Première partie

Documentation TEI

L'édition électronique de la correspondance d'Armand Horel, est une édition XML-TEI. Cette partie est relative à la documentation concernant l'encodage et la production de fichiers.

La Bibliothèque de documentation internationale contemporaine (BDIC) s'est lancée dans un important programme de numérisation de ses collections, accessibles depuis sa bibliothèque numérique [1]. Proposer une édition électronique à partir de ce corpus numérisés, permet de mettre en valeur certains fonds, tout en offrant des outils de recherche que les versions fac-similaires ne permettent pas.

L'édition électronique de la correspondance d'Armand Horel s'inscrit, plus largement, dans un projet visant à offrir, aux chercheurs qui le souhaitent, une solution leur permettant de réaliser ce type d'exercice dans le cadre de leur travaux.

1. voir : http ://argonnaute.u-paris10.fr/

Chapitre 1

Introduction

Procéder à une édition en ligne nécessite se s'interroger sur le langage à utiliser pour l'encodage du texte. Une publication web repose sur des pages HTML et nécessite donc, *a minima*, un balisage présentationnel de la source. Toutefois, depuis la publication en 2009 du Référentiel Général d'Interopérabilité (RGI) par la Direction générale de modernisation de l'État, c'est la technologie XML (langage de balisage descriptif) qui est favorisée en ce qu'elle assure la pérennisation et interopérabilité de l'information. Parallèlement, le dialecte XML-TEI, dont le champ d'application est maintenant relativement étendu, est devenu un standard concernant l'édition scientifique des sources primaires. C'est donc d'elle même que cette solution s'est imposée pour le développement d'un modèle d'édition de correspondance.

Ce choix n'est pas pour autant une solution applicable directement. Le balisage du texte nécessite de réfléchir à l'élaboration d'un schéma, contenant les règles à respecter concernant l'usage de la TEI et qui oblige à procéder avant toute chose à une analyse structurelle et intellectuelle de la source.

Chapitre 2

Langage de balisage et Text encoding initiative (TEI)

2.1 La balisage descriptif

D'une manière générale, les langages de balisage sont des dialectes informatiques adaptés à l'enrichissement d'informations textuelles. De plus, sa structure dans laquelle une balise correspond à une unité syntaxique délimitant une séquence au sein d'un flux de caractères est compréhensible par la machine. On dénombre trois catégorie de langage de balisage :

- les langages procéduraux dans lesquels les balises correspondent à des instructions exécutables par un programme informatique,
- les langages présentationnels qui ont pour fonction de mettre en forme le texte,
- les langages descriptifs qui présentent l'avantage de distinguer le contenu de la forme permettant ainsi d'en séparer le traitement. Ce dernier garantie ainsi une meilleure portabilité des fichiers numériques.

Un langage descriptif réside donc avant tout dans l'analyse de la source textuelle au sein de laquelle il convient d'identifier sa structure sémantique nonobstant tout traitement présentationnel. Il s'agit dont enrichir le document d'informations sémantiques afin de pouvoir proposer par exemple plusieurs lectures d'un document. trois grande étapes sont nécessaires à la réalisation de ce travail :

- L'analyse sémantique du document : Il s'agit d'identifier les différent éléments qui composent le texte (paragraphe, signature, date, noms de personne...
- Le choix des balises : quelle balise peut-on appliquer à un élément donné.
- L'encodage.

Ce type de langage descriptif repose sur une structure hiérarchique. Comme nous l'avons vu, le balisage repose sur l'identification d'une séquence au sein d'un flux de carac-

tère. C'est hiérarchique parce que les éléments (paragraphes, phrases...) sont imbriqués les uns dans les autres et sont donc liés entre eux par une relation linéaire.

Les langages à balises descriptifs présentent au final un certain nombre d'avantages :

— La processus d'établissement du texte est simplifié car il ne repose que sur identification du contenu non pas sur sa présentation ou la compréhension du programme.
— Le document est indépendant de l'apparence formelle que l'on souhaite lui donner.
— Ce sont des langages interopérables qui facilitent le partage de données.

Ce type de langage est tout à l'avantage des éditeurs dans la mesure où ; on limite les risques d'incompatibilité, ce type de fichier permet de proposer plusieurs éditions à partir du même artefact numérique, enfin, il est possible de générer de manière automatique les informations bibliographiques du document évitant ainsi des erreurs ou autorisant leur versement automatique dans des bases de données en ligne.

En revanche ce choix implique bien souvent un cadre technique de travail moins confortable que l'utilisation d'un traitement de texte que tout un chacun maîtrise plus ou moins.

La technologie XML (eXtensible markup Language), est un métalangage de balisage structuré, c'est à dire qu'il respecte une structure hiérarchique formant une arborescence. Elle permet le développement de dialectes descriptifs interopérables. En effet, si sa structure est compréhensible par la machine, ce n'est pour autant pas un langage à proprement parlé et ne propose donc pas une véritablement sémantique. XML se contente d'énoncer un ensemble de règles sur ce que doit être un document bien formé et valide ; il nécessite donc d'élaborer, ou de choisir un vocabulaire spécialisé, comme la TEI.

2.2 Text Encoding Initiative (TEI)

La TEI est un groupement international qui a pour finalité de développer et maintenir un standard pour l'édition de texte sous forme numérique. Il s'agit donc d'un vocabulaire XML spécialisé dans l'édition des sources primaires. Son champ d'application est maintenant relativement étendue, ouvrages imprimés anciens, textes médiévaux, chartes et documents, cours écrits ou oraux etc. Toutefois, ce que propose la TEI n'est pas tant un schéma général qu'un cadre de développement composé d'un dialecte explicité par une documentation, le tout réuni dans des *Guidelines*.

A la fins des années 1980, des chercheurs et universitaires, déjà impliqués dans la production de textes sous forme numérique font le constat d'un manque de solution

concernant l'échange de texte résultant de leur recherches. En 1987, lors d'une rencontre organisée par l'Association for Computers and Humanities (ACH), au Vassar College de Poughkeepsie, une trentaine de chercheurs et professionnels s'accordent sur le besoin de développer un cadre de pratique commune. Ils aboutissent à la formulation des principes de Poughkeepsie, dont la finalité est l'élaboration de *Guidelines* (recommandations) avec pour objectifs principaux de :

— Fournir un format d'échange standard de données pour la recherche en Humanités,
— Proposer des principes d'encodage de texte dans ce même format,
— Définir une syntaxe et un schéma,
— Garantir autant que possible la compatibilité avec les standards existants.

L'ACH, rejoint par l'Association for Literary and Linguistic Computing et l'Association for Computational Linguistic établirent la Text Encoding Initiative (TEI) afin de mener le projet, dans plusieurs langues et à un niveau international. Les premières *Guidelines* furent publiées en 1993. Basé à l'origine sur la technologie SGML (Standard Generalized Markup Language), la TEI embrassera XML dès sa création en 1996. Nous sommes actuellement à la cinquièeme version de la TEI dénommée P5 ; elle présente maintenant un certaine nombre d'avantage :

— Adaptabilité à toute forme de document
— Expressivité de part la granularité qu'elle propose
— C'est une standard internationnale assurant interopérabilité et pérennité.
— Elle est basé sur la technologie XML,
— Particulièrement adapté à l'édition électronique.

2.2.1 La production de source primaire

Plus que la publication d'un standard, la TEI émet des recommandations relativement simples et compréhensibles sur les conventions d'encodage à adopter. De plus sa structure "modulaire" lui permet de répondre à différentes problématiques particulières. Si l'on ajoute le fait, comme nous l'avons vu, que la TEI repose sur la syntaxe XML standardisée par le World Wide Web Consortium (W3C) on comprend qu'elle se soit rapidement imposée comme un standard pour l'édition électronique de sources primaires.

Un schéma généraliste

La TEI se veut suffisamment riche pour être adaptée à la multitude des champs des humanités. Cependant, si l'objectif affiché de proposer un standard est censé assurer interopérabilité et pérennité, la généricité de ce dernier, oblige à faire des choix devenant par la même occasion un frein à ces objectifs. Ces choix sont nécessaires pour cadrer au maximum une pratique d'encodage et permettre ainsi un traitement automatisé des

fichier. Ainsi l'encodage TEI, n'est pas un fin en soi et il ne suffit donc pas qu'un texte soit encodé avec ce langage pour assurer sa compatibilité. Les documents doivent être avant issue d'une même logique éditoriale.

Lecture du texte

Une modélisation répond avant tout à un besoin éditorial et à la représentation que l'on veut donner du texte parmi la multitude de lecture possible. La TEI, parce qu'elle est basée sur la structure hiérarchique de XML, impose une certaine vision du texte. Cette syntaxe XML permet de répondre aux besoins les plus généraux tout en facilitant par la suite le traitement et la transformation des éléments TEI en éléments HTML. Dans d'autres cas l'arborescence imposé par le modèle XML, peut sembler inadaptée voir lourde car elle impose tout de même une interprétation éditoriale de la source. Bien qu'imparfaite cette solution reste dans l'immense majorité des cas acceptable et suffisante ; de plus elle permet de bénéficier des puissants outils XML. C'est donc de l'interprétation du texte qu'il est question. Il faut bien avoir à l'esprit que la TEI, bien que considérée comme le standard d'édition dans le domaine académique, n'est pas pour autant une solution générique pouvant répondre à tous les cas de figures, une modélisation, même si elle peut répondre à plusieurs problématiques, propose déjà une interprétation du texte et donc une lecture, "reflet des des questionnements à l'œuvre au moment de l'acte d'édition".

Chapitre 3

Modélisation et TEI

3.1 Modélisation TEI et personnalisation

3.1.1 Modélisation

Ce que propose la TEI n'est pas tant un schéma qu'un cadre de développement. Tout d'abord parce qu'elle procure à l'éditeur un ensemble de recommandation qu'il est parfois nécessaire d'adapter aux spécificité du projet d'édition. Ensuite parce qu'elle adopte une organisation "modulaire". Chaque module répondant plus ou moins à un type de problématique. C'est ce qui permet à la TEI d'être particulièrement flexible, même si cela peut dérouter l'encodeur inexpérimenté. Il existe souvent plusieurs solutions d'encodage pour un objet donné. Ce choix doit alors être pris en ayant à l'esprit tous les "tenants et aboutissants, impliquant des retours incessants vers les recommandations de la TEI"

Ceci explique cette nécessité de personnaliser le modèle TEI afin de l'adapter au mieux aux spécificités du texte que l'on encode. On imagine aisément que les besoins ne sont pas les mêmes pour l'encodage d'un texte médiéval et celui d'un correspondance contemporaine.

3.1.2 Personnalisation

Réalisation d'un schéma

La TEI a été imaginée comme un ensemble de module, que l'on peut assembler afin de répondre aux spécificités d'un projet. Toute modélisation est établie par un schéma correspondant à un sous ensemble de la TEI et répondant aux besoins de son projet. Ce schéma permet à la fois de contrôler la production d'un fichier, mais aussi de valider le contenu d'un fichier en associant ce dernier au schéma.

Un schéma d'établi généralement à partir d'un échantillon représentatif du corpus à encodé. il n'est pour autant susceptible d'évoluer au fur et à mesure de l'avancée de

travaux.

En revanche ce n'est pas tant un schéma qui est proposé qu'un cadre de développement. La TEI adopte une organisation modulaire. Chaque module répond à une problématique spécifique en définissant un certain nombre des composantes d'une modélisation (éléments, attributs, classes. . .). Ils sont documentés par un chapitre dans les Guidelines [1] et chacun est libre des les inclure ou non, en fonction de ses besoins dans sa personnalisation.

1. Recommandation TEI