Du document électronique à son usage : le rôle central de la normalisation

Jean-Luc Benoit, Charles Bernet, Patrice Bonhomme, Laurent Romary, Nadia Viscogliosi

Du document électronique à son usage : le rôle central de la normalisation

J.-L. Benoit*, Ch. Bernet*, P. Bonhomme, L. Romary**, N.Viscogliosi****

*INaLF (Institut National de la Langue Française)
Institut National de la Langue Française
44, av. de la Libération
B.P.30687-54063 NANCY CEDEX
Tél. (33) 03 83 21 76
http://www.inalf.cnrs.fr/
email : jean-luc.benoit@inalf.cnrs.fr
email : charles.bernet@inalf.cnrs.fr

**LORIA (Laboratoire Lorrain de Recherche en Informatique et ses Applications)
Campus Scientifique - BP 239
54506 VANDOEUVRE-LES-NANCY CEDEX
Tél. : (33) 03 83 59 20 37
http://www.loria.fr
email : patrice.bonhomme@loria.fr
email : laurent.romary@loria.fr
email : nadia.viscogliosi@loria.fr

1 Introduction

L'objectif de cet article est de mettre en évidence, au travers d'expériences concrètes, le rôle essentiel d'une démarche de normalisation lorsque l'on a à gérer des fonds textuels informatisés. Il s'agit pour nous de montrer qu'il est possible d'adopter une politique éditoriale rationnelle qui permette à la fois de prendre en compte les patrimoines existants et d'en assurer une certaine pérennité par le biais de pratiques reproductibles et documentées. Il faut être capable, lorsque l'on manipule des textes électroniques, de tenir compte à la fois du passé et de l'avenir. En effet, il existe de nombreuses sources de documents électroniques qui ont adopté des modes de représentation très différents et souvent incompatibles (format propre à un traitement de texte, HTML, codage propre à une institution, etc.); il faut espérer que l'on ne tombe pas dans les mêmes travers méthodologiques et que l'on soit véritablement en mesure de transmettre un patrimoine informatisé réutilisable par d'autres.

1.1 Principes généraux

Au centre d'une telle démarche éditoriale se trouve bien évidemment la notion de texte, ou plus précisément de document, qui englobe le texte lui-même ainsi qu'un ensemble d'informations que l'on souhaite pouvoir lui attacher. De fait, on ne peut mettre en œuvre une base de documents textuels sans s'être posé au préalable un certain nombre de questions relatives aux objectifs d'un tel projet.

Présentation ou représentation — Il est important de dissocier, lorsque l'on passe d'un format imprimé à une version électronique, les informations d'ordre purement typographique de celles qui renseignent sur l'organisation logique du texte ou l'identification d'éléments particuliers. Identifier une suite de mots comme étant en italiques dans le texte original ne permet pas de faire une différence entre une emphase, une expression d'origine étrangère ou, dans le cas du théâtre, une indication scénique. En fonction de ce que l'on souhaite effectivement garder de l'édition d'origine, on peut vouloir ajouter quelques marques typographiques, mais il est clair qu'il faut disposer d'un moyen d'exprimer des informations relevant du contenu.

Indépendance des données vis-à-vis du logiciel — Pérenniser un document électronique, c'est entre autres s'affranchir des contraintes liées à un logiciel particulier qui imposerait un format propriétaire de représentation des données. De tels formats sont trop sujets à évolution et le risque est grand (l'expérience le prouve) de ne plus pouvoir accéder à une information parce que le logiciel qui a servi à la créer n'est plus disponible. Par ailleurs, il est indispensable que tout projet académique soit capable de s'appuyer sur une offre logicielle variée comprenant en particulier des logiciels dits libres.

Comparer et transmettre des documents — L'indépendance vis-à-vis du logiciel doit s'accompagner d'une véritable démarche de normalisation pour réellement harmoniser les pratiques de gestion de ressources textuelles entre les différentes institutions ou laboratoires concernés. L'objectif est d'aboutir à un traitement *uniforme* et *cohérent* de textes relevant du même domaine ou appartenant au même genre. Il s'agit en effet, d'une part, d'être capable de comparer des données d'origines différentes et de façon duale de pouvoir échanger ces données avec d'autres sites en utilisant un protocole "aveugle", c'est-à-dire ne nécessitant pas une connaissance préalable du codage employé.

Des données qui se suffisent à elles-mêmes — On arrive ainsi à envisager un mode de gestion des documents textuels qui repose sur un codage normalisé, issu si possible de pratiques normatives existantes, et surtout qui permette à un document électronique d'être parfaitement autonome, tant du point de vue de la plate-forme logicielle qui va l'accueillir que du point de vue de l'utilisateur qui doit se l'approprier pour un usage peut-être différent de celui qui a conduit à sa création. D'un point de vue théorique, on est proche des notions de document semi-structuré (Buneman et alii, 1996 ; Abiteboul, 1997 ; Abiteboul et alii, 1997) qui, par opposition aux modèles classiques de bases de données, ne nécessitent pas de connaissance préalable d'un schéma abstrait décrivant leur organisation. Ce sont exactement ces notions qui ont d'abord conduit à la mise en œuvre de la norme SGML dans les années 80 puis à la définition simplifiée de son cousin XML dans le cadre du développement de ces techniques au sein du réseau Internet. La cohérence entre les contraintes que nous nous fixons et les perspectives technologiques proposées par de tels standards expliquent les choix que nous avons pu faire dans le cadre des travaux présentés dans cet article.

1.2 Cadre de ce travail

La réflexion et les applications présentées ici sont le fruit d'une collaboration de plusieurs mois entre le *Service des Bases Textuelles* de l'INaLF et l'équipe *Langue et Dialogue* du LORIA visant à mettre en commun une expertise philologique et technique sur le texte et son informatisation. Fruit des expériences acquises au sein du projet Silfide[1] (cf. Romary et alii, 1999; Bonhomme et alii, 1998), cette collaboration a pour objectifs :

- d'unifier les pratiques de codage entre nos équipes et de diffuser ces pratiques en direction d'autres institutions ou laboratoires qui se poseraient des problèmes similaires;
- de valider ces pratiques au travers de textes de références, issus notamment des anciens fonds informatisés de l'INaLF, pour en réaliser des éditions informatiques nettoyées et balisées selon de nouvelles règles;
- d'assurer la diffusion des résultats par le biais, d'une part, de la plate-forme Silfide pour des accès en ligne, et d'autre part sous des formats directement accessibles sur des supports électroniques "traditionnels".

2 Codage d'un document textuel : éléments éditoriaux et techniques

2.1 Les informations à prendre en compte

La mise en œuvre d'un projet éditorial lié au codage de documents textuels informatisés requiert une analyse précise des éléments d'information à prendre en compte. On ne peut en effet se limiter à la simple transcription, sous la forme d'une suite de caractères ASCII, des textes d'origine. De fait, on peut dégager quatre grandes classes de données susceptibles d'être représentées dans un document textuel :

a) en premier lieu, il faut pouvoir associer à tout texte informatisé, des *informations documentaires* précises qui renseignent à la fois sur l'origine du contenu textuel (auteur, titre, édition de référence), mais aussi sur le contenu électronique proprement dit (responsabilité de

---

[1] Silfide (**S**erveur **I**nteractif pour la **L**angue **F**rançaise, son **I**dentité, sa **D**iffusion et son **E**tude) est un projet du CNRS et de l'AUPELF-UREF. Son objectif principal est de permettre l'accès d'une manière conviviale et raisonnée à des ressources textuelles (quelle que soit leur origine, écrite ou orale) à l'ensemble de la communauté universitaire travaillant à partir de la langue (linguistes, enseignants, informaticiens,...) à travers un réseau de serveurs informatiques et d'actions en alimentant les fonctions. http://www.loria.fr/projets/Silfide.

l'édition électronique, conditions de distribution du document, niveaux et choix de codage, etc.);

b) il est parfois utile, on le verra dans le cas du théâtre, d'associer des *éléments connexes* au texte lui-même, tels que préfaces, introductions, etc. En fonction de l'usage que l'on veut faire du document informatisé, il faudra se demander si l'on garde des éléments tels que les index ou les tables des matières qui sont soit faciles à reconstituer, soit redondants par rapport à des modes d'accès plus élaborés qu'autorise le texte informatisé;

c) l'une des étapes essentielles dans l'informatisation d'un texte reste malgré tout l'identification de sa *structure*. Celle-ci peut être considérée à différents niveaux : il y a bien sûr le découpage en grandes divisions et sous-divisions qui introduit une hiérarchisation du texte ; il y a le niveau du paragraphe qui, en fonction du genre, peut s'exprimer sous la forme de strophes, de tours de parole etc. Enfin, il faut repérer les marques plus fines des ruptures de lignes quand celles-ci sont significatives, soit explicitement dans un poème, soit de façon plus informelle (verset claudelien, par exemple). A ce stade, notons que selon l'édition de référence, ou simplement à des fins de vérification, on pourra garder certaines informations liées à la structure physique du texte, comme les marques de rupture de page;

d) enfin, on peut mettre en évidence une quatrième classe d'informations, plus fluctuante, constituée d'éléments qui ne sont pas nécessairement explicites dans le texte d'origine. L'identification ("annotation") de ces éléments peut néanmoins faciliter la gestion et surtout l'accès au texte informatisé. Suivant la perspective éditoriale adoptée, mais également l'usage qui sera fait du document, on pourra ainsi marquer la présence de noms propres (voir par exemple Bruneseaux, 1998), d'abréviations, d'expressions d'origine étrangère, etc. Même si, dans certains cas, ces informations sont repérées par des éléments typographiques dans le texte source, le travail d'annotation reposera toujours sur une certaine subjectivité de celui qui effectue l'opération.

2.2 Quel cadre adopter pour baliser un texte

2.2.1 Les premiers pas

Comme on l'aura compris, nous défendons dans cet article une approche qui vise à associer toute entreprise de numérisation à une démarche de normalisation, pour d'une part ne pas refaire le travail effectué par d'autres au préalable, et surtout assurer un bon niveau de survie aux documents

sur lesquels nous travaillons. Il existe ainsi depuis plus de 10 ans une norme, appelée SGML[2], largement utilisée dans les milieux de l'édition (notamment scientifique) et de la gestion documentaire d'entreprise. Cette norme a été pendant plusieurs années la seule référence dans le domaine de la représentation des documents électroniques, et les grandes sociétés savantes du domaine de l'informatique pour les sciences humaines ont décidé de s'y conformer pour organiser les propositions de la Text Encoding Initiative (TEI). Avant de présenter cette initiative (cf. infra), il faut signaler que dans sa toute généralité, SGML est une norme lourde et complexe à mettre en œuvre, ce qui a fortement limité sa diffusion et son usage dans les milieux académiques.

2.2.2 XML[3], le « chaînon manquant » ?

"Chaînon manquant", "pierre de Rosette", les métaphores sont souvent éloquentes pour parler de XML. Mais il s'agit d'une éloquence de bon aloi tant on pressent que l'on tient là le langage de normalisation. En effet, la mise en œuvre de la norme SGML d'un point de vue éditorial est souvent lourde et complexe à réaliser pour l'utilisateur de base. La recommandation XML, qui est un sous-ensemble de la norme SGML, offre les mêmes possibilités que la norme SGML sans sa complexité pour le codage de données structurées. XML a été conçu pour mettre à la disposition des développeurs un environnement structuré permettant la création de DTD[4] convenant à tous types d'informations.

XML ouvre en fait les portes de l'Internet aux documents structurées et comble ainsi les lacunes et les déviances des applications purement basées sur le langage HTML. XML, qui fournit une méthode uniforme pour la description et l'échange de l'information structurée sur le Web, établit le pont entre la richesse et la complexité de SGML et la pauvreté sémantique d'HTML.

Parmi les différences notables entre XML et SGML, la principale concerne les notions de document valide et de document bien formé. En effet, un document XML peut ne pas contenir de déclaration de type de document (DTD). Nous parlerons dans ce cas de *document bien formé,* à la différence des *documents valides* qui eux, tout comme avec SGML, possèdent une DTD et en respectent le schéma de codage. L'avantage principal de la notion de document bien formé est la possibilité d'extraire des fragments de document et de les traiter ou de les échanger sans leur attribuer de DTD. Pour résumer, on édite en valide et on transmet en bien formé.

---

[2] SGML (**S**tandard **G**eneralized **M**arkup **L**angage), le grand-père de tous les langages balisés, a été créé en 1960 sous l'impulsion d'IBM pour répondre aux problèmes liés au portage des documents d'une plate-forme logicielle ou matérielle vers une autre. En 1986 ce langage devint un standard officiel (ISO 8879:1986).

[3] XML : e**X**tended **M**arkup **L**anguage. http://www.w3.org/TR/REC

[4] DTD : **D**ocument **T**ype **D**efinition

2.2.3 De la représentation avec la TEI

XML étant uniquement un *méta-langage*, il est important de fonder la normalisation de ses données sur des pratiques partagées avec sa propre communauté. Dans le cadre de la collaboration entre le LORIA et l'INaLF, comme d'ailleurs dans les autres projets de l'équipe Langue et Dialogue, le choix s'est porté sur l'utilisation des directives de codage de la TEI comme format normalisé de codage et d'échange. Les textes ainsi codés et enrichis sont rendus disponibles, par exemple, sur le serveur Silfide.

La TEI est un projet international d'un système de codage commun des textes, en SGML, qu'il était devenu urgent de mettre au point dans la diversité chaotique des systèmes utilisés dans les années 80. Le projet, lancé en 1987, aboutit à un système de conventions qui répond aux besoins fondamentaux de la plupart des projets d'encodage de textes, quelles qu'en soient la nature et la taille. En 1994, paraît un guide d'encodage : le *TEI Guidelines for Electonic Text Encoding and Interchange* [5] qui est une aide détaillée et régulièrement mise à jour. Le consortium TEI, nouvellement créé, s'est penché sur la mise en conformité des directives avec la recommandation XML.

2.2.4 De la présentation avec XSL[6]

XSL est une application de XML permettant de définir des *feuilles de style* grâce auxquelles tout document XML pourra être mis en forme, en vue d'être affiché et/ou imprimé, ou bien transformé en un autre document XML. Ainsi, pour un même document de départ, on peut élaborer plusieurs feuilles de style pour obtenir plusieurs résultats à l'arrivée : afficher le document en entier ou en partie, réorganiser le contenu du document, etc.

De même, une seule feuille de style pourra être appliquée à l'ensemble des documents d'un corpus respectant les mêmes règles de codage. Afin, par exemple, de modifier l'apparence à l'affichage de l'ensemble des documents, il suffit de modifier cette seule feuille de style.

XSL se compose de deux langages :

- XSLT, un puissant langage de transformation pour sélectionner et/ou réorganiser l'information en fonction des besoins. Cette transformation s'appuie sur la représentation hiérarchique des documents XML, les noms des éléments, les noms et valeurs des attributs ainsi que sur le contenu lui même.
- XSL FO, un langage autorisant la création d'objets de formatage et de description de leurs propriétés. Ce langage permet par exemple de créer des objets de type paragraphe, séquence,

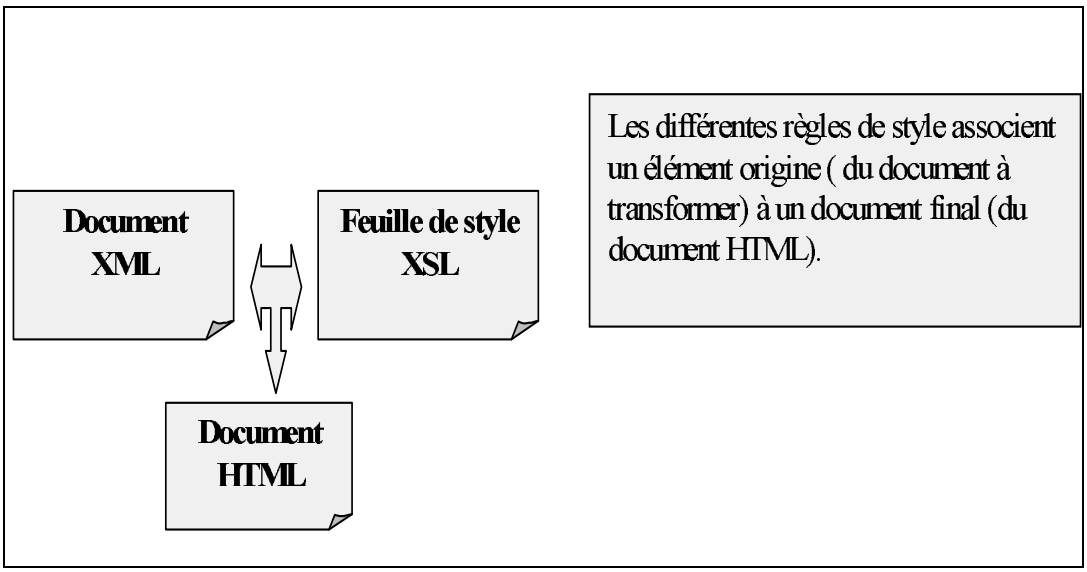
---

[5] http://etext.virginia.edu/TEI.html

[6] XSL : e**X**tensible **S**tylesheet **L**anguage

tableau, cellule, image etc. et de leur attribuer des propriétés d'affichage telles que l'espace avant et après, la fonte, les couleurs, etc. Ce langage est totalement indépendant du support de sortie.

Un tel mécanisme permet de conserver toute la richesse sémantique du codage initial dans un seul fichier, dont on peut extraire, selon les besoins, de nombreux documents différents.



3 Un texte bien documenté : l'importance de l'en-tête

Un texte électronique de qualité, bien saisi, bien codé, enrichi d'informations sémantiques, linguistiques ou autres, peut perdre considérablement de son intérêt et de sa valeur s'il n'est pas, par ailleurs, décrit correctement et clairement situé parmi un ensemble de documents. Ainsi par exemple, de nombreuses œuvres littéraires sont accessibles en texte intégral sur Internet mais bien peu citent une édition de référence, ce qui est un obstacle à leur utilisation dans la Recherche.

S'il va de soi qu'un document comporte au minimum un titre, le plus souvent un ou plusieurs auteurs, voire un éditeur, une date de parution etc., il peut être utile, sinon nécessaire, de rechercher d'autres informations parfois implicites, n'appartenant pas au texte lui-même. Par exemple, des requêtes telles que "la littérature féminine" ou "*telle* édition de *tel* document" ne peuvent aboutir si on ne dispose pas d'indications aussi simples que le sexe de l'auteur et la source précise du texte.

Toutes ces informations sont recueillies dans un en-tête, ou *header*, qui précède le texte proprement dit. Selon le type de document considéré et l'objectif visé lors de la collecte et du codage, elles peuvent représenter une grande diversité qu'il devient important d'organiser.

3.1 L'en-tête TEI

La TEI rend très bien compte de cette diversité et permet avec une grande souplesse d'élaborer des en-têtes pour un grand nombre de types de documents.

Constitué de quatre parties essentielles, cet en-tête permet de décrire en profondeur :

- le fichier informatique lui-même, contenant le texte et l'en-tête (avec la possibilité d'indiquer les responsabilités attachées à ce fichier, sa taille, les modalités de sa constitution et de sa diffusion, la ou les sources utilisées,…),
- les règles de codage et les choix éditoriaux appliqués au contenu (niveau de balisage, balises utilisées, corrections, etc.),
- les caractéristiques du texte codé : mots-clés, nature du texte (oral vs écrit, original vs traduction, genre littéraire), contexte, lieux et personnes impliqués, ...,
- l'historique du fichier, permettant de suivre les mises à jour, ajouts et corrections successives apportées au fichier.

3.2 L'en-tête Silfide

Les possibilités proposées par la TEI pour structurer ces données, nous l'avons vu, sont très étendues. Il est facile de concevoir qu'à partir d'un même document et des mêmes objectifs d'encodage, le résultat pourra être différent selon les utilisateurs. Il nous a donc semblé nécessaire dans le cadre de notre collaboration et, à terme, pour envisager de nouveaux projets, de définir un *en-tête type* adapté à la nature des documents considérés. En effet, la manipulation de ces documents (en particulier leur affichage et l'application d'outils de recherche) implique la présence obligatoire de certains éléments et le respect d'une sémantique rigoureuse de ces éléments. De même, l'application ultérieure d'une feuille de style implique une régularité dans la description des données.

Les exemples suivants, tirés d'un corpus réel[7], illustrent la "dérive" de l'utilisation de balises, par ailleurs conforme à la TEI, provenant de l'interprétation du sens de ces balises par différents membres d'un même projet. La balise <title> contient à la fois le titre, la langue et l'auteur dans un ordre et une syntaxe variable.

Exemples :
<title>the Republic an electronic version (in Bulgarian) </title>
<author>Plato</author>

[7] projet TELRI : alignement d'un corpus multilingue (18 langues) de "La République" de Platon.

<title>Plato's Republic - New Translation</title>

<title>Plato's Republic in German</title>

De ce constat et d'une réflexion poussée sur les éléments pertinents à prendre en compte pour décrire un document est né l'*en-tête Silfide minimal*. Il s'agissait au départ d'élaborer un modèle d'en-tête applicable à l'ensemble du fonds Silfide et contenant les informations nécessaires et suffisantes, structurées, permettant d'exploiter au mieux les ressources décrites. La difficulté pour constituer un tel en-tête générique résidait dans la résolution du paradoxe suivant :

- être, justement, suffisamment générique pour suffire à décrire la diversité des documents présents dans la base Silfide, qu'il s'agisse de romans, de textes législatifs, d'articles de presse, de bande dessinée, de transcription d'oral, etc.
- permettre de rendre compte de la spécificité de ces documents (textes écrits, transcription d'oral, genres littéraires, traductions, etc.).

Un soin particulier a été apporté à la gestion des langues (langues du texte, du titre, des en-têtes, des versions originales,…) dans la mesure où une part des développements autour de Silfide, serveur d'abord francophone, concerne l'alignement de textes multilingues.

Le travail d'harmonisation a porté aussi bien sur le type d'informations que sur leur syntaxe. En effet, les recommandations de la TEI n'imposent aucune contrainte sur le contenu des balises; or la forme de l'information peut être aussi importante que l'information elle-même, comme le montrent les exemples suivants :

Exemples :
<name>Paul Claudel</name>
<name>Claudel, Paul</name>
<name>Claudel P.</name>

Dans ces trois cas, conformes à la TEI, la variation de la syntaxe pourrait empêcher, par exemple, la constitution d'un index de noms d'auteurs automatique.

Un *manuel d'élaboration d'en-tête Silfide* intégrant tous ces aspects, et dont une version de travail est d'ores et déjà accessible[8], est actuellement en cours de rédaction.

[8] http://www.loria.fr/projets/Silfide/informations/Docs/Divers/header.html

4 L'exemple des textes de théâtre

4.1 Les pièces de théâtre : un modèle de structuration

Le théâtre est, parmi les genres littéraires, l'un de ceux qui gagne le plus à être balisé en raison, d'une part, de l'agencement de la plupart des pièces en structures à plusieurs niveaux et, d'autre part, de l'imbrication de types de discours différents dans le corps du texte.

Mais avant tout, les éditions des pièces de théâtre peuvent, comme toutes les autres publications, comporter des textes liminaires ou annexes qui doivent être marqués comme tels pour les distinguer du texte principal. En effet, il est fréquent de trouver, selon les époques ou selon les auteurs, un avant-propos, une dédicace, une épître dédicatoire, une introduction, une préface ou une postface. Certains textes annexes sont souvent présents dans les éditions des pièces de théâtre. Ce sont par exemple l'*argument,* qui expose brièvement le sujet et l'action de la pièce, parfois remplacé par un *avertissement* lorsque l'auteur ne se borne pas à parler du sujet qu'il traite, mais de sa manière de le traiter. Ces textes accessoires ont souvent une importance reconnue par l'histoire littéraire et doivent absolument figurer dans les saisies électroniques des pièces auxquelles ils sont associés. C'est le cas notamment des préfaces de certaines pièces classiques, dans lesquelles les dramaturges se situent par rapport à leurs sources, par rapport à leurs rivaux et donnent des précisions sur la réception de leurs œuvres par le public (cf. les deux préfaces successives de *Britannicus* de Racine) ou encore des *examens* des pièces de Corneille, écrits pendant une période de retraite, dans lesquels le dramaturge livre ses réflexions critiques et doctrinales. Les *examens* sont généralement donnés, dans les éditions critiques de l'œuvre de Corneille, non pas à leur place selon la chronologie (1660-1662) mais en association avec chacune des pièces concernées.

Dans le texte des pièces proprement dit, le balisage normalisé est le moyen d'accéder de manière sélective à différents types de données ou à différents niveaux de structure.

Le texte des pièces de théâtre est, par nature, composite. Il comporte, hormis le texte de scène proprement dit, une liste des personnages, des indications de structure et des indications scéniques, c'est-à-dire les *didascalies* (Ubersfeld, 1996).

Dès qu'elles atteignent une certaine longueur, les pièces de théâtre sont structurées. La forme la plus fréquente est, à l'image des pièces classiques, un découpage en *actes* et en *scènes*, mais on rencontre aussi des pièces articulées en *tableaux* (l'acte II de *Rhinocéros* d'Eugène Ionesco est découpé en deux tableaux) ou en *parties*, par exemple dans l'œuvre de Paul Claudel, elles-mêmes découpées en scènes. Un découpage en *journées* comme dans *le Soulier de satin* est exceptionnel.

Ces structures plus ou moins régulières sont parfois précédées d'un *prologue* (*La Toison d'Or* de Corneille, *Amphitryon* et *le Malade imaginaire* de Molière, *Esther* de Racine) ou suivies d'un *épilogue* (*le Soulier de Satin* de P. Claudel), ceux-ci appartenant de plein droit au texte scénique.

Les prologues du théâtre classique constituent généralement une pièce avant la pièce avec des rôles distincts de ceux de l'action principale.

Entre les actes, il arrive que se glissent des *intermèdes,* des *ballets* (par exemple les intermèdes et les ballets du *Bourgeois gentilhomme* et les entrées de ballets du *Malade imaginaire* de Molière).

Le niveau ultime de la structuration du texte de scène est la *réplique*, qui est nécessairement associée à un personnage de la pièce.

Les dialogues peuvent prendre différentes formes : on peut identifier les tirades (suites de paroles plus longues) en les opposant aux stichomythies (courtes répliques de même longueur). Les monologues ne sont pas toujours faciles à définir. D'une manière abusive, ils désignent les scènes dans lesquels le personnage parle seul. Mais le personnage (Oreste, dans la dernière scène d'*Andromaque*) ne l'est pas toujours. De plus, peut-on parler de monologue lorsqu'un personnage s'adresse à un autre qui reste muet ? Cas particuliers des monologues, les *apartés* dont le commencement est noté par une didascalie et, dans le théâtre ancien, les stances, qui sont distribuées en strophes, comme un poème lyrique.

Le balisage peut donc restituer chacun de ces niveaux, ainsi encore que les *parties chantées*, notamment dans les chœurs de certaines tragédies, telles les tragédies bibliques de Racine ou dans les intermèdes de certaines comédies, comme *le Bourgeois gentilhomme* déjà mentionné. Lorsque les pièces sont en vers – et c'est fréquemment le cas dans le théâtre français jusqu'au 19^{e} siècle – un balisage normalisé des fins de vers ou des associations de mots (ou de phonèmes) à la rime ouvre la possibilité de mener des études de poétique à l'aide de l'édition électronique des textes saisis.

4.2 Le corpus de travail

Un corpus de 400 pièces, rassemblé par un groupe de travail à l'INaLF sous le nom de "théâtre français", avait fait l'objet d'un premier codage selon des règles internes[9] dans le but de développer des outils d'interrogation. Ainsi les divisions en actes et scènes, la structure en vers étaient indiquées, tout comme l'alternance entre les prises de paroles, l'identité des locuteurs, les didascalies, les vers coupés entre plusieurs locuteurs et les avatars du texte.

Cependant, la perspective d'échanger et de développer ces données, impossible avec un format interne, induisait le recours à un standard. Un travail de *rétroconversion* vers le format TEI a donc été entrepris, largement facilité par la similitude entre les deux approches. Seules quelques balises

[9] Ces règles initiales ont été établies dans le cadre d'un groupe de travail à l'INaLF avec la collaboration de M. Chauvet (théâtre de la période préclassique), J. Dendien et J.-Y. Hamon (informatique) et Fr. Surdel (bibliographie).

qui avaient pour but de donner une description physique du texte n'ont pas été conservées. À titre d'exemple la balise , considérée comme purement éditoriale, a été supprimée.

Lorsque l'on définit un langage de balisage comme un langage capable de décrire la structure et le contenu d'un document, cela ne signifie pas – est-il besoin de le rappeler ? – qu'il établisse une quelconque séparation entre forme et fond, entre signifié et signifiant. Lorsque Claudel utilise des polices de caractères différentes pour décrire l'affiche de recrutement des caravelles dans *Le Livre de Christophe Colomb*, ce n'est pas anodin et il faut le mentionner (coll. de la Pléiade, Théâtre, t.2, p.1159). Le balisage d'un vers ne saurait ignorer la place de la césure et passer sous silence un rejet ou un enjambement. La place de chaque mot n'y est pas interchangeable. Au baliseur d'isoler les formes signifiantes et d'y appliquer le balisage convenable : la TEI possède un arsenal de balises capable de faire face.

Le travail de balisage se rapproche donc plus de la lecture que d'un travail éditorial mécanique. Baliser un texte, c'est d'abord bien en connaître la structure et bien le comprendre, pour en faciliter les lectures ultérieures.

4.3 Cromwell *à l'épreuve de la TEI*

4.3.1 *Cromwell* de Balzac

L'année1999 a marqué le bicentenaire de la naissance de Balzac. A cette occasion nous avons choisi, dans le corpus "théâtre français", de travailler sur cet auteur en nous attachant particulièrement à un genre qui ne resta guère attaché à son nom dans l'histoire littéraire : le théâtre.

4.3.2 La TEI et le théâtre

Le *TEI Guidelines for Electonic Text Encoding and Interchange* consacre un chapitre à l'encodage du théâtre (cf. Base Tag Set for Drama, chap. 10)[10] dans lequel on a puisé tous les éléments nécessaires au balisage de notre pièce. Les valeurs des attributs, non contraintes, permettent d'apporter des précisions sur l'utilisation des balises et de s'adapter à la réalité du texte.

Un en-tête de théâtre - En plus des indications classiques de catalogage (titre, édition considérée, date, …) une pièce de théâtre possède des caractéristiques qui lui sont propres et que l'on peut souhaiter faire apparaître, lorsque l'information est disponible. On ne saurait passer sous silence, par exemple, des éléments bibliographiques tels que la date et le lieu de la première représentation de la pièce, les acteurs ayant créé la pièce, etc. D'une façon plus générale, on fera figurer les mentions des réécritures, coupures et ajouts, ainsi que les reprises de la pièce jugées notables.

L'avant-texte - L'élément <front> contient toutes les parties qui précèdent le texte de la pièce elle-même, telles qu'elles apparaissent dans la version imprimée : prologue, introduction, dédicaces, préfaces…

On y trouvera en particulier : la description de la page de titre, la liste des personnages, des indications de temps et de lieu. A titre d'exemples, nous détaillons ici le codage de la liste des personnages, qui illustre différents mécanismes de la TEI.

- **La liste des personnages : <castList>**

<castList> contient la liste des personnages ou des groupes de personnages (y compris muets), parfois le lien qui les unit entre eux, éventuellement le nom de l'acteur qui a tenu le rôle à la première de la pièce.

On déclare ici le nom des personnages auxquels on attribue un identifiant ("id") qui permet de les repérer tout au long de la pièce.

Ex :

```
< castList>
<head>PERSONNAGES</head>

<castItem>
        <role id="CH"><name>CHARLES Ier</name></role>
        <roleDesc>roi d'Angleterre</roleDesc>
</castItem>
 [...]
<castItem>
        <role id="CR"><name>CROMWELL</name></role>
</castItem>
[...]
<castGroup>
<head rend="accolade">principaux amis de Cromwell. Personnages muets.</head>
        <castItem><role><name>FLEETVOLD</name></role></castItem>
        <castItem><role><name>BARCLAY</name></role></castItem>
        [...]
</castGroup>
[...]
<castItem>
```

---

[10] http://etext.lib.virginia.edu/tei-tocs3.html

<roleDesc>TOUS LES MEMBRES DU PARLEMENT</roleDesc>

</castItem>

</castlist>

Le texte de la pièce proprement dit (le corps du texte) - Dans l'encodage du théâtre, sont caractéristiques dans le corps du texte (<body>) : la division du texte en actes et scènes, les didascalies, les dialogues.

- **Le codage des divisions**

La même balise est utilisée pour coder tous les niveaux, qui sont distingués grâce à l'utilisation des attributs.

Ex :

* *

<head>ACTE PREMIER</head>

* *

<head>SCÈNE PREMIÈRE</head>

[…]

* *

[…]

* *

- **Les indications de jeux de scène, ou "didascalies"**

Les didascalies, uniformément codées par l'élément <stage>, recouvrent des cas de figure différents qui sont clairement distingués par les valeurs de l'attribut "type".

Ainsi, <stage> sera utilisé :

- au début d'une scène, pour identifier les interlocuteurs en présence (type="personnage").

Ex :

* *

<head>SCÈNE PREMIÈRE</head>

<stage type="personnage">LA REINE, STRAFFORD</stage>

[…]

* *

- pour indiquer un mouvement, un élément de mise en scène (type="mouvement") :

Ex :

[…]

<stage type="mouvement">Elle s'assied.</stage>

[…]

- pour préciser, dans une prise de parole, une situation particulière du locuteur, un aparté, etc. (type="phatique") :

Ex :

[…]

<speaker>CROMWELL</speaker><stage type="phatique">, aux conjurés.</stage>

[…]

<speaker>LE ROI</speaker><stage type="phatique">, seul.</stage>

[…]

Interlocuteurs et prises de paroles - <sp> et <speaker> permettent de baliser les interventions des personnages et les tours de parole.

Les vers, puisqu'il s'agit d'une pièce en vers, sont codés et numérotés grâce à l'élément <l>.

Exemple (extrait de l'Acte I, scène II) :

* *

<head>SCENE II</head>

<stage>IRETON, CROMWELL, STRAFFORD</stage>

[…]

<sp who="CR">

<speaker>CROMWELL</speaker>

<l n="201">Vous me semblez surpris, Seigneur, par ma présence?</l>

<l n="202">Mes efforts cependant servent votre vengeance.</l>

<l n="203">Quel était l'entretien qu'a troublé mon abord?</l>

</sp>

<sp who="ST">

<speaker>STRAFFORD</speaker>

<l n="204">Voici la liberté que vous vantez si fort?</l>

<l n="205">Bientôt l'on ne pourra dans toute l'Angleterre</l>

<l n="206">Sans l'ordre de Cromwell, ou parler ou se taire;</l>

<l n="207">L'amour de la vengeance est peu fait pour mon cœur,</l>

<l n="208">Je prétends, aujourd'hui, vous le prouver, Seigneur.</l>

</sp>

[…]

<stage>Il sort.</stage>

* *

L'attribut "who" de l'élément <sp> prend la valeur de l'identifiant, tel qu'il a été déclaré dans l'en-tête (cf. supra), du personnage qui s'exprime; ceci est fort utile lorsqu'un personnage apparaît sous plusieurs désignations.

Exemple :
<sp who="CH"><speaker>CHARLES</speaker></sp>
<sp who="CH"><speaker>LE ROI</speaker></sp>

4.3.3 Application d'une feuille de style

Comme nous l'avons montré, le codage met l'accent sur la valeur sémantique des informations et non sur leur mise en page. Une feuille de style contient quelques règles simples qui vont interpréter ces éléments et leur appliquer un style au choix, qui va rendre le document lisible, présentable voire imprimable.

Ainsi, l'extrait suivant du fichier XML :

```
<div type="acte" n="1">
<head>ACTE PREMIER</head>

<div type="scene" n="1">
<head>SCÈNE PREMIÈRE</head>
<stage>LA REINE, STRAFFORD</stage>
<sp who="MH"><speaker>LA REINE</speaker>
<l n="1">Arrêtons-nous, Strafford, je me soutiens à peine!...</l>
<stage>Elle s'assied.</stage>
<l n="2">En l'état où je suis, qui me croirait la Reine?</l>
<l n="3">Moi qui reçus le jour pour imposer des lois,</l>
<l n="4">Il faut, en abordant le palais de vos Rois,</l>
<l n="5">À l'heureuse indigence emprunter sa livrée!...</l>
[…]
</sp>
```

pourrait ressembler, une fois interprété par une feuille de style XSL, à ceci :

ACTE I

SCÈNE PREMIÈRE

LA REINE, STRAFFORD

LA REINE

Arrêtons-nous, Strafford, je me soutiens à peine!...

Elle s'assied.

En l'état où je suis, qui me croirait la Reine?

Moi qui reçus le jour pour imposer des lois,

Il faut, en abordant le palais de vos Rois,

À l'heureuse indigence emprunter sa livrée!...

La feuille de style écrite le théâtre ainsi que la pièce *Cromwell* publiée intégralement sont consultables sur le site de l'INaLF[11].

5 Conclusion et perspectives

5.1 Quelle stratégie pour faire évoluer les grands fonds ?

Nous avons essayé, par l'intermédiaire de ce document, de faire partager notre expérience acquise dans le cadre de nombreux projets nationaux et internationaux. Tous avaient comme point commun l'accès et la distribution de ressources linguistiques structurées en élaborant, à partir de grands fonds de ressources assemblés dans le passé. A terme, cette tâche s'est révélée, certes nécessaire, mais souvent lourde et laborieuse à mener à bien.

Il est donc important, lors du montage de ce type de projet, de clairement définir les objectifs à atteindre afin d'opter pour une stratégie adéquate permettant, dans le laps de temps imparti et avec les forces en présence, de faire évoluer la plus grande partie des fonds existants vers des ensembles de ressources normalisées de référence. Cette stratégie doit tenir compte des besoins et des objectifs finaux, de la granularité du codage, de la pérennité des données à construire et ne doit en aucun cas tenir compte du type d'application qui exploitera les données. En effet, il est primordial de séparer les ressources à utiliser de l'application même si, parfois, la frontière entre les données et les fonctionnalités est très étroite. L'avantage de cette démarche est d'obtenir des ressources textuelles indépendantes de toute application. On évite ainsi la duplication des données et on ouvre la porte à de futurs applications et besoins. Dans le même état d'esprit, cela revient à considérer une ressource primaire comme une entité vivante, qui va s'enrichir et s'améliorer avec le temps, par l'intermédiaire d'un système d'annotations externes.

[11] http://zeus.inalf.cnrs.fr/sbt/

5.2 Illustrations directes

Afin d'illustrer ces propos, voici une brève description de projets retenus dans le cadre du programme Européen MLIS[12].

5.2.1 Le projet DHYDRO

DHYDRO[13], projet de MLIS, a pour objet de créer sur Internet un espace terminologique multilingue spécialisé dans le domaine de l'hydrographie (Blampain et alii, 1999). Il s'agit de développer à la fois des outils d'aide à la rédaction et des interfaces de consultation et d'interrogation.

La normalisation est ici doublement au cœur des préoccupations puisqu'il s'agit, d'une part, de faire converger les outils et méthodes de travail de rédacteurs répartis sur plusieurs continents et, d'autre part, d'appliquer un modèle normatif à un type de données particulier : les terminologies. Dans ce domaine, le niveau de structuration des informations est extrêmement élevé; l'élaboration de normes terminologiques, d'ailleurs, a préoccupé les terminologues avant les informaticiens, et dans les mêmes termes : rendre les documents réutilisables par d'autres, échangeables, et maintenant interprétables par des automates standards. L'avènement du document électronique permet de faire converger les deux aspects et d'offrir aux terminologues des outils et des techniques appropriées pour atteindre ces objectifs.

Parmi les formats disponibles, le choix pour DHYDRO s'est porté sur MARTIF[14] (norme ISO-12200) qui s'appuie sur SGML et sur les travaux de la TEI. A l'heure actuelle, MARTIF permet de représenter l'ensemble des données de DHYDRO en préservant toutes les informations existantes. Ce format ajoute même une richesse de codage qui ouvre des perspectives nouvelles en termes de fonctionnalités d'un outil de saisie et en termes de finesse de consultation.

5.2.2 Le projet ELAN

Le but du projet ELAN est de fournir une alternative à la lourdeur de mise en œuvre d'une procédure de normalisation de grands fonds de ressource linguistique. Les objectifs du projet sont de définir :

1. CQL, un langage de requête commun
2. une architecture distribuée permettant un accès aisé à de grande base de données linguistiques hétérogènes réparties sur des sites distants.

---

[12] MLIS : **M**ulti**L**ingual **I**nformation **S**ociety

[13] http://www.loria.fr/projets/MLIS/DHYDRO

[14] MARTIF : **Ma**chine-**R**eadable **T**erminology **I**nterchange **F**ormat

La normalisation des données intervient non pas en amont du projet et de façon permanente, mais est prise en compte lors de la demande des données par l'utilisateur via le langage de requête et de manière transitoire. La spécification du codage ne prend pas en charge le codage des ressources primaires (textes éventuellement annotés) mais plutôt le codage des requêtes et surtout celui des résultats provenant des différents serveurs en réponse à chacune des requêtes.

6 Références

Abiteboul S. 1997, Querying semi-structured data. Actes *ICDT'97*.

Abiteboul S., Quass D., McHugh J., Widom J, et Weiner J. L. 1997. The lorel query language for semi-structured data. *Journal of Digital Libraries,* 1(1).

Blampain D., Descotte S., Husson J.-L., Rohde H., Romary L., Van Campenhoudt M. et Viscogliosi N. 1999, Le projet européen DHYDRO : la normalisation à l'épreuve d'un forum terminologique. *Conférence sur la Coopération dans le Domaine de la Terminologie en Europe* de l'AET (Paris, 17-19 mai 1999).

Bonhomme P., Bourion E., Cruz-Lara S., Jadelot C., Rastier F., Romary L., de Saint-Rat C. et Viscogliosi N. 1998, Rapport d'étape Silfide 1998.

Bruneseaux F. 1998, Noms propres, syntagmes nominaux, expressions référentielles : repérage et codage., *Langues, 1*, 1, pp. 46-60.

Buneman P., Davidson S., Hillebrand G., et Suciu D. 1996, A query language and optimization techniques for unstructured data. *Actes d'ACM-SIGMOD International Conference on Management of Data,* pp. 505-516, Montreal, Canada.

MARTIF, ISO/FDIS 12200 1998, Applications informatiques en terminologie - Format de transfert de données terminologiques exploitables par la machine (MARTIF) - Transfert négocié, Genève, ISO.

Jakobson 1970, *Essais de linguistique générale*, , Paris, Ed. du Seuil.

Romary L., Bonhomme P., Bruneseaux F. et Pierrel J.-M. 1999, Silfide : a system for open access and distributed delivery of TEI encoded documents. *Computers and Humanities*, 33, pp.31-38.

SGML, ISO 8879 1986, Information processing — Text and office systems — Standard Generalized Markup Language (SGML).

TEI-P3, Association for Computers and the Humanities (ACH), Associatin for Computational Linguistics (ACL) and Association for Literary and Linguistic

Computing (ALLC) 1994, Guidelines for Electronic Text Encoding and Interchange (TEI-P3), 2 vol..Ed.C.M. Sperberg-McQueen and Lou Burnard, Chicago, Oxford : Text Encoding Initiative.

Ubersfeld A. 1996, *Lire le théâtre I*, Paris, Belin, 1996, p. 17-18.

XML, Extensible Markup Language (XML) 1.0, W3C Recommendation 10-February-1998, Bray T., Paoli J., Sperberg-McQueen C.M.