de données : cartographie du corpus...

Le XML en bref, c'est quoi ?

```
<place xml:id="kobe">
    <placeName>Kobe</placeName>
    <location>
        <geo>135.1943764 34.6932379</geo>
    </location>
</place>
```

Le XML en bref, c'est quoi ?

```
<place xml:id="kobe">
    <placeName>Kobe</placeName>
    <location>
        <geo>135.1943764 34.6932379</geo>
    </location>
</place>
```

- Une **syntaxe** pour structurer du texte de façon **hiérarchique**, en éléments/sous-éléments

- Le texte est structuré à l'aide de **balises**

- Un élément XML commence avec une balise **ouvrante** (<a>) et termine avec une balise **fermante** ()

- Il peut aussi y avoir des balises **autofermantes** (<a/>)

Le XML en bref, c'est quoi ?

```
<place xml:id="kobe">
    <placeName>Kobe</placeName>
    <location>
        <geo>135.1943764 34.6932379</geo>
    </location>
</place>
```

- Un élement XML est **défini par son nom** (en rose dans l'exemple : <place>)

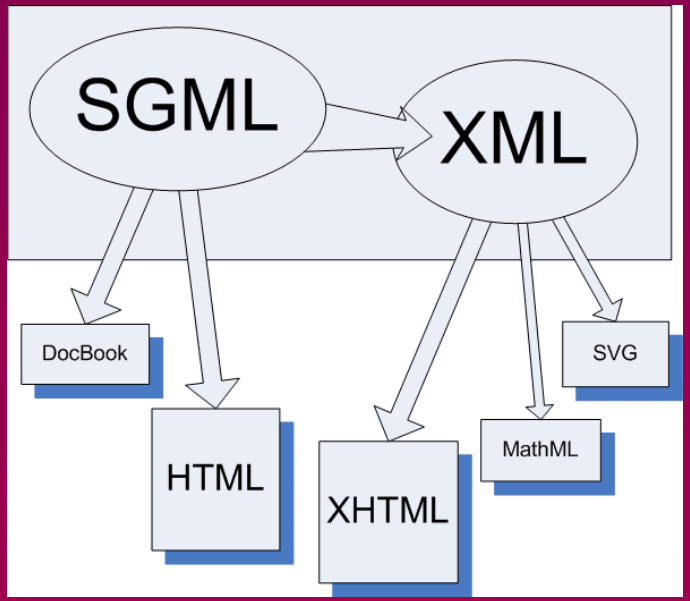
- Un élément XML peut contenir :
  - des **attributs** dans la balise ouvrante (en vert dans l'exemple : xml:id="kobe"). Un attribut sert à qualifier l'élément XML, à contenir des informations supplémentaires
  - du **texte**, entre la balise ouvrante et la balise fermante (en blanc dans l'exemple : Kobe)
  - d'autres **éléments XML** (dans l'exemple, placeName est contenu dans place)

- Deux éléments XML **ne peuvent pas se chevaucher** : c'est impossible de faire ça : <a>

=> à la base, c'est tout !

L'évolution des standards, en quelques dates et acronymes

- 1969 : GML (*Generalized Markup Langage*) créé chez IBM
- 1986 : SGML (*Standard Generalized Markup Langage*) remplace le GML
- 1987 : TEI (*Text Encoding Initiative*), basé sur le SGML
- 1999 : XML (*eXtensible Markup Langage*), standard principal pour l'encodage du texte aujourd'hui



Et la TEI dans tout ça ?

Les *Guidelines*, ou la bible des balises: https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html

Le XML définit une syntaxe, la TEI définit une sémantique :
- le **XML dit comment structurer** un document (organisation hiérarchique d'éléments, avec une balise ouvrante, une balise fermante, du texte et des attributs)
- la **TEI définit quelles balises** utiliser, leur signification, quels attributs utilisés, comment imbriquer les éléments entre eux…
- la TEI définit enfin l'**ODD (*One Document Does it all*)**, un document XML validé par la TEI. L'ODD est un document TEI qui documente et vérifie la structure d'autres documents TEI.
    - elle permet de définir ses propres schémas/modèles d'encodage et de les documenter : quels éléments sont autorisés, quels attributs, quelles combinaisons d'éléments et de valeurs…
    - il est alors possible de créer sa propre spécification de la TEI, et de l'augmenter avec des éléments non standards

Les principes de la TEI sont les suivants :
- « **balisage sémantique** » : c'est le sens, la sémantique du texte qui prime sur la forme
- **modèle communautaire**, lié à une communauté de chercheur.euse.s en /sciences du texte =>
    - **très large domaine d'application** : la TEI doit permettre d'encoder tous les textes (textes de différentes époques, langues, zones d'origines…)
    - vous pouvez **proposer vos propres éléments et améliorations** à ajouter à la TEI :)

Mais comment en est-on arrivé là ?

La création du GML, du SGML et enfin du XML ne viennent pas des « humanités » : ce sont des technologies crées par des informaticien.ne.s pour stocker des données textuelles, et pas pour représenter du texte au sens où nous l'entendons.

Il y a donc une **histoire de l'adoption du XML dans les humanités**, et des moments où des choix conscients ont été faits pour adapter cette technologie à la littérature, à l'histoire ou à l'archivistique.

OHCO : *One ordrered collection of objects*

Cette expression résume la théorie qui se développe au tournant des années 1990 : tout texte peut être représenté comme *une collection ordonnée d'objets*. Tout un programme :

- **One** : un encodage représente une seule organisation d'un texte
- **Ordrered** : un encodage représente le texte de façon ordonnée, structurée
- **Collection** : le texte est séparé en plusieurs petits blocs
- **Objects** : on ne parle plus de texte, mais d'objets => éloignement de la textualité, remplacé par le paradigme informatique « d'objet »

Théorie exprimée entre autres dans : DeRose (Steven), Durand (David), Mylonas (Elli) et Renear (Allen), «What is Text, Really ?», Journal of Computing in Higher Education, I–2 (1990), p. 3-26, url :https://cs.brown.edu/courses/cs195-1/reading/WhatIsTextReally.pdf

Critiques de l'OHCO

Cette théorie est critiquée dès ses débuts, y compris par les auteur.ice.s de l'article *What is text, really ?*

- des **critiques pratiques** : la structure de l'OHCO ne colle pas aux textes :
  - le problème des hiérarchies multiples
  - l'OHCO ne permet pas de représenter toutes les facettes d'un texte (cf. schéma)
- des **critiques théoriques** :
  - produit d'une mentalité positiviste et universaliste
  - l'OHCO influence la perception du texte, impose une signification unique à un texte
  - l'OHCO a été pensé dans un contexte occidental et ne prend pas en compte les formes non-occidentales du texte
  - l'OHCO prend le texte comme un élément isolé (problème de la référentialité)

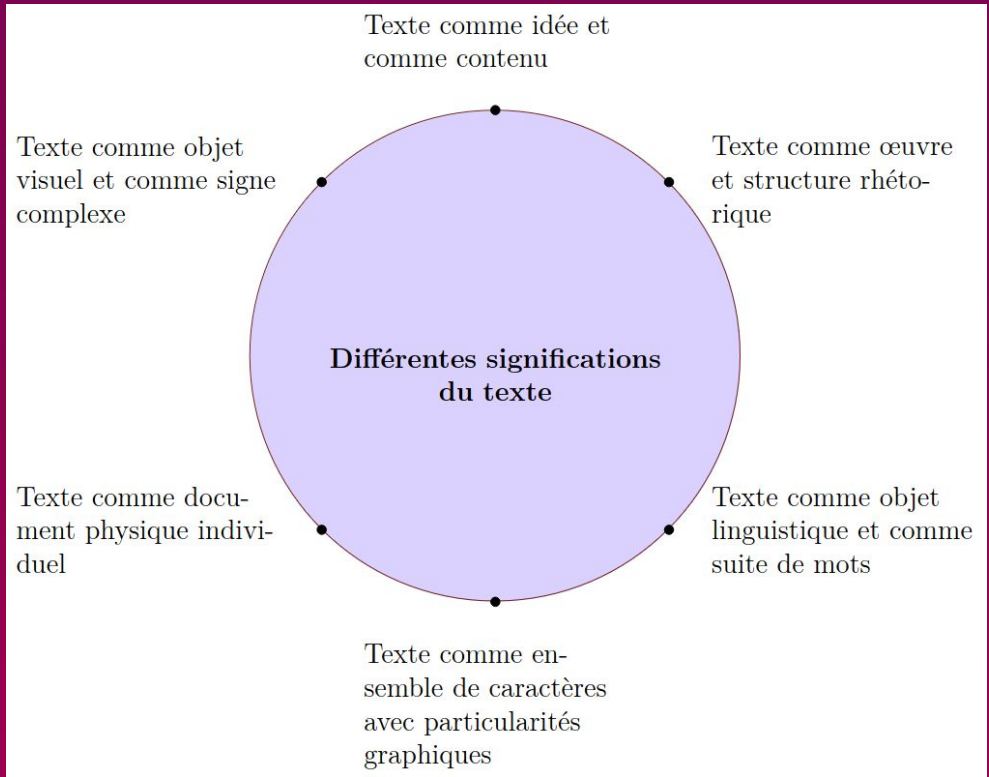


FIGURE 2.6 – Les différentes signification d'un texte selon P. Sahle (2016)

Renear (Allen), Mylonas (Elli) et Durand (David), « Refining our Notion of What Text Really Is : The Problem of Overlapping Hierarchies », dans *Research in Humanities Computing*, dir. Nancy Ide et Susan Hockey, Oxford, 1996

Critiques de l'OHCO

Il y a donc eu des tentatives de **syntaxes alternatives**:
- le texte comme graphe de citations (*Homer MultiText*)
- les syntaxes alternatives (LMNL)

Mais surtout, il est apparu qu'**une modélisation XML est toujours une représentation d'un texte** (pas le texte lui-même). C'est toujours quelque chose de subjectif, une interprétation qui construit la manière dont un texte est perçu et qui détermine ce qu'on pourra faire du document encodé.

L'impact des débats sur OHCO dans la TEI

L'interprétativité de la modélisation est un principe central de la TEI :
- pas de modélisation « parfaite » pour un texte
- possibilité de contraindre l'encodage et de documenter ses choix d'encodage via l'ODD
- développements spécifiques pour différents types de textes, anciens et extra-occidentaux

Chez les utilisateur.ice.s, de nombreux exemples d'**utilisations critiques de la TEI** : par exemple, un encodage croisant différents brouillons d'un chapitre du *Portrait de Dorian Gray* d'Oscar Wilde qui montre comment les sous-entendus homosexuels sont progressivement édulcorés :
- https://gofilipa.github.io/dorian_encoded/
- https://github.com/gofilipa/dorian_encoded

Bibliographie

Très bonne approche refléxive et critique sur la modélisation : le livre de Julia Flanders & Fotis Jannidis
- Jannidis (Fotis) et Flanders (Julia), « A gentle introduction to data modelling », dans *The shape of data in the digital humanities : modeling texts and text-based resources*, dir. Julia Flanders et Fotis Jannidis, London ; New York, 2019 (Digital research in the arts and humanities), p. 26-95.
- — « Data modelling in a digital humanities context », dans The shape of data in the digital humanities : modeling texts and text-based resources, dir. Julia Flanders et Fotis Jannidis, London ; New York, 2019 (Digital research in the arts and humanities), p. 3-25

Sur la TEI (voir surtout les articles de Lou Burnard, l'un des créateur.ices de la TEI!)
- TEI Consortium, P5 : Guidelines for Electronic Text Encoding and Interchange, Text Encoding Initiative, Version 4.4.0, 2022, url : https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html
- Burnard (Lou), « How modelling standards evolve. The case of the TEI », dans *The shape of data in the digital humanities : modeling texts and text-based resources*, dir. Julia Flanders et Fotis Jannidis, London ; New York, 2019 (Digital research in the arts and humanities), p. 99-116.
- — « What is TEI conformance, and why should you care ? », *Journal of the Text Encoding Initiative*, 12 (2019), doi : https://doi.org/10.4000/jtei.1777.
- Sahle (Patrick), « Digital modelling. Modelling the digital edition », dans *Medieval and modern manuscript studies in the digital age*, London/Cambridge, 2016, url : https://dixit.uni-koeln.de/wp-content/uploads/2015/04/Camp1-Patrick_Sahle_-_Digital_Modelling.pdf

Sur l'OHCO
- DeRose (Steven), Durand (David), Mylonas (Elli) et Renear (Allen), «What is Text, Really ?», Journal of Computing in Higher Education, I–2 (1990), p. 3-26, url : https://cs.brown.edu/courses/cs195-1/reading/WhatIsTextReally.pdf
- Renear (Allen), Mylonas (Elli) et Durand (David), « Refining our Notion of What Text Really Is : The Problem of Overlapping Hierarchies », dans Research in Humanities Computing, dir. Nancy Ide et Susan Hockey, Oxford, 1996, url : https://cds.library.brown.edu/resources/stg/monographs/ohco.html

Propositions alternatives au XML pour modéliser du texte
- Bleeker (Elli), Haentjens Dekker (Ronald) et Buitendijk (Bram), « Texts as Hypergraphs : An Intuitive Representation of Interpretations of Text », Journal of the Text Encoding Initiative, 14 (2021), doi : https://doi.org/10.4000/jtei.3919.
- Smith (David Neel) et Blackwell (Christopher W.), « Four URLs, limitless apps : Separation of concerns in the Homer Multitext architecture », dans Donum natalicium digitaliter confectum Gre- gorio Nagy septuagenario a discipulis collegis familiaribus oblatum : A Virtual Birthday Gift Presented to Gregory Nagy on Turning Seventy by His Students, Colleagues, and Friends, 2012, url : https://chs.harvard.edu/d-n-smith-c-w-blackwell-four-urls-limitless-apps-separation-of-concerns-in-the-homer-multitext-architecture/