I – Introduction à la XML–TEI : pourquoi encoder ses corpus ?

A – Historique de la création de la XML–TEI

Dans les années 1980, le développement de l'utilisation des technologies numériques pour des projets d'étude ou d'archives s'est vu confronté à la prolifération de systèmes de représentation numérique de texte. Plus des projets de création de textes électroniques se mettaient en place, plus il y avait de nouveaux systèmes, qui se trouvaient être incompatibles les uns avec les autres. Cela représentait un véritable frein dans l'avancée technologique dans cette discipline, puisque cela faisait découler des problèmes de conservation, de pérennisation et de partage de données. Ainsi, en 1987, une réunion eut lieu afin d'aborder ces problèmes. Regroupant des spécialistes de diverses disciplines et des professionnels des archives ou bibliothèques, venant d'Europe, Amérique du Nord et Asie, cette réunion marque la genèse de la TEI. Le travail fut confié à trois organisations sponsors et trois comités de travail furent aménagés pour lancer la mise en place de règles d'encodage.

Le premier jet des Guidelines (P1) est produit en juin 1990. D'autres versions sortiront par la suite, aidés par la création de nouveaux groupes de travail et par de nombreuses révisions, extensions et modifications qui mènent à la première version officielle des Guidelines, la P3 en mai 1994. À l'époque, on ne parle pas encore de XML-TEI puisque le format XML n'existe pas. La TEI s'écrivait, à cette période, sous la forme de SGML ou *Standard Generalized Markup Language*, que l'on peut considérer comme l'ancêtre de la XML. Ce langage de balisage fonctionnait avec une DTD, c'est-à-dire un moyen de définir la structure et les éléments du document.

En janvier 1999, une demande fut faite par deux universités pour établir la création d'une organisation internationale, appelé le TEI Consortium, qui aurait pour objectif le maintien, le développement et la promotion de la TEI. Après diverses démarches, le consortium fut officiellement établi en janvier 2001. L'intérêt de ce consortium était de maintenir la TEI dans une organisation autosuffisante, non lucrative, indépendant économiquement et académiquement. De plus, elle avait pour but de se créer une communauté d'utilisateurs qui travailleraient constamment au développement et à l'utilisation généralisée de la TEI. Ces deux objectifs du consortium ont été atteints, ce qui s'observe notamment par l'impact de la TEI aujourd'hui, qui est reconnu internationalement, utilisé comme premier choix pour la production d'éditions numérique et de nombreux autres projets, qui sont pour une grande partie listée sur le site de la TEI.

Cette popularité de la TEI a entraîné la nécessité d'une publication d'une nouvelle version des guidelines. Tout d'abord, en juin 2022, la version P4 sort : c'est celle qui va introduire le XML au sein de la TEI. Le format est né quelque années auparavant, en 1999 et est considéré comme une version simplifiée du SGML, et qui permet de décrire des documents

structurés. À cet époque, on garde aussi la SGML dans l'utilisation de la TEI. Après cela, au regard de la dizaine d'années qu'avait déjà la version P3, il a été décidé le développement d'une nouvelle version révisée, améliorée et comprenant de nouveaux développements dans un certain nombre de domaines, tels que la description de manuscrits ou l'encodage d'images. Cette version, appelée les P5 Guidelines, est sortie en novembre 2007 et le SGML y est abandonné. Cette version est celle encore utilisée aujourd'hui, mais agrémentée par les divers changements et ajouts qui ont eu lieu au fur et à mesure des années.

B – Qu'est–ce que le format XML ?

a) Qu'est-ce que le format XML ?

XML ou *eXtensible Markup Language* (langage de balisage extensible) est un standard soutenu par le consortium World Wide Web pour le balisage de documents. Son but principal, en tant que langage, est le partage de données entre différents systèmes, tel que l'Internet. Cela lui est possible notamment grâce à son format souple qui lui permet d'être adapté à divers contextes. Il ne peut cependant pas être utilisé comme langage de programmation. Il sert simplement à stocker et à transporter de la donnée.

C'est un langage pour des documents textuels dont les données sont délimitées par des balises, que l'on appelle *élément* et qui se caractérisent par la présence de chevrons (<, >)

Il a la particularité d'être un langage extensible, c'est-à-dire qu'il n'a pas de balises prédéfinies. Il est donc possible à la fois de créer et de définir son propre jeu d'étiquettes (les éléments) mais aussi d'étendre ce jeu d'étiquettes. Il est même possible d'en mêler les uns avec les autres, en spécifiant lesquels sont utilisés, par le biais d'espaces de noms qui permettent d'éviter la confusion.

b) Comment ça fonctionne ?

La XML fonctionne en suivant une hiérarchie que l'on désigne sous le nom d'arbre. Il se compose d'un élément racine (qui se caractérise par la première paire de balises de l'arbre) puis de balises imbriquées. Elles seront désignées sous le terme de balises frère/sœur si elles sont au même niveau et de parent et enfant si elles ne le sont pas.

Bien que cela ne soit pas considéré comme un élément obligatoire dans l'arbre XML, il est tout d'abord possible de mettre le prologue du fichier en premier ligne du document XML. Ce prologue se présentera presque toujours de la manière suivante :

```
<?xml version="1.0" encoding="UTF-8"?>
```

Cela indique la version de XML qui sert de base à l'écriture (il n'y en a qu'une seule et donc, cela sera toujours 1.0) et le codage des caractères dans le document. La majorité des

documents sont encodés en UTF-8, ainsi, c'est la donnée que l'on trouve principalement. Il est important de bien rentrer les attributs dans cet ordre-là pour le prologue.
Avec ce prologue, le programme qui lira le fichier saura qu'il s'agit d'un document XML dont il devra lire des caractères encodés en UTF-8. Le prologue n'appartient pas vraiment à l'arbre, il n'a pas besoin d'être fermé et il ne peut pas être considéré comme l'élément racine.
→ Lorsque vous travaillerez sur l'encodage avec Oxygen, ce prologue sera automatiquement écrit lorsque vous choisirez de faire un document XML.

XML est un langage extensible alors, il est toujours possible de créer son propre vocabulaire et d'encoder avec des balises que l'on a créées soi-même. Cependant, il existe tout de même certaines règles impératives à suivre dans la mise en place d'un arbre XML.
→ Les exemples qui seront montrés par la suite ont été créés avec l'éditeur de texte *Sublime Text*, ce qui nous donne ainsi les couleurs proposées par cet éditeur. Cette coloration peut être présente pour tout autre éditeur de texte, à condition de spécifier que l'on est en XML.

1. Une balise ouvrante devra toujours être suivie quelque part dans l'arbre de sa balise fermante, qui se caractérisera par la présence d'un antislash au début.

```
<titre>Présentation de personnes</titre>
```

Exception faite des balises auto fermantes, aussi appelées balises vides, qui ne comporteront pas de textes, mais qui devront dans ce cas avoir l'antislash à la fin de la balise, telle la balise de saut de ligne *
*

2. XML est sensible à la casse, alors il est important qu'une balise fermante ait exactement la même casse que sa balise ouvrante.

Ainsi, cela est valide :

```
<titre>Présentation de personnes</titre>
```

Mais cela serait erroné pour XML :

```
<titre>Présentation de personnes</Titre>
```

3. L'imbrication doit être rigoureusement respecté. Les éléments doivent être fermés dans leur ordre d'ouverture pour qu'il n'y ait pas de chevauchement et que l'arbre soit juste et bien formé.

Bonne version :

```
<personne><nom>Martin</nom></personne>
```

Mauvaise version :

```
<personne><nom>Martin</personne></nom>
```

4. Il est possible de rajouter des attributs avec des valeurs à certains éléments pour leur donner plus de précision. L'attribut ne se place que sur la balise ouvrante ; il ne doit pas être remis sur la balise fermante. Il est possible de mettre plusieurs attributs sur un même élément et l'ordre n'a pas d'importance.

Exemple :

```
<personne n="1" type="homme" profession="professeur">
    <nom>Martin</nom>
    <prenom>Jean</prenom>
</personne>
```

Prénom sans accent, car à éviter lors de la création de balises, puisque cela pourrait être mal interprété par certains programmes.

Ces valeurs doivent obligatoirement être mises entre des apostrophes ou entre des guillemets. Ainsi, cela sera incorrect :

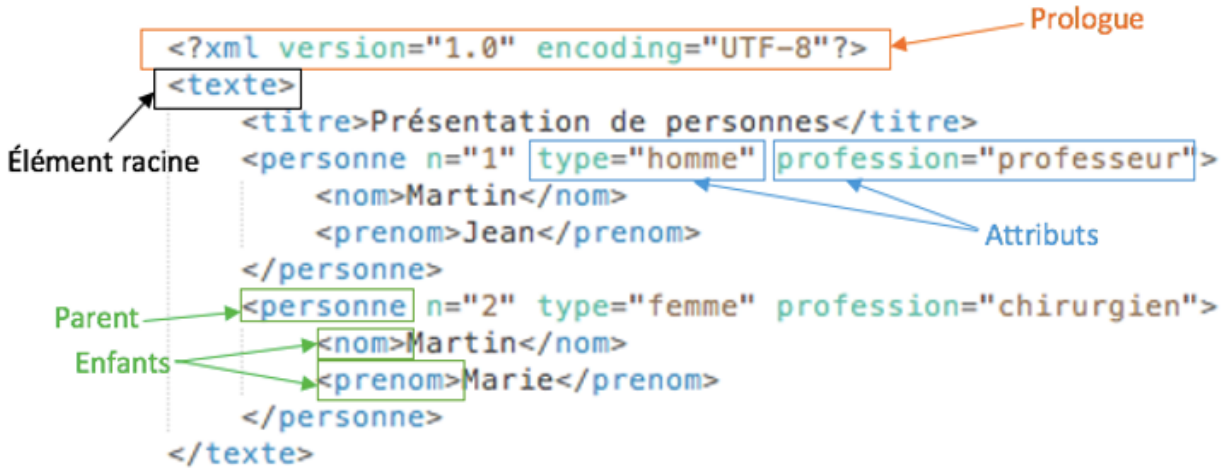
```
<personne type=homme>
```

Mais cela sera correct :

```
<personne type="homme">
```

On peut voir d'ailleurs ici l'impact que cela a puisque l'encodage a été fait avec un éditeur de texte et comme on peut l'observer, la valeur de l'attribut prend une couleur quand elle est correctement encodée alors qu'elle reste noire et donc mal prise en compte quand les guillemets ne sont pas présents.

Démonstration d'un arbre bien formé (l'imbrication est bonne et les balises sont toutes correctement fermées) et des parties importantes (prologue, élément racine, parent et enfants, attributs).



C – Qu'est–ce que la TEI ?

Je vous ai présenté, en première partie, un historique de la création de la TEI et du XML. Comme je l'ai dit, la TEI englobe un certain nombre d'éléments, et notamment des guidelines et un consortium, que je vais vous présenter plus en détail dès maintenant.

a) Les TEI Guidelines

La TEI est un standard pour la représentation de texte dans leur format numérique et il est donc nécessaire d'avoir une série de règles à suivre pour pouvoir être conforme à ce standard : c'est ce que sont les TEI Guidelines.

Représentée par une très longue liste détaillée et extensive, elles spécifient les règles d'encodage pour des documents lisibles par machine, notamment dans le domaine des sciences humaines et sociales et de la linguistique. De plus, elles sont en constante évolution grâce au travail de certains groupes, que je développerai par la suite et sont mises à jour dès que cela est nécessaire.

Le site internet qui contient ces guidelines est d'autant plus pratique, qu'il permet une navigation générale, mais également par thème, des règles d'encodage. Il est ainsi possible, au lieu d'avoir à chercher en vain une balise dont on ne connait même pas le nom, d'aller dans le chapitre qui convient pour trouver la balise que l'on cherche. Les premiers chapitres permettent d'avoir les balises assez génériques qui composent la TEI, tel que ce qui devra être utilisé pour encoder les métadonnées du header ou les éléments principaux qui encodent le corps du texte. Mais il est aussi possible d'aller plus en détail : il existe des séries de balises faites pour l'encodage de poèmes et autres textes composés de vers (Chap. 6 Verse), pour l'encodage de métadonnées lié à des manuscrits historiques (Chap. 10 Manuscript Description) ou encore l'encodage de tout ce qui a trait aux entités nommées d'un texte, c'est-à-dire les personnes, lieux, organisations, etc. (Chap. 13 Names, Dates, People and Places).

Depuis ces chapitres, mais aussi avec la page d'accueil qui recense tous les éléments et les attributs, il est également permis d'aller plus en profondeur dans la recherche et de regarder tout ce que peut faire un élément spécial. Une fois la page spécifique d'un élément atteint, on obtient un très grand nombre d'informations dessus : le module auquel il appartient (utile lors de la documentation); les attributs qu'il peut contenir; les balises qu'il peut contenir et à l'inverse les seules balises dans lequel on peut le trouver ; et enfin, des exemples d'utilisation de la balise.

b) Consortium collaboratif

Comme cela fut expliqué lors de l'histoire de la création de la TEI, le consortium a pour objectif de maintenir, développer et promouvoir la TEI. Cela peut notamment s'observer par

la mise en place de certains services et outils, développés par le consortium. Je vais en présenter quatre différents, qui ont chacun un objectif particulier : la promotion de la TEI pour le journal et pour les conférences, le développement de la TEI pour les SIG, et l'assistance avec la mailing-list.

1. *JTEI : Journal de la TEI*

Le Journal de la *Text Encoding Initiative* est le journal officiel du consortium TEI. Il regroupe des articles d'état de l'art sur la TEI, d'utilisation actuelle, notamment sur des projets ou encore des découvertes et nouvelles utilisations de la TEI. Le journal vise à être également un lieu de discussion entre la TEI et les auteurs de la communauté, à travers les articles qui sont postés. Cela représente un moyen adéquat pour connaitre les nouveautés sur la TEI et les innovations qui se développent.
Le journal est hébergé sur Open Edition et les articles publiés font au préalable l'objet d'une lecture, relecture et vérification approfondie par un examen rigoureux des pairs de la communauté.

2. *Conference/Annual Meetings*

Chaque année, depuis près de 20 ans, le consortium TEI organise une conférence se déroulant sur environ une semaine et qui regroupe des personnes venant des quatre coins du monde, afin de participer à des ateliers, des réunions particulières (comme les SIGs que je présenterais juste après) et des sessions de courts et longs papiers. Cela représente une occasion pour les différents membres de la communauté TEI de se retrouver, d'échanger, de découvrir ce qui se fait actuellement et de prévoir de nouveaux développements. C'est également l'occasion pour les responsables de la TEI d'avoir leur réunion annuelle afin d'élire ou de réélire leurs membres, de discuter des affaires du bureaux et de présenter le bilan de l'année. Cela conduit parfois aussi à la publication d'une édition spéciale dans le jTEI, qui reprend certaines des présentations faites durant cette semaine.

3. *Special Interests Groups (SIGs)*

Les SIGs sont des groupes de travail mis en place par le consortium TEI et qui permettent de regrouper des amoureux de la TEI qui voudraient échanger sur certains sujets spécifiques à propos de points particuliers. Ces groupes sont ouverts à tous et ils permettent de discuter avec des pairs de sujets qui tiennent à cœur. TEI donne plusieurs moyens pour développer ces groupes, dont notamment un espace web et une *mailing list*.
Les groupes ont tendance à se concentrer sur un aspect particulier de la TEI, dont ils peuvent ensuite débattre et échanger, pour possiblement donner naissance à quelque chose

de nouveau. Le groupe peut ne servir qu'à échanger, mais il est aussi possible de développer une documentation ou même une extension des Guidelines parfois.
Pour vous donner un exemple du type de réflexion faite par les SIGs, je peux vous en présenter brièvement trois :

- *Computer-Mediated Communication SIG* : la communication virtuelle (messagerie instantanée, réseaux sociaux, etc.) est de plus en plus présente et peut être utilisé dans du travail d'humanités numériques. Cependant, la TEI n'a pas de standard de représentation pour ce type de textes et le SIG travaille donc à des suggestions pour adapter la TEI.
- *Correspondence SIG* : les membres de ce groupe ont pour objectif de fournir plus de balises afin de mieux représenter la composition d'une correspondance, quel que soit son type (e-mails, journal intime, lettre avec enveloppe, télégramme), sur le point de vue de l'apparence et de la structure du contenu. Ce groupe a notamment abouti à la création d'une balise <corrispDesc> accompagné d'un <corrispAction> et <corrispContext> qui permettent de donner des informations sur les expéditeur et destinataire de la lettre et sa chronologie dans un corpus.
- *East Asian/Japanese SIG* : L'encodage de documents écrit en langue japonaise ou d'Asie de l'Est se développe beaucoup depuis un certain nombre d'années et les chercheurs font face à un manque d'éléments et d'attributs appropriés. Le SIG a donc pour objectif de pallier ce manque, de créer des guidelines appropriées pour les textes en japonais et dans la longueur, d'étendre le groupe à plus de langages d'Asie de l'Est.

4. TEI Mailing-List

La TEI est une vaste communauté composée de personnes avec diverses connaissances. Ils ne sont pas seulement membres de la communauté pour utiliser des standards de la TEI mais également pour aider ceux qui peuvent avoir des difficultés, notamment dans des cas d'utilisation spécifique, mais non spécifié, de balises. La mailing-list est ouverte à tous et il est toujours possible d'envoyer un message qui expose le problème rencontré, ainsi que les solutions possibles envisagées. Toute personne avec une idée de réponse est ensuite encouragée à répondre, puisque cela peut montrer des cas similaires ou au contraire des divergences.

Si nécessaire de présenter un exemple, possible de présenter mon cas précis avec le *closer*/*postscript* ou la lettre dans la lettre (mailing-list correspondance)

De plus, la mailing-list garde trace de tous les fils de discussion, depuis la question posée, jusqu'à la dernière réponse donnée. Cela permet avant même d'avoir à poser une question, de voir si une autre personne n'a pas déjà rencontré le problème et ainsi, tel que dans un forum, ne pas avoir de doublon.