Encoder en TEI

Simon Gabay

Université de Neuchâtel

12 novembre 2018

Edition numérique

Une édition numérique

Qu'est-ce qu'une édition numérique ?

- À une époque où tout le monde travaille sur un odrinateur, est-ce qu'utiliser le medium numérique, à un stade ou un autre (transcription, annotation, index. . .), est suffisant pour parler d'édition numérique ?
- Est-ce que diffuser une édition sur un medium numérique (pdf, clef USB, site internet. . .) est suffisant pour parler d'édition numérique ?
- (Cette question nous renvoie plus généralement au flou qui entoure la définition des humanités numériques.)

Une édition électronique savante

Qu'est-ce qu'une édition électronique savante ?

- Une édition qui utilise les possibilités de l'outil informatique pour proposer...
- ...une représentation critique ou enrichie sémantiquement d'une source...
- ...des fonctionnalités utiles à leur compréhension et leur analyse.

La philologie numérique

L'édition numérique savante s'appuie sur la philologie numérique

- L'ecdotique (l'édition d'un texte) n'est qu'une partie du travail
- La philologie (numérique) s'intéresse aussi à la création des données (transcription) ou leur exploitation (exploitation) en plus de leur diffusion.

Vocabulaire de la philologie

Le travail de philologie ecdotique doit suivre quelques grandes étapes, dont chacune connaît des solutions et des procédures informatiques spécifiques.

- Transcription
- Collation
- Analyse paléographique et codicologique
- Analyse linguistique
- Annotation
- Indexation
- Publication
- (Exploitation)
- Archivage

Une communauté

L'ecodtique est un art complexe et ancien, dont des rudiments sont nécessaires pour éditer un texte, numériquement ou pas. L'utilsation d'outils numériques complexifie encore le travail. Faire de l'édition numérique :

- C'est travailler en équipe
- C'est avoir recours à une communauté de recherche, qui partage des pratiques et une philosophie (cf. Humanistica, Cahier, EADH, DARIAH, Communauté TEI...)

XML-TEI

XML

Le XML

- XML signifie *Extensible Markup Language*
- C'est un **langage** de balisage (vs langage de programmation, de définition de données ou de requête)
- Comme tout langage, il est régi par des règles

Règles principales

Le langage de balisage fonctionne de manière simple

```
<élément attribut="valeur">donnée</élément>
```

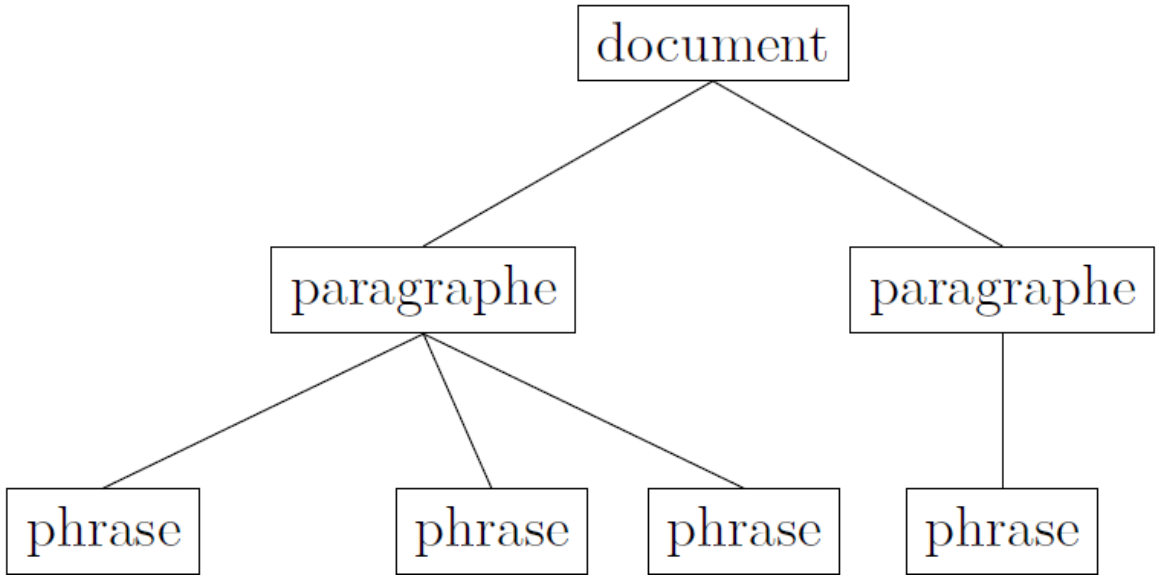
- Un <élément> est entre chevrons
- Une <balise> doit être fermé </balise>
- Une <balise1> ne doit <balise2> pas être croisé </balise1> avec un autre </balise2>
- Une <balise/> peut être auto-fermante
- Un <élément> peut porter un @attribut (noté avec un @)
- l'@attribut a une "valeur" (entre guillemets)

Du texte à la base de données

Du texte à la base de données

- Une donnée est enfermée entre deux balises. Pour nous il s'agit de textes, de paragraphes, de phrases...
- Les données sont "emboîtées" les unes dans les autres : un document contient des paragraphes, qui contiennent des phrases, qui contiennent des mots...
- On transforme ainsi le texte en base de données

Une structure arborescente



Document

On emploie a priori les italiques pour les termes empruntés à d'autres langues. On emploie les petites capitales pour les noms propres, comme Léopold Delisle. On emploie en revanche généralement le gras pour des raisons coupables.
On retourne à la ligne pour un nouveau paragraphe.

XML comme langage structuré (par des balises)

En pratique, on obtient obtient un document de ce type :

```
<document>
 <paragraphe>
  <phrase>
  On emploie <locutionÉtrangère>a priori</locutionÉtrangère> les
  italiques pour les termes empruntés à d'autres langues.
  </phrase>
  <phrase>
  On emploie les petites capitales pour les noms propres,
  comme <nom>Léopold Delisle</nom> ou <nom>Jules Quicherat</nom>.
  </phrase>
  <phrase>
  On emploie en revanche généralement le gras pour des raisons
  coupables.
  </phrase>
 </paragraphe>
 <paragraphe>
  <phrase>
  On retourne à la ligne pour un nouveau paragraphe.
  </phrase>
 </paragraphe>
</document>
```

Une question fondamentale

Une question fondamentale

- Nous avons ici utilisé <paragraphe> ou <phrase>, mais nous aurions pu choisir d'autres noms.
- Si nous étions italiens, nous aurions choisi <paragrafo> et <frase>
- Mais alors les documents sont encodés différemment: comment choisir des noms pour les <éléments> et les @attributs communs à tous?

La TEI

La TEI

- TEI pour *Text Encoding Initiative*
- Elle est créé en 1987 (donc avant internet)
- La TEI est pilotée par un consortium qui maintient et développe des recommandations pour l'encodage des textes
- Ces recommandations sont en constantes évolutions
- Elles sont disponibles en ligne `http://www.tei-c.org/guidelines/`

La TEI

Entre vocabulaire et langage

- Il existe d'autres vocabulaires XML que la TEI, comme l'EAD pour les archivistes...
- ...ou Dublin Core (DC) pour les bibliothécaires
- Ces vocabulaires peuvent d'ailleurs être exprimés avec d'autres langages (RDF-DC).
- Pour cette raison, on parle de XML-TEI, (ainsi il a existé un SGML-TEI).

Trois particularités

Trois particularités de la TEI

- Le vocabulaire est en anglais : on utilise une balise `<w>` (*word*) pour un `<w>mot</w>`
- Ce vocabulaire est limité: on ne peut (presque) pas inventer de nouvelles balises
- Elle propose autant que possible un encodage sémantique (à l'inverse de LaTeX, utilisé pour faire ces diapos)

Sémantique et procédural

On emploie *a priori* les italiques pour les locutions et termes empruntés à d'autres langues.

Procédural

```
On emploie <italique>a priori</italique> les italiques pour les locutions et termes empruntés à d'autres langues.
```

semantique

```
On emploie<locutionEtrangère>a priori</locutionEtrangère> les italiques...
```

semantique II

```
On emploie<latin>a priori</latin> les italiques...
```

La solution en TEI

En XML-TEI, on encoderait le document ainsi :

La solution en TEI

```
 On emploie<foreign xml:lang="la">a priori</foreign> les
italiques...
```

Avant toute chose

Modélisation

La première étape est la Modélisation

Définition

Opération par laquelle on établit le modèle d'un système complexe, afin d'étudier plus commodément et de mesurer les effets sur ce système des variations de tel ou tel de ses éléments composants. [a]

a. J. Giraud, P. Pamart, J. Riverain, *Les nouveaux mots « dans le vent »*, Paris, France, 1974).

Modélisation

Il s'agit de définir un modèle adapté

- Aux documents que l'on édite
- À nos questions de recherche
- Aux moyens (techniques, financiers. . .) dont on dispose

Attention ! Il est souvent coûteux et compliqué de revenir sur certains choix. Il s'agit donc de bien réfléchir !

Modélisation

Concrètement, pour un philologue, les premières questions sont les suivantes ?

- Quels passages du textes doivent être balisés ? Les noms ? les locutions étrangères tous les mots ? Doit-on mettre la catégorie morpho-syntaxique et le lemme ?
- Doit-on représenter la structure physique du document (folios, pages. . .) ou la structure logique (chapitres, parties. . .)
- Attention, il est (presque) impossible de tout faire : il faut choisir !

Modéliser

Deux encodages différents d'un même texte

Structure logique

```
<document>
 <paragraphe>
  <phrase>
   On emploie <locutionÉtrangère>a priori</locutionÉtrangère> les
   italiques pour les termes empruntés à d'autres langues.
  </phrase>
  <phrase>
   On emploie les petites capitales pour les noms propres,
   comme <nom>Léopold Delisle</nom> ou <nom>Jules Quicherat</nom>.
  </phrase>
 </paragraphe>
</document>
```

Structure physique

```
<document>
 <pb n="1"/>
 On emploie <locutionÉtrangère>a priori</locutionÉtrangère> les
 italiques pour les termes empruntés à d'autres langues. On
 <pb n="2"/>
 emploie les petites capitales pour les noms propres, comme
 <nom>Léopold Delisle</nom> ou <nom>Jules Quicherat</nom>.
</document>
```

Granularité

La granularité

Définition

Degré de finesse ou précision d'un modèle, conçu comme le niveau de son plus petit composant. Plus la granularité est grande, plus on descend dans la modélisation (niveau phrase, mot, graphème, etc.) – et plus on ajoute de balises.

Modéliser

Deux encodages différents d'un même texte

Faible granularité

```
<document>
 <paragraphe>
  <phrase>
   On emploie a priori les
   italiques pour les termes empruntés à d'autres langues.
  </phrase>
  <phrase>
   On emploie les petites capitales pour les noms propres,
   comme <nom>Léopold Delisle</nom> ou <nom>Jules Quicherat</nom>.
  </phrase>
 </paragraphe>
</document>
```

Moyenne granularité

```
<document>
 <paragraphe>
  <phrase>
   On emploie <locutionÉtrangère>a priori</locutionÉtrangère> les
  italiques pour les termes empruntés à d'autres langues.
  </phrase>
  <phrase>
   On emploie les petites capitales pour les noms propres,
   comme <nom>Léopold Delisle</nom> ou <nom>Jules Quicherat</nom>.
  </phrase>
 </paragraphe>
```

Modéliser

On peut aller très loin

```
<document>
 <paragraphe>
  <phrase>
   <w lemme="on" POS="PROper">On<w/>
   <w lemme="employer" POS="VERcjg">emploie<w>
   ...
```

Pourquoi la TEI ?

Défauts de la TEI

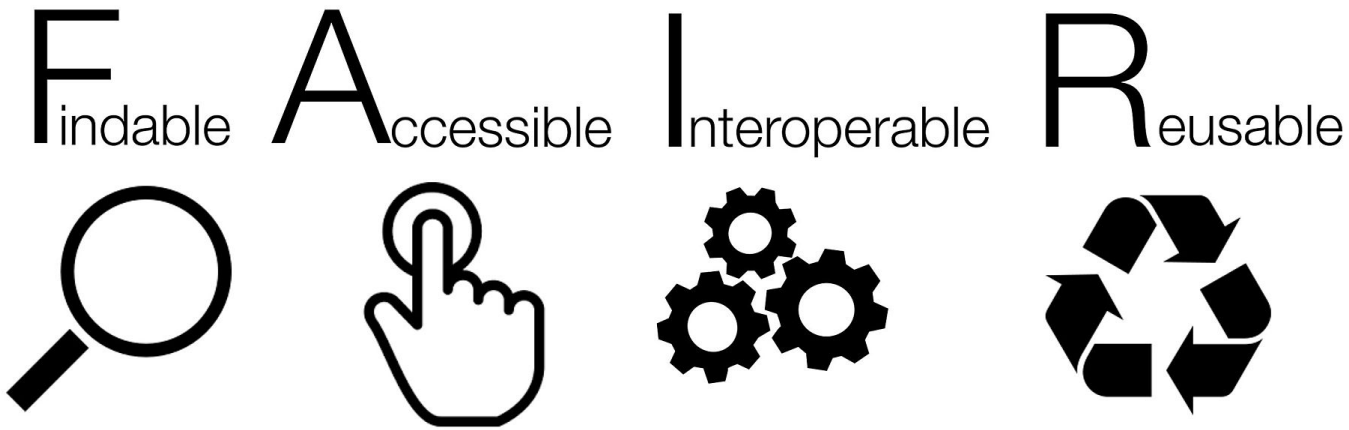
La TEI pose des problèmes

- Elle force à utiliser un standard, par définition générique, et qui ne convient pas forcément exactement à nos données
- Elle nécessite un apprentissage, notamment pour respecter le sémantisme du vocabulaire

Alors pourquoi la TEI ?

Fair Data

La TEI permet de répondre (en partie) aux principes du Fair Data



(Img. SangyaPundir CC BY-SA 4.0)

Les principes soutenus par les États et agences de recherche (ANR, ERC, FNS, DFG. . .)

Fair Data

Les quatre grands principes du Fair Data :

- *Findable* : Les données sont faciles à trouver en ligne, y compris sur le long terme (archivage pérenne) ;
- *Accessible* : Les métadonnées accompagnent les données, y compris quand ces dernières ne peuvent être diffusées (*open science*)
- *Interoperable* : Le format doit être ouvert, libre, documenté et compréhensible ;
- *Reusable* : La modalités techniques et légales de réutilisation sont claires.

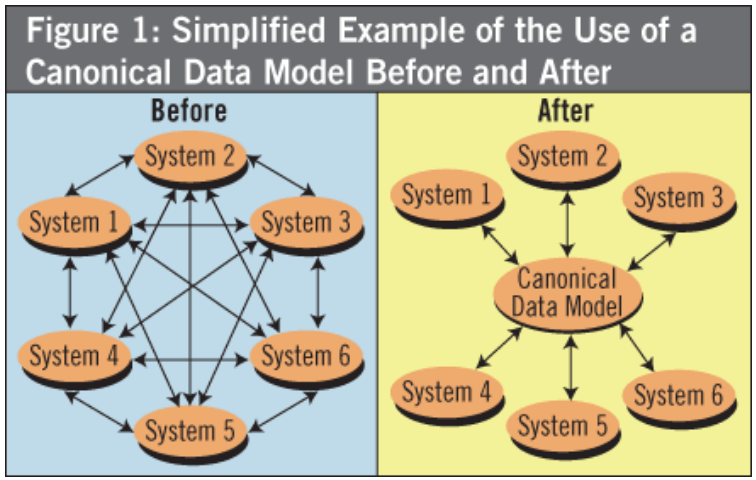
Bonnes pratiques

La TEI encourage (force ?) à avoir des bonnes pratiques

- Utilisation de métadonnées (des données sur les données)
- Identification claire des auteurs et de leurs responsabilités
- Propose une documentation rigoureuse

Format pivot

La TEI est un format pivot



extr. de Steve Hoberman, « Canonical Data Model », *Information Management Magazine*, en ligne : `http://www.information-management.com/issues/2007_50/10001733-1.html`.

Chaîne de traitement

La TEI est une étape de la chaîne de traitement

- La chaîne de traitement numérique des documents (*Workflow*) est une suite de tâches ou opérations
- Elle implique l'utilisation de différentes solutions informatiques, par exemple :
  - XML-TXM pour l'analyse linguistique
  - HTML pour la publication web
  - LaTeX pour l'édition papier ou pdf
  - ...
- LA TEI n'est pas la réponse à tout
- Mais une attention particulière est portée à son intégration dans la chaîne de traitement

Conclusion

Conclusion

Conclusion

- La TEI n'est qu'un (petit) morceau de la philologie numérique
- La philologie numérique est avant tout de la philologie
- La philologie numérique nécessite des compétences en humanités numériques
- D'autres formations existent