Humanités Numériques

Cours 5 : Linguistique et Informatique partie 2

Loïc Grobol

2021-10-19

Les grands corpus

Premières initiatives :

- Corpus de Brown (1964)
- *British National Corpus* (1991)

En français :

- Frantext
- *Corpus for Idiolectal Research* (CIDRE)
- Corpus du Français Parlé Parisien
- *Multicultural Paris French* Corpus
- Corpus d'Étude pour le Français Contemporain
- ESLO

À l'heure du web

- *Common Crawl*
- *Open Super-large Crawled Aggregated coRpus* (OSCAR)

Les documents structurés

- Documents de base : texte comme flux de caractères
- Manquent :
  - Métadonnées
  - *Structure**s*** : paragraphes, sections, chapitres, en-têtes…
  - Pas les mêmes types selon les docs : roman, journaux, dictionnaires, juridique…

HTML

***H**yper**t**ext **M**arkup **L**anguage*

- Créé autour de 1991 par Tim Berners-Lee
  - Avec les adresses URL et le protocole HTTP c'est ce qui fait le web
  - Objectif : un format simple pour décrire des documents textuels structurés, incluant des ressources multimédia et **liés** entre eux.
- À l'origine un cas particulier de SGML, s'en éloigne avec le temps

— Toujours en évolution pour s'adapter aux usages

```html
<!DOCTYPE html>
<html>
  <head>
    <title>This is a title</title>
  </head>
  <body>
    <p>Bonjour, tout le <a href="https://fr.wikipedia.org/wiki/Hello_world">monde</a></p>
  </body>
</html>
```

Exercice Ouvrez un éditeur de texte (par exemple le bloc note) et entrez (ou collez…) le fragment de code précédent. Ouvrez-le dans votre navigateur et contemplez.

Vous pouvez y faire des modifications, ajouter des éléments, changer la cible du lien. La documentation MDN est souvent d'un grand secours.

— Les éléments entre *chevrons* < et > sont des **balises** (*tags*, *elements*)
— Les balises structurent le document
  — Indiquent la fonction de chaque partie de texte
  — Séparent contenu et méta-données
  — Le résultat peut être vu comme un arbre
— Les balises peuvent contenir des métadonnées, comme la cible d'un lien
— En principe, informations purement *sémantique* : on ne code pas la mise en forme

Pourquoi des chevrons ?

C'est une convention historique d'annotation, recyclée

— Dans son usage philologique, s'utilise par paires et sert à ajouter un mot, un groupe de mots, un élément conjecturel dans une édition généralement scientifique.
— Sert de parenthèse qui permet d'isoler une portion de texte en l'isolant de la situation de communication.

XML

*E**x**tensible **M**arkup **L**anguage*

— Créé en 1996 sous les auspices du *World Wide Web Consortium* (W3C)
— Objectif : description structurée de tout type de données encodables comme du texte
— Au contraire de HTML, l'ensemble des balises n'est pas prédéfini et chaque document peut définir les siennes

---

```xml
<div type="act" n="II" xml:id="II"><head>Acte II</head>
   <div type="scene" n="2" xml:id="II2"><head>Scène 2</head>
     <sp><speaker>Rodrigue</speaker>
         <l part="I">À moi, comte, deux mots.</l></sp>
     <sp><speaker>Comte</speaker>
         <l part="M">Parle</l></sp>
     <sp><speaker>Rodrique</speaker>
         <l part="F">Ôte-moi d'un doute</l></sp>
     <sp><speaker>Comte</speaker>
         <l part="I">Connais-tu bien Don Diègue ?</l></sp>
     <sp><speaker>Comte</speaker>
         <l part="M">Oui</l></sp>
     <sp><speaker>Rodrigue</speaker>
       <l part="F">Parlons bas, écoute.</l>
       <l>Sais-tu que ce vieillard fut la même vertu,</l>
       <l>La vaillance et l'honneur de son temps ? Le sais-tu ?</l></sp>
   </div>
 </div>
```

La TEI

> La « Text Encoding Initiative » (TEI) est l'un des projets les plus durables et influents du champ aujourd'hui appelé « humanités numériques ». Son but est de fournir des recommandations pour la création et la gestion sous forme numérique de tout type de données créées et utilisées par les chercheurs en sciences humaines, comme les sources historiques, les manuscrits, les documents d'archives, les inscriptions anciennes et bien d'autres.

(Burnard, 2015)

— Un format de balisage et une communauté académique internationale
— Recommandations pour l'encodage de ressources numériques, et plus particulièrement de documents textuels.

La TEI met l'accent sur ce qui est partagé par tous les types de documents, qu'ils soient représentés physiquement sous une forme numérique sur un disque ou une carte mémoire, sous une forme imprimée comme un livre ou un journal, sous une forme écrite comme un manuscrit ou un codex, ou sous une forme inscrite dans la pierre ou sur une tablette de cire.

Cette continuité facilite la migration du texte depuis des manifestations plus anciennes, comme l'imprimé ou le manuscrit, vers d'autres plus récentes comme le disque ou l'écran.

C'est pourquoi la vision de la TEI de ce qu'est le texte est largement conditionnée par ce que le texte a été dans le passé, sans toutefois trop compromettre ce que le texte peut devenir dans le futur. Elle essaie de traiter tous les types de documents numériques de la même façon, qu'ils soient « nativement numériques » ou non.

La TEI fournit le nom et la définition de centaines de balises, en même temps que des règles sur la façon dont elles peuvent être combinées. Plus précisément, les Guidelines de la TEI définissent cinq ou six cents concepts différents, avec les spécifications détaillées des éléments et classes d'éléments XML qui peuvent être utilisés pour les représenter.

— La plupart des documents TEI n'a besoin que d'une petite partie de ce qui est proposé
— Mais rien n'est inutile, simplement différents documents ont différents besoins
  — Théâtre
  — Épigraphie, Paléographie
  — Dictionnaires
  — Transcriptions de la parole…

Bibliographie

Lou Burnard. 2015. *Qu'est-ce que la Text Encoding Initiative ?*Encyclopédie numérique. OpenEdition Press, Marseille, éditions, octobre.