Décrire et manipuler un document numérique

Cours 9 : TEI

Loïc Grobol <lgrobol@parisnanterre.fr>

2022-03-21

Sources et compléments

Ce cours s'inspire entre autre

— Des transparents de la formation « Introduction à la Text Encoding Initiative » de Florian Chiffoleau.
— Du cours « Introduction à la TEI » de Jean-Baptiste Camps.
— Du tutoriel interactif « TEI by example ».

Principes de la TEI

Historique

— Novembre 1987 : Création de la Text Encoding Initiative (TEI)
  — → Pallier à la prolifération de systèmes divers et incompatibles pour la représentation numérique de textes
— Mai 1994 : Publication de la première version officielle des TEI Guidelines (P3)
  — → Règles produites par le travail combiné de nombreux ateliers de réflexions et par les révisions et extensions faites à la version P1 produite en 1990

— Janvier 1999-Janvier 2001 : Création du consortium TEI
  — → Avoir une organisation officielle qui maintient, développe et promeut la TEI
— Novembre 2007 : Publication de la dernière version des TEI Guidelines (P5)
  — → Version révisée de P4 avec de nouveaux développements pour un certain nombre de domaines
  — → Version disponible en XML
  — → Mises à jour deux fois par an depuis

Les *Guidelines*

https://tei-c.org/guidelines/p5/

— Règles définies pour l'encodage de tous les textes sous leur format numérique et lisibles par machine.
— Une liste longue, détaillée et en constante évolution.
— De nombreuses balises pour encoder des corpus de type, langue et structure différentes (romans, pièces de théâtre, poème, lettres, rapport officiel, etc.).
— Page unique pour chaque balise et attribut, afin d'avoir toutes les informations sur la manière dont ils s'utilisent.
  — Exemple : la balise `<p>`

Le consortium

Special Interest Groups (SIGs) :

- Groupe de réflexion, d'échanges et de débats sur des sujets spécifiques liés à la TEI
- Aide au développement et à l'évolution de la TEI
- Trois exemples de SIGs : Computer-Mediated Communication SIG, Correspondence SIG, East Asian/Japanese SIG

---

jTEI, le journal de la TEI :

- Journal officiel du consortium TEI
- Articles sur l'état de l'art, des innovations en matière de TEI et des exploitations dans des projets
- Articles rigoureusement évalué avant publication sur la plateforme Open Edition

---

TEI-L mailing-list :

- Ouverte à toute la communauté TEI, pour poser des questions ou y répondre
- Partage d'expertise ou d'expérience
- Archivage de tous les problèmes précédemment rencontrés

En pratique

Concrètement, l'idée de la TEI c'est de fournir une boîte à outils pour la représentation numérique de documents textuels (au sens assez large).

- Plein de gens ont besoin de créer des normes différentes pour des besoins différents.
- « Si vous avez besoin de représenter tel type d'information, voici une façon standard »
- Standardisation :
  - Évite de dupliquer le travail : encodage **et** exploitation.
  - Rend plus facile l'interopérabilité.
- On garde la flexibilité : les *guidelines*, ce n'est pas une norme, mais des **briques* pour faire des normes.

Applications à plein de domaines : les *guidelines* sont très volumineuses et sophistiquées.

- Découpage en modules : par exemple msdescription pour la représentation des textes manuscrites.
- Regroupement des éléments en classes : par exemple le groupe `att.written` regroupe les éléments pour lesquels ça peut faire sens de préciser un scribe. Ces éléments possèdent tous l'attribut `hand` qui sert à identifier ce scribe.

ODD et Roma

- Les normes TEI sont formalisables dans les formats habituels de spécification XML
- Mais c'est souvent très verbeux et complexe
  - → compter environ 2k lignes pour une DTD basique
- On préfére en général les écrire dans le format ODD « One Document Does it all », spécifique à la TEI, qui donne à la fois une spécification et sa documentation.
  - Les documents ODD sont eux-même écrits dans une norme TEI.
  - Et les guidelines sont elles écrites en ODD.
- Et plutôt que de l'écrire à la main, on conseille très fortement d'utiliser l'outil ROMA.

Structure des documents TEI

OK, mais à quoi ça ressemble **concrètement** un document TEI

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
  <!---...-->
  </teiHeader>
  <text>
  <!--...-->
  </text>
</TEI>
```

Espaces de noms

`xmlns` ?

C'est ce qu'on appelle un « espace de nom », un *namespace*.

En fait les éléments TEI on des noms à rallonge. Par exemple `<p>`, c'est le petit nom de `<http://www.tei-c.org/ns/1.0:p>`.

Pourquoi à votre avis ?

Plusieurs spécifications XML peuvent partager des attributs du même nom court. Par exemple `<p>` existe aussi dans la norme XHTML (une tentative ratée de décrire HTML en XML) mais avec des caractéristiques (par exemple des attributs) différentes.

Le fait d'avoir un nom long permet de ne pas les confondre.

Ainsi le `<p>` de XHTML est en fait un `<http://www.w3.org/1999/xhtml:p>`.

Évidemment on ne veut pas écrire ces noms à rallonge dans nos fichiers. Pour ça on peut déclarer des raccourcis :

```
<tei:TEI xmlns:tei="http://www.tei-c.org/ns/1.0">
  <tei:teiHeader>
  <!--...-->
  </tei:teiHeader>
  <tei:text>
  <!--...-->
  </tei:text>
</tei:TEI>
```

Ici `tei` est déclaré comme abréviation pour `http://www.tei-c.org/ns/1.0`.

Et on peut aussi déclarer un espace de nom implicite, qui sera utilisé par défaut pour tous les noms pour lesquels on en précise pas d'autre :

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
  <!--...-->
  </teiHeader>
  <text>
  <!--...-->
  </text>
</TEI>
```

Dans les documents TEI habituels, on utilise en général seulement l'espace de noms `http://www.tei-c.org/ns/1.0` et `xml:http://www.w3.org/XML/1998/namespace` qui est déclaré implicitement en XML et contient les attributs suivants :

- `xml:lang` pour la langue d'un élément.
- `xml:space` pour spécifier si les caractères blancs (espaces) sont pris en compte.
- `xml:base` pour spécifier l'URL de base pour les URL d'un élément.
- `xml:id` l'identifiant par défaut d'un élément.

En-tête

Comme en HTML, les documents TEI comprennent obligatoirement un en-tête (*header*) qui embarque les métadonnées du document. Il est beaucoup plus comple(t|xe) que celui de HTML.

L'en-tête TEI se compose au minimum d'une description du fichier électronique `<fileDesc>` composé de trois sections obligatoires :

- `<titleStmt>` : titre (`<title>`), auteurice(s) (`<author>`) et responsables de la production du fichier.
- `<publicationStmt>` : détail de la publication du fichier, peut contenir des paragraphes
- `<sourceDesc>` : origine du document électronique, par exemple s'il s'agit d'une transcription d'un document papier.

```xml
<teiHeader xmlns="http://www.tei-c.org/ns/1.0">
  <fileDesc>
    <titleStmt>
      <title>The Strange Adventures of Dr. Burt Diddledygook: a machine-readable transcription</title>
      <respStmt>
        <resp>editor</resp>
        <name xml:id="EV">Edward Vanhoutte</name>
      </respStmt>
    </titleStmt>
    <publicationStmt>
      <p>Not for distribution.</p>
    </publicationStmt>
    <sourceDesc>
      <p>Transcribed from the diaries of the late Dr. Roy Offire.</p>
    </sourceDesc>
  </fileDesc>
</teiHeader>
```

Corps

Un document TEI contient en général un (ou plusieurs) texte `<text>`.

Toujours comme en HTML, un texte a un corps (`<body>`), qui contient des structures textuelles de base comme les paragraphes (`<p>`) et des structures spécifiques suivant les genres :

— Lignes (`<l>`) pour la poésie.
— Répliques (`<sp>`) pour le théatre.
— …

```xml
<text xmlns="http://www.tei-c.org/ns/1.0">
  <body>
  <p>For the first time in twenty-five years, Dr Burt Diddledygook decided not to turn up to the annual
  </body>
</text>
```

Front

Un texte peut aussi optionnellement contenir un `<front>`, avec préface, table des matières, dédicace…

```xml
<front xmlns="http://www.tei-c.org/ns/1.0">
  <div type="dedication">
    <p>In memory of Lisa Wheeman.</p>
  </div>
  <div type="contents">
    <head>Table of Contents</head>
```

```
    <list>
      <item>I. The Decision</item>
      <item>II. The Fuss</item>
      <item>III. The Celebration</item>
    </list>
  </div>
</front>
```

Back

```
<back xmlns="http://www.tei-c.org/ns/1.0">
  <div type="colophon">
    <p>Typeset in Haselfoot 37 and Henry 8. Printed and bound by Whistleshout, South Africa.</p>
  </div>
</back>
```

Et maintenant ?

Maintenant, quand vous aurez besoin de représenter des documents textuels (et plus) sous forme numérique, vous pourrez utiliser un format TEI en :

— Identifiant dans les *guidelines* les élémenst qui vous intéressent et les modules associés
— Concevant une spécification à l'aide de ROMA
— Encoder vos documents, idéalement dans un éditeur adapté, mais à défaut dans n'import quel éditeur de texte.

Et comment on affiche toutes ces données ? C'est une histoire pour une prochaine fois

Exercice

Pour chacun des chapitre 5 à 13 des *guidelines*, dire rapidement ce que contient le module décrit dans ce chapitre et quel(s) type(s) de documents il peut concerner.

Il n'est probablement pas nécessaire de lire tous ces chapitres en entier.