

Introduction à la *Text Encoding Initiative*

1. Le « projet TEI » : genèse

Initié en 1987 par trois associations anglo-saxonnes¹, le projet de recherche « *Text Encoding Initiative* » a pour objectif de normaliser l'encodage des textes au format numérique dans les sciences humaines, afin d'en faciliter l'analyse et la manipulation. L'encodage d'un texte électronique permet d'ajouter au fichier des données dites « de marquage » qui, dans les langages SGML et XML, permettent de décrire non seulement le contenu mais aussi la structure logique du document. Les disciplines des sciences humaines tirent partie de ce type d'encodage qui facilite l'analyse et le traitement de vastes corpus de textes. Cependant, la grande diversité des pratiques nuit à l'échange, à l'importation et au traitement des fichiers sur des applications multiples.

Un modèle général de structuration des données, la TEI, a été élaboré pour harmoniser ces pratiques et répondre ainsi aux besoins de la communauté scientifique. **Ce projet en open source a connu plusieurs versions, de la TEI P1 en 1990 à la TEI P5 en 2007**, dernière en date, qui comporte d'importantes améliorations. **Le TEI consortium, créé en 1999, est devenu une fondation internationale à but non lucratif dont plus de 90 institutions sont membres aujourd'hui.** En France, quelques établissements de recherche et d'enseignement y prennent part (telles l'Ecole nationale des Chartes, l'enssib, deux unités de recherche du CNRS) et Nancy est devenue en 2005 le centre européen de support au consortium TEI.

2. La théorie ...

La TEI est une norme de balisage des textes électroniques fondée sur XML. Utilisable pour tous types de textes littéraires, ce modèle s'applique quelle que soit la langue dans laquelle le texte est écrit et prend en compte toutes ses spécificités, que ce soit le document même (chapitres, paragraphes, strophes, ...) ou son appareil critique (commentaire éditorial, interprétation, analyse, ...).

Cette norme est constituée :

- d'un ensemble d'éléments² organisés en modules distincts, les « **tag sets** », que l'on sélectionne en fonction de ses besoins pour former une *Document Type Definition* (DTD).
- d'un ensemble de **Recommandations** (les « *Guidelines* ») qui expliquent comment utiliser la DTD.

L'intérêt d'utiliser une telle norme réside dans sa richesse et sa souplesse. Selon Lou Burnard, **la TEI est un système extensible, modulaire et polymorphe constituant un modèle abstrait.** Sa modularité constitue une de ses grandes forces, lui permettant d'être un modèle international et interdisciplinaire. Ainsi, on encode un document à partir d'un **module-noyau obligatoire**, le « *core tagset* », qui rassemble les éléments communs à tous les types d'information. On choisit ensuite un **jeu de base** (« *base tagset* ») parmi les sept proposés : TEI.prose pour les textes en prose, TEI.verse pour les textes en vers, TEI.spoken pour la transcription d'interview, etc. Enfin, on ajoute les **modules additionnels** (« *additional tagset* ») dont on a besoin : TEI.figures permet de décrire les images, illustrations, tables et formules, TEI.analysis fournit des éléments simples pour la description de l'analyse des textes, TEI.transcr pour la transcription des sources historiques écrites et de bibliographie analytique, etc.

On peut donc concevoir autant de DTD TEI que ce que les combinaisons de modules le permettent. Les textes encodés selon les recommandations peuvent être réutilisés dans un très grand nombre d'applications actuelles, et la norme prévoit aussi leur usage future dans des applications qu'il reste encore à développer.

La TEI toutefois comporte certaines limites. **Paradoxalement, les atouts que sont sa richesse et sa généricité la rendent complexe à utiliser.** La majorité des utilisateurs emploient la version simplifiée de la TEI, la « *TEI Lite* », qui suffit à les satisfaire. En France, certains projets de normalisations sont menés à l'heure actuelle, mais le modèle TEI reste encore peu connu. De façon symptomatique, la traduction en français de toutes les recommandations n'a pas encore été menée à bien.

3. ... et la pratique

Un document respectant la norme TEI se présente ainsi³ :

¹ The Association of Computers in the Humanities, the Association of Computational Linguistics et the Association for Literary and Linguistic Computing.

² Concept du langage SGML. Des connaissances basiques sur XML sont nécessaires pour bien comprendre le fonctionnement de la TEI.

³ <http://xml.coverpages.org/beaudry-TEI.html>

```

<!DOCTYPE tei [ <!ENTITY TEI.prose "INCLUDE">
<tei>
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Le plus petit document conforme à la TEI</title>
      </titleStmt>
      <publicationStmt>
        <p>Ce document n'est pas publié.</p>
      </publicationStmt>
      <sourceDesc>
        <p> Ce document est original.</p>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <text>
    <body>
      <p>Voici le document conforme à la TEI le plus court qu'on puisse imaginer. </p>
    </body>
  </text>
</tei>

```

Un **document conforme à la TEI doit comporter, outre la transcription même du texte, un « header »**. Cet ensemble de balises obligatoires, inspiré de l'ISBD, contient des métadonnées sur le document balisé (le titre, l'auteur, la source, les principes retenus pour le balisage, des informations sur l'histoire du texte, en particulier ses révisions et ses modifications). Des outils sont disponibles sur le site du TEI consortium pour aider les utilisateurs à générer des documents et des schémas conformes à la TEI : Pizza Chef, conçu par le TEI consortium pour faire comprendre le concept de modularité, puis **ROMA**.

La TEI s'adresse en priorité aux chercheurs qui travaillent sur des corpus électroniques, mais peut aussi intéresser des bibliothèques, des musées ou des éditeurs. En France, plusieurs projets sont en cours : l'Université Lyon 2 l'utilise pour dématérialiser, publier et archiver ses thèses électroniques, l'enssib pour encoder le *Bulletin des Bibliothèques de France*. L'Ecole nationale des Chartes choisit systématiquement la TEI pour les textes de sa collection Editions électroniques.

Liens	<p>Références incontournables :</p> <p>ECOLE NATIONALE DES CHARTES. Formation TEI à l'École nationale des chartes, juin 2009 [en ligne]. Disponible sur : <http://www.enc.sorbonne.fr/formation-TEI-juin-2009/index.htm> ressources pédagogiques d'une formation sur la TEI à l'Ecole des Chartes en juin 2009.</p> <p>LOISEAU, Sylvain. « Les standards : autour d'XML et de la TEI » in <i>La Manufacture</i> [en ligne]. Disponible sur : http://www.revue-texto.net/Corpus/Manufacture/ [consulté le 10 septembre 2009]</p> <p>POUPEAU, Gauthier. <i>Les petites cases</i> [en ligne]. Disponible sur : <http://www.lespetitescases.net/> blog de Gauthier Poupeau qui comporte plusieurs billets sur la TEI et le travail de l'ENC sur le cartulaire blanc de l'Abbaye de Saint Denis, ainsi que des liens vers d'autres ressources.</p> <p>TEXT ENCODING INITIATIVE CONSORTIUM. <i>TEI: Text Encoding Initiative</i> [en ligne] Disponible sur : <http://www.tei-c.org/index.xml> le site web du TEI consortium, qui comporte les <i>Guidelines</i>, des tutoriels, les projets en cours, l'agenda des activités des groupes de travail...</p> <p>Quelques exemples d'utilisation :</p> <p>CLAVAUD, Florence. « L'Ecole nationale des Chartes et les « humanités numériques » : premiers bilans et perspectives » [en ligne] in <i>Ecole de bibliothéconomie et des Sciences de l'information</i>. Disponible sur : <http://www.ebsi.umontreal.ca/confmidi/2009/confmidi-fclavaud.pdf> intervention de Florence Clavaud à l'EBSI sur l'utilisation du XML et de la TEI, en particulier à l'Ecole des Chartes.</p> <p>DUFOURNAUD Nicole, FEKETE Jean-Daniel. « Analyse historique de sources manuscrites : application de TEI à un corpus de lettres de rémission du XVIème siècle » in <i>Document Numérique</i>, 1999, vol.3, 1-2, p. 117-134. Aussi disponible sur CAIRN.</p> <p>HEIDEN Serge, LAVRENTIEV Alexei. « Ressources électroniques pour l'étude des textes médiévaux : approches et outils », in <i>Revue Française de Linguistique Appliquée</i> 2004/1, Volume IX, p. 99-118.</p>
--------------	---

Fiche réalisée par : Béatrice Crassous
Créée le : 6 septembre 2009