Cahiers GUTenberg

PRÉSENTATION DE LA TEI

Nancy IDE, Jean VÉRONIS

Cahiers GUTenberg, nº 24 (1996), p. 4-10.

<http://cahiers.gutenberg.eu.org/fitem?id=CG_1996___24_4_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.

Présentation de la TEI : *Text Encoding Initiative*

Nancy IDE[a b] et Jean VÉRONIS[b]

[a] *Department of Computer Science*
Vassar College
Poughkeepsie
New York, NY 12601, USA
ide@cs.vassar.edu

[b] *Laboratoire Parole et Langage*
Université de Provence et CNRS
29, avenue Robert Schuman
13621 Aix en Provence Cedex 1, France
veronis@univ-aix.fr

1. Historique

La *Text Encoding Initiative* (TEI) est un projet international qui vise à la mise au point d'un ensemble de normes pour la préparation et l'échange de textes électroniques. La TEI est née lors d'une réunion organisée en novembre 1987 au Vassar College (Poughkeepsie, New York) par l'un des présents auteurs (Nancy Ide) et à laquelle ont participé diverses personnalités travaillant dans le domaine de l'archivage, de la structuration ou de l'analyse des textes électroniques (voir un historique détaillé dans [5]). À l'époque, l'énorme variété des formats de codage et de représentation des textes (à peu près tous mutuellement incompatibles) était perçue comme un obstacle majeur à l'échange des données et à la recherche. Les chercheurs présents à Vassar sont tombés d'accord sur la nécessité de travailler à la définition d'un nouveau format de codage des textes électroniques et en ont posé les principes de base. Le nouveau format devait:

- être aussi complet que possible,
- être simple, clair et concret,
- être facile à utiliser sans logiciel particulier,
- être rigoureusement défini,
- permettre un traitement efficace,
- être ouvert à des extensions définies par les utilisateurs,

– être compatible avec les standards existants ou en développement.

Le TEI a été créée officiellement en 1988 sous l'égide de l'*Association for Computers and the Humanities*, de l'*Association for Computational Linguistics* et de l'*Association for Literary and Linguistic Computing*. Le projet a été financé par le *U.S. National Endowment for the Humanities*, la Commission Européenne (DG XIII), la fondation Andrew W. Mellon et le *Social Science and Humanities Research Council du Canada*. De nombreux chercheurs à travers le monde ont travaillé regroupés dans des comités traitant chacun d'un thème précis. L'ensemble a été coordonné par un Comité de Pilotage (présidé successivement par Nancy Ide, Don Walker, Susan Hockey et David Barnard) et deux éditeurs (Michael Sperberg-McQueen et Lou Burnard).

En mai 1994, le travail effectué par les différents comités a été publié sous forme de *Guidelines for Electronic Text Encoding and Interchange* (« Recommandations pour le codage et l'échange des textes informatisés »), aussi connues sous le nom de « TEI P3 »[6, 8]. Ces *Recommandations* proposent un ensemble de conventions de codage utilisables dans une grande variété d'applications : publication électronique, analyse littéraire et historique, lexicographie, traitement automatique des langues, recherche documentaire, hypertexte, etc. Les *Recommandations* concernent les textes écrits ou parlés, sans restriction de langue, de période, de genre ou de contenu et répondent aux besoins fondamentaux de nombreux utilisateurs, lexicographes, linguistes, philologues, bibliothécaires et, de manière générale, de tout ceux qui sont concernés par l'archivage et l'accès à des documents électroniques.

2. Principes

Les règles et recommandations proposées dans les *Recommandations* sont basées sur le langage SGML *(Standard Generalized Markup Language)* qui est un standard international [7] d'un usage de plus en plus répandu. SGML est un méta-langage qui précise des règles permettant la définition de systèmes de balises pour chaque type de texte. En règle générale, les éléments du texte sont encadrés par des balises ouvrantes et fermantes, du type `<balise>... </balise>`. Ces balises peuvent contenir des attributs fournissant une description de l'élément textuel concerné et qui se placent sur la balise ouvrante :

```
<balise attribut=valeur>... </balise>
```

.

SGML permet d'associer à chaque type de texte une « Définition de Type de Document» (DTD) qui précise les balises autorisées et les agencements légaux de ces balises (voir par exemple [2, 3] et, en français, [1, 4, 9].

La DTD TEI est construite de façon modulaire :

- un jeu de balises «noyau» *(core tag set)* composé d'éléments communs à tous les types de textes (divisions, paragraphes, etc.) ;
- des ensembles de balises de base *(base tag sets)* pour chaque type particulier de texte (prose, poésie en vers, etc.) ;
- des jeux de balises additionnelles *(additional tag sets)* pour des mécanismes particuliers qui peuvent se superposer à n'importe quel type de texte (liens hypertextuels, etc.).

Les balises du noyau sont toujours présentes, mais les jeux de balises de base ou additionnelles sont « chargées » dans la DTD en fonction des besoins des utilisateurs. Un ensemble de mécanismes est aussi fourni pour permettre aux utilisateurs de rajouter leurs propres balises ou les balises existantes sans avoir à réécrire la DTD. D'une certaine manière on peut dire qu'il n'y a pas une DTD TEI, mais plutôt une famille extrêmement diverse de DTD modulaires, le commun dénominateur de toutes les instances de documents TEI étant le jeu de balises-noyau.

Le noyau comprend en particulier l' «en-tête» TEI *(TEI header)* qui doit obligatoirement figurer en début de toute instance de document TEI. Cet en-tête est le premier mécanisme systématique qui ait été développé pour la documentation en ligne des textes électroniques. Elle permet de décrire aussi bien les aspects bibliographiques habituels du document source (auteur, éditeur, édition, etc.) que ceux propres à la version électronique (aspects bibliographiques mais aussi spécificités du codage, historique des révisions, etc.). L'en-tête TEI s'est révélé de la plus grande utilité pour le catalogage des documents électroniques et leur échange.

3. Futur de la TEI

Avec la publication des *Recommandations* en 1994, la TEI est entrée dans une nouvelle phase. Des erreurs ou omissions sont possibles sur un travail de cette ampleur et il convient peut-être de rationaliser ou de simplifier certaines options de codage. Par ailleurs, des jeux de balises doivent être développés pour de nouvelles catégories de textes et des «personnalisations» pour des applications particulières sont nécessaires.

De nombreux projets ont adopté les *Recommandations* de la TEI pour le codage de leurs textes *(British National Corpus, Stockholm-Umea Corpus of modern Swedish, EAGLES, PAROLE, MULTEXT, Lingua Parallel Concordancing Project, Memoria, Oxford Text Archive, Project Runeberg, Thesaurus Linguae Latinae, Model*

Editions Partnership, Brown University Women Writers Project, Corpus de Referencia del Espanol Actual, Corpus Diacronico del Espanol, etc.). La TEI a commencé un cycle de travail collaboratif avec ces projets, de manière à valider les *Recommandations* et à proposer diverses extensions et révisions. Tous les utilisateurs de la TEI sont cordialement invités à prendre contact avec le comité de pilotage afin de participer à cet effort.

4. Information sur la TEI

– Les *Guidelines for Electronic Text Encoding and Interchange* (Les « Recommandation ») sont disponibles en version imprimée (1300 pages, 2 volumes) ou sur CD-ROM, au prix de 75$ ou 50 livres sterling, auprès de :

TEI Orders
Oxford University Computing Services
13 Banbury Road
Oxford OX2 6NN
Royaume Uni

Un bon de commande est disponible à l'adresse :
`http://www-tei.uic.edu/orgs/tei/info/p3order.html`

Les *Recommandations* sont aussi disponibles en version électronique navigable sur le World Wide Web à l'adresse :
`http://www-tei.uic.edu/orgs/tei`
ou peuvent être téléchargées par ftp anonyme aux adresses :
`ftp-tei.uic.edu (pub/tei)`
`info.ex.ac.uk (pub/SGML/TEI)`
`TEI.IPC.Chiba-u.ac.jp (TEI/P3)`
`ftp.ifi.uio.no (pub/SGML/TEI)`

– D'autres informations sur la TEI sont accessibles sur le World Wide Web à l'adresse indiquée ci-dessus. Une liste de discussion est également ouverte à tous les utilisateurs de la TEI. Pour souscrire, il convient d'envoyer un message électronique à
`listserv@uicvm.uic.edu`
avec la ligne de texte suivante :
`subscribe TEI-L Prenom Nom`
(où «Prenom» et «Nom» doivent bien sûr être remplacés de façon adéquate).

– On pourra également consulter le site de Robin Cover consacré à SGML :
`http://www.sil.org/sgml/sgml.html`

– Une version simplifiée de la TEI, dite TEI Lite, fait l'objet du présent *Cahier GUTenberg* qui sera accessible d'ici peu dans :
`http://www.univ-rennes1.fr/pub/GUTenberg/publications`
`http://distb.mesr.fr/norm/`

– Enfin, un ouvrage regroupe des articles écrits par les principaux acteurs des comités de travail de la TEI et essaie de donner le contexte et les raisons ayant abouti aux choix décrits dans les *Recommandations* : [6].

Bibliographie

[1] Jacques ANDRÉ, «Balises, structures et TEI», *Cahiers GUTenberg*, n° 24 (ce cahier), juin 1996, 11–22.

[2] Lou BURNARD, «What is SGML and how does it help?», in Ide and Véronis (1995) (Eds.) *The Text Encoding Initiative: Background and Context*, Kluwer Academic Publishers, Dordrecht, 1995, 41-50.

[3] Charles GOLDFARB, *The SGML Handbook*, Oxford University Press, 1990.

[4] Michel GOOSSENS, «Introduction pratique à SGML», *Cahiers GUTenberg*, n° 19, janvier 1995, p. 27–58.

[5] Nancy IDE, C.M. SPERBERG-MCQUEEN, *The Text Encoding Initiative: Its History, Goals, and Future Development*, in Ide and Véronis (1995) (Eds.) *The Text Encoding Initiative: Background and Context*, Kluwer Academic Publishers, Dordrecht, 1995, 5-15.

[6] Nancy IDE and Jean VÉRONIS, *The Text Encoding Initiative: Background and Context*, Kluwer Academic Publishers, Dordrecht, 1995, 242 p. ISBN : 0-7923-3689-5.

[7] International Organisation for Standardization, *Langage normalisé de balises généralisé (SGML)*, ISO 8879-1986 (F), Genève, 1986 ; voir aussi AFNOR Z 71-010.

[8] C.M. SPERBERG-MCQUEEN and Lou BURNARD, *Guidelines For Electronic Text Encoding and Interchange (TEI P3)*, ACH-ACL-ALLC Text Encoding Initiative, 1994.

[9] Eric VAN HERWIJNEN, *SGML pratique*, International Thomson publishing France, 1995, 400 p. ISBN : 2-84180-009-1.

Annexe

Traduction de la table des matières des *Recommendations pour le codage et l'échange des textes informatisés (Guidelines)* de la TEI.