

Cahiers **GUT** *enberg*

🔗 ÉDITORIAL : TEI : TEXT ENCODING
INITIATIVE

👤 François ROLE

Cahiers GUTenberg, n° 24 (1996), p. 1-3.

<http://cahiers.gutenberg.eu.org/fitem?id=CG_1996__24_1_0>

© Association GUTenberg, 1996, tous droits réservés.

L'accès aux articles des *Cahiers GUTenberg*

(<http://cahiers.gutenberg.eu.org/>),

implique l'accord avec les conditions générales

d'utilisation (<http://cahiers.gutenberg.eu.org/legal.html>).

Toute utilisation commerciale ou impression systématique

est constitutive d'une infraction pénale. Toute copie ou impression

de ce fichier doit contenir la présente mention de copyright.

Éditorial :

TEI – Text Encoding Initiative

Les *Cahiers GUTenberg* traitent souvent du logiciel de formatage de textes $\text{T}_{\text{E}}\text{X}$ et de son utilisation pour les sciences mathématiques. Mais ce n'est pas son seul domaine et dès les premiers numéros le ton était donné avec des articles sur le multilinguisme (numéro 0, avril 1988) ou la critique littéraire (article de R.P. Poggenbur sur Baudelaire, numéro 1, avril 1989). L'une des toutes premières manifestations de l'association GUTenberg, avant même qu'elle s'appelle ainsi, a d'ailleurs été une journée « $\text{T}_{\text{E}}\text{X}$ et les sciences humaines » organisée en commun avec l'IRHT (Institut de Recherche sur l'Histoire des Textes) à Paris en mai 1987. Cette association et ses *Cahiers* se veulent donc ouverts au codage électronique des textes. C'est dans ce contexte qu'il nous a été demandé de préparer ce numéro sur un sujet qui ne relève pas directement de $\text{T}_{\text{E}}\text{X}$, mais dont les habitués de $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$ verront la similitude.

Lors d'une conférence internationale tenue en 1987 au Vassar College (Poughkeepsie, New York) des chercheurs représentatifs d'un grand nombre de disciplines ont jeté les bases d'un projet connu depuis sous le nom de *Text Encoding Initiative (TEI)*.

Comme son nom le suggère, la TEI a pour objectif de faciliter l'échange et la circulation des documents électroniques au sein de la communauté scientifique. La variété des équipements informatiques et des codages associés constitue en effet souvent un obstacle pour les équipes de recherche qui souhaitent partager des volumes de données importants (par exemple de larges corpus linguistiques ou éditoriaux). En recommandant l'utilisation de conventions de codage neutres par rapport aux équipements et largement appliqués la TEI veut contribuer à améliorer cette situation.

Les travaux de la TEI¹ ont débouché sur la publication d'un ensemble de recommandations dont la dernière édition a été publiée en 1994 sous le titre *Guidelines for Electronic Text Encoding and Interchange*². Ces recommandations ont

1. La présentation – pages suivantes – de Nancy Ide et de Jean Veronis, qui ont suivi les travaux de la TEI depuis le début, rappelle dans quelles conditions se sont déroulés ces travaux durant la phase la plus active entre 1987 et 1994

2. Dans la suite de ce numéro nous désignons ce document par l'expression « les *Recommandations de la TEI* » ou tout simplement « les recommandations ».

au cours des dernières années été mises en œuvre dans le cadre de plusieurs projets importants.

Les recommandations de la TEI s'appuient sur la norme SGML. De fait, elles contiennent une ou plutôt un ensemble de DTD très riches, accompagnées de commentaires très détaillés sur l'usage des balises et des attributs.

Compte tenu de l'ampleur des recommandations de la TEI³, il n'est pas possible d'en donner une présentation complète. Autour d'un noyau de balises commun à tous les types de textes, la TEI contient en effet de nombreux modules additionnels adaptés à une grande variété de types de textes.

Pour parvenir à une présentation cohérente, nous nous sommes un peu inspirés de cette structure modulaire en composant ce cahier. Après une introduction de Nancy Ide et Jean Véronis et des rappels de Jacques André sur les documents structurés, ce numéro spécial s'ouvre en effet sur une traduction de la *TEI Lite*, un document d'initiation rédigé l'an dernier par deux des pères fondateurs de la TEI, Lou Burnard et C. M. Sperberg-McQueen. Cette traduction permet au lecteur francophone d'avoir un aperçu global de la TEI et notamment des conventions de codage les plus générales.

Les articles suivants donnent des exemples d'utilisation spécialisée (codage de dictionnaires, de corpus linguistiques ou d'apparats critiques, mise en page de documents bilingues, etc.). Tous ces articles sont basés sur des recherches faites en France ce qui prouve que si la TEI y a été longtemps boudée, elle intéresse maintenant toute une classe de spécialistes des textes et de l'industrie de la langue.

En conclusion, nous souhaitons que ce numéro contribue à mieux faire connaître ce projet au public francophone, et qu'il donne envie au lecteur intéressé de se reporter à la version intégrale des *Recommandations*.

Remerciements : merci à Lou Burnard qui nous a autorisé à publier la traduction française de ce qui fait la plus grosse partie de ce *Cahier*; la réalisation pratique de celui-ci doit beaucoup à ceux qui ont collaboré à sa préparation, à sa mise en page et à sa relecture, notamment Jacques André, Emmanuel Saint-James et Jean Véronis.

François ROLE
DISTNB
Ministère de l'Éducation nationale,
de l'Enseignement supérieur
et de la Recherche

3. 1300 pages dans la dernière édition!

1. On trouvera les références et les moyens de se procurer la TEI et sa documentation dans les divers articles qui suivent cet éditorial, notamment dans la présentation de Nancy Ide et Jean Véronis (pages 4 sqq) et dans celui de Lou Burnard et C.M. Sperberg-McQueen (pages 23 sqq).
2. Pour valider des documents codés selon la TEI, il suffit d'utiliser des «parseurs» SGML classiques comme SGMLS ou nSGMLS dont on trouvera un inventaire dans:
<http://www.sil.org/sgml/publicSW.html#parsers>
et notamment dans:
<ftp://ftp.jclark.com/pub/sp/>
3. Pour effectuer des traitements (notamment des conversions de format, ou des extractions de données) on peut utiliser des outils SGML du domaine public permettant de traiter les sorties des parser sgmls et nsgmls (ce sont en général des «package» Perl5). Citons SGMLS.PM:
<http://aix1.uottawa.ca/~dmeggins/SGMLSpm/sgmlspm.html>
et dtd-fragments:
<http://ftp.ifi.uio.no/pub/SGML/demo/>
4. Pour consulter des documents TEI, on peut utiliser les logiciels d'affichage SGML courants. Mais on privilégiera des logiciels qui peuvent s'intégrer aux navigateurs WEB comme, par exemple, le logiciel *Panorama* développé par SoftQuad et le NCSA:
<http://www.oclc.org:5046/oclc/research/panorama/panorama.html>
<http://www.sq.com/products/panorama/pan-free.html>
5. Un traducteur TEI Lite => \LaTeX est en cours d'écriture par Jean-Daniel Fekete et sera très prochainement à l'url:
<http://ftp.lri.fr/LRI/soft/ihm/tei2latex-0.1.tar.gz>
tei2latex est un programme Perl 5 qui traduit un document *TEI Lite* en $\LaTeX 2_{\epsilon}$. Un utilisateur peut utiliser plusieurs niveaux de configuration pour adapter la traduction d'un document *TEI Lite* à ses besoins et goûts. *tei2latex* repose sur le module Perl5 *SGMLS.pm* de David Megginson (disponible à: <ftp://aix1.uottawa.ca/pub/dmeggins> et à <ftp://ftp.lri.fr/LRI/soft/ihm/SGMLSpm-1.03ii.tar.gz>) qui utilise la sortie de l'analyseur SGML *nsgmls* ou *sgmls* de James Clark (disponible sur plusieurs sites).