
Codage, documentation et diffusion de ressources textuelles

Patrice BONHOMME, Florence BRUNESSEAU et Laurent ROMARY

*Projet Dialogue, CRIN-CNRS & INRIA-Lorraine, Bâtiment LORIA
Campus scientifique, B.P. 239, 54506 Vandœuvre-lès-Nancy Cedex
{bonhomme, brunesea, romary}@loria.fr*

Résumé. Dans la communauté des sciences humaines ainsi que dans le domaine de l'informatique linguistique, les ressources textuelles sous format électronique sont à la base de bon nombre de recherches ou de réflexions. La grande diversité des ressources, des domaines et des architectures informatiques compliquent considérablement le traitement de ces données. Cet article présente brièvement les solutions et l'aide qu'apporte la TEI, une instance de la norme SGML, et illustre par quelques exemples d'applications développées au sein de notre équipe, les traitements effectués sur des données au format TEI.

Mots clef : corpus, ressources textuelles, serveur, alignement, SGML

1. Pourquoi Utiliser la TEI?

1.1. Normalisation des Données

L'expérience montre que l'une des principales difficultés posées par la manipulation de textes sous une forme électronique est le choix du codage ou du format utilisé. De ce choix dépend, en grande partie, le succès d'une étude, d'une recherche linguistique, du stockage ou de l'échange de toutes ressources textuelles.

Échange des Données – Au moment où les autoroutes de l'information sont en plein essor, il est important de ne pas oublier la diffusion ou l'échange de ces ressources. Mis à part le Web, le support peut être aussi varié qu'une disquette, une bande magnétique ou un CD ROM. L'échange doit alors être effectué dans les meilleures conditions possibles afin de ne pas perdre d'informations ou de ne pas détériorer la structure de ces données. Le choix du

codage devra alors se porter vers un format portable d'un support à un autre, d'une architecture à une autre.

Définition d'Outils – Quelle que soit la forme sous laquelle se présente le texte, celui-ci ne sera utilisable que s'il existe des outils pour le manipuler. Un certain niveau de normalisation permettra la mise à disposition de tout utilisateur d'un ensemble d'outils standardisés (commandes spécifiques), et évitera la duplication d'outils ayant la même fonction mais sur des formats différents parfois même exotiques. Depuis quelques temps déjà, arrive sur le marché des bibliothèques d'outils¹, ou même des environnements logiciels complets supportant la totalité ou une partie seulement de la norme SGML.

Documentation des Données - Un document conforme à la TEI a l'avantage, en plus de la transcription du texte proprement dit, de contenir un en-tête. Celui-ci est sûrement le point fort de la TEI car il apporte une documentation très précise, en fournissant des indications sur la version électronique du texte, le codage utilisé, l'origine du texte source (bibliographie) et les modifications apportées au codage depuis la création de la forme électronique. Il permet également de gérer de manière rigoureuse les droits et les conditions d'utilisation de la version électronique.

TEI vs HTML - Au même titre que HTML, utilisé sur le Web par un grand nombre de personnes, la TEI est une application de la norme SGML, bref une grosse DTD! Elle contient plus de 500 éléments (50 pour HTML) et autant d'attributs. C'est dire si les formes textuelles susceptibles d'être codées à l'aide de la TEI sont nombreuses et variées comme le roman, la poésie, le théâtre, l'article scientifique ou technique mais également le dictionnaire, le dialogue multimodal (parole et geste) et bien d'autre encore. Le point crucial à retenir est que contrairement à HTML, qui a un objectif de présentation (de l'encre sur une page!), la TEI normalise l'étape de représentation des données.

1.2. Proposition de la TEI

La Text Encoding Initiative (TEI) est un projet de norme proposant des directives de codage de texte pour l'élaboration et l'échange de documents électroniques. La TEI n'a rien inventé. Elle a simplement recensé les besoins des utilisateurs, s'inspirant des normes existantes lorsqu'elle ne les a pas, tout simplement, englobées (AAP, ISO12083, HyTime, MARC, ...). La richesse de la TEI offre trois niveaux d'utilisation des balises : obligatoires, conseillées et optionnelles. Il ne reste plus qu'à l'utilisateur à définir ses besoins en fonction de ses propres applications.

1. <http://www.falch.no/~pepper/SGML-Tools.html>, un large éventail d'outils SGML

Pour plus d'information, le lecteur intéressé peut se référer à l'ouvrage coordonné par Jean Véronis et Nancy Ide, *Text Encoding Initiative: Background and Context*, Kluwer Academic Publishers, 1995 et aux divers articles de ce *Cahier GUTenberg*.

2. Exemples d'Applications

Spécifique : L'Alignement Multilingue - Le principe de l'alignement multilingue est de mettre en parallèle des portions assez fines de deux textes (source et cible), l'un étant la traduction de l'autre, en utilisant comme principal critère le nombre de caractères de ces portions. L'algorithme que nous avons développé dans notre équipe² repose sur un découpage en chapitres, paragraphes, phrases et utilise la TEI pour repérer cette structure (balises <div>, <p> et <s>). Mais le calcul de cet alignement pouvant être long sur de gros textes, nous utilisons également la TEI pour stocker les alignements sous la forme de balises de type <xptr> et <link> à l'intérieur du texte source.

Service : Le Serveur Silfide - Le serveur Silfide³ est un outil de mise en commun, convivial et raisonné, des connaissances sur différents aspects de la langue française. Il s'agit d'une plate-forme de ressources textuelles, garantissant les conditions d'accès et d'utilisation défini par les fournisseurs et pouvant répondre aux interrogations des utilisateurs telles que, lieux et conditions d'accès aux données, format, degré de validation et outils disponibles. Quel que soit le format d'origine des ressources, la TEI – son en-tête et ses possibilités de codage – s'est avéré répondre aux besoins du serveur pour la documentation, l'archivage, l'interrogation et la distribution des données.

Intégration : L'Environnement XCorpus Afin de faciliter l'utilisation de ces applications spécifiques, nous développons un environnement graphique XCorpus⁴ dédié à la manipulation de corpus textuels représentés sous forme SGML, en respectant les directives générales de codage proposées par la TEI. Outre l'importation des textes de différentes origines vers un format TEI, XCorpus permet l'édition et la gestion de corpus mono et multilingues afin d'effectuer une série d'opérations, alignements (figure 1), calculs statistiques, lexique, concordances, mais également la génération automatiquement d'une navigation multilingue pour le Web à base de pages HTML en important les éléments

2. Dans le cadre d'un projet Européen Lingua, <http://www.loria.fr/~bonhomme/lingua>

3. Serveur Interactif pour la Langue Française, son Identité, sa Diffusion et son Étude – projet sous l'égide de l'Aupelf/Uref et du CNRS - <http://www.loria.fr/Projet/Silfide>

4. <http://www.loria.fr/Projet/XCorpus>

textuels appropriés sans avoir à s'occuper de la langue effective dans laquelle sera lu le document.

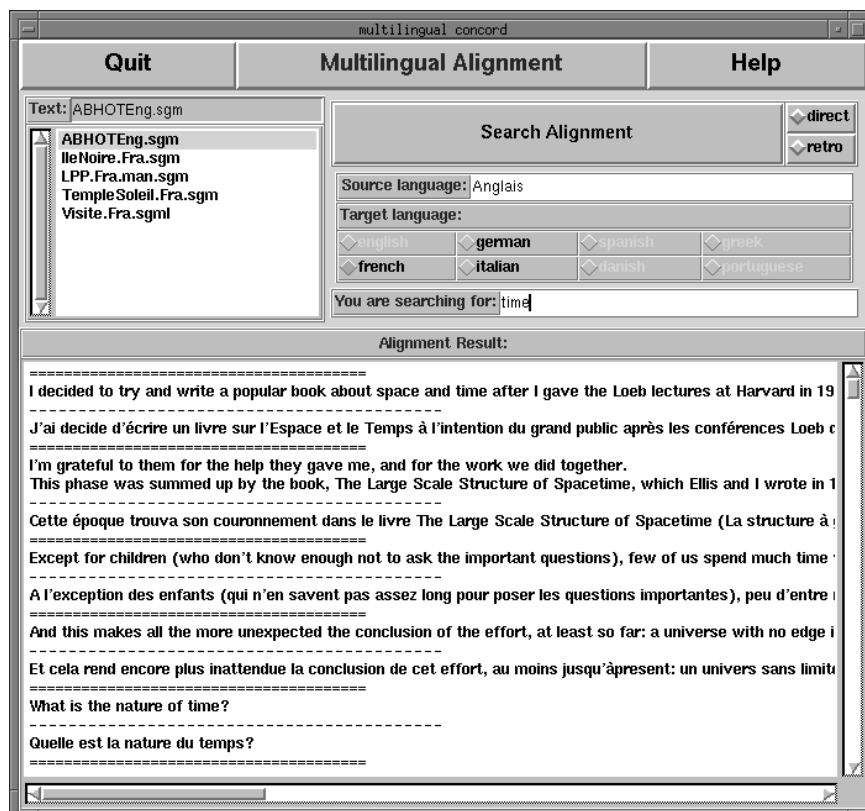


FIGURE 1 – Exemple d'alignement Français-Anglais avec XCorpus

L'utilisation faite de SGML par la TEI est ambitieuse mais ne diffère en rien de toute autre application de SGML, et par conséquent un logiciel SGML sera capable de traiter un texte conforme à la TEI quel qu'il soit.